UNIVERSITY OF BELGRADE

FACULTY OF CIVIL ENGINEERING

Milutin M. Pejović

GEOSTATISTICAL MODELING OF GEOCHEMICAL VARIABLES IN 3D

DOCTORAL DISSERTATION

Belgrade, 2016

UNIVERZITET U BEOGRADU

GRAĐEVINSKI FAKULTET

Milutin M. Pejović

GEOSTATISTIČKO MODELIRANJE GEOHEMIJSKIH PROMENLJIVIH U 3D PROSTORU

DOKTORSKA DISERTACIJA

Beograd, 2016

Mentori:

v. prof. dr Branislav Bajat, dipl. inž. geod., Građevinski fakultet, Beograd

v. prof. dr Zagorka Gospavić, dipl. inž. geod., Građevinski fakultet, Beograd

Članovi komisije:

doc. dr Milan Kilibarda, dipl. inž. geod., Građevinski fakultet, Beograd

dr Dragan Čakmak, viši naučni saradnik, Institut za Zemljiste, Beograd

dr Tomislav Hengl, dipl. inž. šum, Građevinski fakultet, Beograd (gostujući prof.) / ISRIC, Wageningen University and Research, the Netherlands

Datum odbrane:

To my beloved wife, Ksenija

Acknowledgements

First of all, I would like to express my sincere gratitude to my supervisor Prof. Branislav Bajat for the continuous support, patience and motivation during my Ph.D studies. Without his mentorship and guidance, it would not have been possible to complete the work in time.

My sincere gratitude goes to my co-advisor Prof. Zagorka Gospavić for her encouragement and continuous support.

Special thanks goes to my friend and a member of committee, prof. Milan Kilibarda, who introduced me to the fantastic world of R, spatial statistics and science. His contribution to this work is invaluable.

I would like to thank Professor Dragan Čakmak for helping me to better understand the nature of soil. Thanks also goes to the Institute for Soil Science in Belgrade, for providing me with valuable soil data.

I am very grateful to Tom Hengl, Mladen Nikolić and Gerard Heuvelink for helping me define my ideas, and for resolving the main methodological issues in this work.

Big thanks also goes to my Prof. Branko Milovanović for always being willing to share his valuable scientific knowledge and professional experience with me.

I also want to thank my friend Sofija Nemet for language corrections. Any remaining errors that might still be present are my responsibility and are due to last minute changes.

I feel the need to thank all of my close friends for their invaluable emotional support.

I am deeply thankful to my family, my father, mother and my brother, for their faith in me and their unconditional love and support, throughout my whole life.

This last word of acknowledgment I have saved for my beloved wife, Ksenija, to whom this thesis is dedicated. Without you, I would never be as happy as I am now.

Milutin Pejović

UNIVERSITY OF BELGRADE

Abstract

Faculty of Civil Engineering
Department of Geodesy and Geoinformatics

Geostatistical Modeling of Geochemical Variables in 3D

Geostatistical mapping of soil properties in 3D refers to the application of geostatistical methods to the soil data in order to produce maps of soil properties at different depths. Through two separate studies, this thesis elaborates on two different approaches for 3D soil mapping. At first, the well established Spline-Than-Krige approach for the mapping of soil pollutants atmospherically deposited from the copper smelting plant, was used. In the absence of the monitoring data, which can be used for a detailed characterization of the plume spreading process, this study was confined to the consideration of terrain exposure to explain spatial trend in arsenic distribution at different depths. This study aims to explore the extent to which the commonly available information, such as the prevailing wind direction, or the location of the source of pollution, in combination with the digital terrain model, can be used to quantify the terrain exposure, and hence to improve the spatial prediction of the arsenic concentration at several soil depths.

Next, the innovative geostatistical approach to 3D mapping of soil properties, based on soil profile data, was proposed. It provides the semi-automatic way for 3D modeling of soil variables, prediction over the regular grids (rasters) and also the evaluation of prediction accuracy. Methodologically, this approach operates within the 3D regression kriging framework. 3D trend model is conceptualized as hierarchical or non-hierarchical linear interaction model. This means that the model includes the interactions between the spatial covariates and depth in the hiearchial or non-hierarchial manner. The trend modeling is based on the application of the penalized regression technique, *lasso*. The lasso uses a specific regularization penalty in a fitting procedure to enable the efficient parameter estimation and variable selection (including interaction terms) at the same time. Special attention has been paid to accuracy assessment. The proposed approach implements the nested cross-validation procedure as a tool for the evaluation of the overall prediction accuracy. The obtained results show that taking the interaction into account can improve the predictive capabilities of the trend model up to 20%. As expected, the greatest improvement was achieved with variables that have a strong decreasing trend along the depth, as well as a higher variation in the surface soil layers. In addition, the inclusion of interactions between spatial covariates and depth has lead to models with the more sparse structure. The complete computational framework was implemented in the set of R functions, with the aim to constitute an R package (penint3D) for 3D soil mapping.

Key words: 3D soil maping, 3D regression kriging, Spline-Than-Krige, lasso, nested cross-validation, pollution assessment, topographic exposure.

Scientific area: Geodesy

Scientific sub-area: Modeling and Management in Geodesy

UDC number: 528:005(043.3)

UNIVERZITET U BEOGRADU

Rezime

Građevinski fakultet

Odsek za geodeziju i geoinformatiku

Geostatističko modeliranje geohemijskih promenljivih u 3D prostoru

Geostatističko kartiranje zemljišta u 3D odnosi se na primenu geostatističkih metoda na zemljišnim podacima u cilju izrade karata zemljišnih karakteristika jednog područja, koje se odnose na različite dubine zemljišta. U okviru dve nezavisne studije, ova doktorska disertacija razmatra dva različita pristupa geostatističkog modeliranja zemljišta u 3D. U okviru prve studije, "Spline-Than-Krige" metod je korišćen za kartiranje koncentracije arsena u zemljištu, u blizini Rudarsko-topioničarskog basena Bor, na tri različite dubine (0-5 cm, 5-15 cm i 15-30 cm). Dugogodišnje emitovanje neprečišćenih materija iz topionice rudnika u atmosferu, dovelo je do zagadjenja zemljišta u okolini, taloženjem štetnih materija nošenih vetrom. U odsustvu podataka kojima bi se detaljnije mogao opisati proces raspršivanja štetnih materija, ova studija se ograničila na analizu izloženosti terena uticaju vetra, a time i procesu zagađenja. Predstavljen je inovativan pristup kvantifikaciji izloženosti terena izvoru zagađenja. Na osnovu opšte dostupnih podataka, kreirano je nekoliko parametara kojima se kvantifikuje geometrijska i topografska izloženost svake tačke terena izvoru zagađenja. Tako kreirani parametri, iskorišćeni su za opisivanje prostornog trenda koncentracije arsena na tri različite dubine. Definisani trendovi, korišćeni su u okviru regresionog kriginga, za prostornu predikciju. Na taj način pokušalo se odgovoriti na pitanje, u kojoj meri, opšte dostupni podaci, kao što su pravac dominantnog vetra ili poznavanje tačne lokacije izvora zagadjenja u kombinaciji sa digitalnim modelom terena, mogu biti iskorišćeni da bi se unapredila preciznost prostorne predikcije zemljišnih zagadjivača, kako na površinskim slojevima tako i na većim dubinama.

U okviru druge studije, predstavljen je inovativni geostatistički pristup 3D kartiranju zemljišnih promenljivih. Metodološki, predloženi pristup je baziran na 3D regresionom krigingu. Model trenda je definisan linearnom funkcijom koja uključuje članove interakcije između površinskih promenljivih i dubine, po hijerarhijskom i nehijerarhijskom principu. Problem izbora modela i ocena parametara rešen je primenom *lasso* regularizacione regresije. Primenom *lasso* regresije omogućen je automatski izbor značajnih prediktora (uključujući i članove interakcije između površinskih pomoćnih promenljivih i dubine). U okviru ove studije preporučeno je korišćenje i način implementacije ugnježdene unakrsne validacije za ocenu preciznosti predikcije modela. Dobijeni rezultati su pokazali da se uvođenjem interakcija može unaprediti model i do 20%. Najznačajnija unapređenja dobijena su za promenljive sa izraženom varijacijom u gornjim slojevima zemljišta. Pored toga, uvođenje interakcija u model, rezultiralo je izborom modela koji uključuje manji broj pomoćnih promenljivih. Predloženi pristup implementiran je u okviru PenInt3D paketa funkcija razvijanih u R okruženju.

Ključne reči: 3D modeliranje zemljišta, 3D regresioni kriging, lasso, ugnježdena unakrsna validacija, procena zagađenosti, topografska izloženost

Naučna oblast: Geodezija

Uža naučna oblast: Modeliranje i menadzment u geodeziji

UDK broj: 528:005(043.3)

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Soil is one of the most important natural resources necessary for life on Earth. It can be defined as a surface layer of the Earth's crust, located between the lithosphere and the atmosphere, formed by the long-term influence of pedogenesis (Kisić, 2012). Soil is a multipurpose resource that can be used in many aspects of human activity, including: geology, agriculture, forestry, construction, and commercial use. Its relevance to a wide range of human activities makes soil even more vulnerable to damage.

Bearing in mind that soil is a non-renewable natural resource, there has been a lot of focus on the issue of soil degradation over the past few decades. As a result, soil protection and sustainable soil usage have become popular topics in the field. The exact causes and types of soil degradation are numerous and complex (Eswaran et al., 2001). Pollution by harmful elements is one of the most serious examples of soil degradation. Such pollution is typical for soils in the vicinity of industrial zones, especially for those under mining and ore processing impact. Mining and smelting activities are recognized as the most effective sources of pollution. It is not uncommon that harmful dust and fumes from smelting plants and waste incinerators are released into the atmosphere without processing. The greatest part of the emitted matter is deposited on the ground through wet and dry depositions, thereby significantly changing the soil's chemical compounds. In these cases, the content of harmful elements in the soil usually shows the typical patterns in both horizontal and vertical (in-depth) sense. Horizontally, the concentration of pollutants typically decreases as the distance from the source of pollution, or the distance from from the major wind direction increases. Regional topography and different land (usage) types may also influence contamination processes thereby making the patterns more complex. Vertically, the concentration of harmful elements typically decreases as the soil depth increases.

The assessment of soil pollution necessarily involves the creation of maps that delineate areas where the pollutants exceed the pre-specified allowable levels. It is inevitably preceded by numerous in-field and laboratory investigation surveys. Moreover, these maps should provide information about the depths of soil contamination. Depending on the remediation technologies, the volume of contaminated soil may also be required. Volumes can then be converted into treatment costs, which allows for the selection of the most cost-effective and applicable remediation technology.

Geostatistics is well established in solving these issues and provides a number of tools for the exploratory data analysis, spatial predictions, risk mapping and the simulation of possible realizations of spatial phenomena (Goovaerts et al., 1997; Goovaerts, 2001; Khalil et al., 2013; Komnitsas and Modis, 2006; Dayani and Mohammadi, 2010; Guastaldi and Del Frate, 2012; Tavares et al., 2008; García-Sánchez et al., 2010). Even if the characterization of soil in 3D is needed, geostatistics is typically used as a tool for horizontal data analysis and mapping. The requirements for the maps related to other (deeper) soil layers are commonly met by the modeling of each horizontal layer independently, i.e. each layer was modeled without considering the soil properties above or below. The most widely used approach for such mapping was proposed by Malone et al. (2009) later to be called the 'Spline-than-Krige' method by (Orton et al., 2016). This method implies the conversion of profile data into a continuous form by fitting a spline function to the profile data (Bishop et al., 1999), prior to fitting the 2D spatial prediction model. The final product is a suite of digital maps of soil properties relating to different soil depths. The drawbacks of this method are twofold: (1) the spline-converted data are estimates with associated errors, which, if used, ultimately create additional source of errors in the model (Hengl and Heuvelink, 2013); and (2) the independent mapping of different layers poses a risk that the maps would show illogical discrepancies when overlapping (Meirvenne et al., 2003).

For these aforementioned reasons, soil should be considered as a 3D body. Soil properties vary in each direction, and also in time. At some scale, these variations are also spatially auto-correlated and it makes sense to treat them with 3D geostatistics. The use of 3D geostatistics in soil science is relatively new and represents the logical continuation of geostatistical advances in soil mapping. Today, 3D soil mapping is recognized as one of the main methodological challenges facing the soil scientists community (Arrouays et al., 2014). Regarding that geostatistical methods do not differ meaningfully if the spatial phenomena are considered in 2D or in 3D, the key difficulties in the application of the 3D geostatistical methods can be caused by the very nature of the soil data, or specific soil properties. This is summarized by Hengl and Heuvelink (2013), as follows:

1. *The differences between sampling intervals and spatial correlation in the horizontal and vertical dimensions are very large. This results in strong anisotropy between the two directions that must be accounted. The estimation of the anisotropy may be hampered by a relatively small number of observations along the vertical profile.*

2. *Soil property values refer to vertical block support (usually because they are composite samples, i.e. the average over a soil horizon), hence some of the local variation (in the vertical dimension) has been smoothed out.*

3. *Soil surveyors systematically under-represent lower depths - surveyors tend to systematically take fewer samples as they assume that deeper horizons are of less importance for management, or because deeper horizons are more expensive to collect, or because deeper horizons are assumed to be more homogeneous and uniform.*

4. *Many soil properties show clear trends along the vertical dimension. It may not be that easy to incorporate a vertical trend because such a trend is generally not consistently similar between different soil types.* In addition, the lack of environmental covariates known in 3D space largely limits the development of 3D spatial models of soil property.

This research is primarily committed to solving the last problem. However, the other issues will also be partially addressed.

1.2 Problem statement

For over more than a hundred years of activity, the exploitation and processing of copper ore in the mining and metallurgical complex Bor, Serbia, caused serious environmental problems (Kovačević et al., 2010; Serbula et al., 2013, 2014). This was mainly due to inefficient control and refinement of toxic fumes during the smelting process. Harmful compounds released in the atmosphere have spread over the surrounding area and changed the soil's geochemistry. A field survey was conducted in 2006 to document the actual state of soil in the vicinity. The survey included the opening of 205 soil profiles that were randomly distributed over the area of 200 km^2 and were spaced 10 km away from the mining complex; see Section 3.1. This area was also selected due to its high potential for further mining investigation. Preliminary data analysis indicated that the soil was indeed affected by a long term atmospheric pollution processes. This revealed the three-dimensional, non-stationary pollution problem with complex spatial patterns that can be connected with many external factors, such as prevailing climatic conditions, soil types, topography, etc.

Mapping such phenomena by geostatistical methods implies the inclusion of external factors into the geostatistical model. Even more, the exclusion of these factors may result in a misleading geostatistical model. However, the incorporation of these into a 3D geostatistical model may reveal new challenges. Considering that the external environmental influences mostly affect the upper soil layers, and that their effect decreases with soil depth, it may be expected that many of the soil characteristics will show a clear trend along the soil depth. Furthermore, it may also be expected that the vertical trend varies spatially due to different soil characteristics across the area. Therefore, the key question of this research is how these influences can be properly approximated and incorporated into a 3D geostatistical model to improve the prediction at any location in a 3D space. A possible solution could be to make a 3D interaction model, i.e. the model that includes the interactions between the environmental factors and soil depth. However, the inclusion of interactions will dramatically increase the number of covariates that should be considered, which imposes the problem of model selection. Another important issue is related to the modeling of spatial correlation structure. The anisotropic correlation model must be found, that includes the anisotropy between the vertical and the horizontal direction, which is a prerequisite for the application of 3D geostatistical methods on soil measurements.

1.3 Objectives

The final objective of this research is to propose an innovative approach for the 3D mapping of soil properties, which combines the advantages of interaction models and 3D geostatistics. This approach would be particularly suitable for soils and soil properties affected by intensive external environmental factors or human or industrial activities. Considering the characteristics of the case study, and the methodological challenges, the specific objectives can be formulated as follows:

1. *To examine how case-specific environmental conditions, like exposure to the source of pollution, can be quantified and mapped based on the limited amount of commonly available information, such as terrain topography, prevailing climatic conditions and spatial relations.*

2. *To determine the contribution of such case-specific environmental layers (maps) to the mapping of pollutants at different soil depths.*

3. *To examine how the important interactions can be automatically recognized and included in a linear 3D trend model.*

4. *To examine the advantages and the disadvantages of the inclusion of the interactions between spatial covariates and depth within the linear 3D model of soil variables.*

5. *To analyze and model the dependency structure of trend residuals in 3D.*

1.4 Approach

Methodologically, the approach relies upon point scale geostatistics. Regression kriging is adopted as a general statistical framework for spatial prediction. Regression kriging is a two-step approach that combines two conceptually different techniques: regression for trend estimation and simple or ordinary kriging to interpolate residuals (Hengl et al., 2007; Bajat et al., 2013). In this study, trend modeling is based on linear regression. The restrictive nature of linear models is relaxed by considering interactions between predictors and their functional expansion in a polynomial form. Environmental factors are represented by a set of one or more continuous or categorical variables known as spatial covariates. In Chapter 4, regression kriging was used within the so called Spline-Than-Krige approach (Malone et al., 2009) (see Section 2.5.2) to map atmospherically deposited arsenic concentration at several soil depths. In the Chapter 5, penalized regression method *'lasso'* (Tibshirani, 1996) and its extension for hierarchical interaction models proposed by Bien et al. (2013) (see sections 2.2.7.1 and 2.2.7.2) were used to optimize the 3D interaction trend model. Subsequently, it was incorporated into the generic framework for 3D soil mapping. A new approach for model accuracy assessment is also an integral part of this framework. Nested n-fold cross-validation was proposed to perform model assessment that preserves the basic principle of predictive modeling, which states that the modeling process has to be completely separated from the validation process.

1.5 Outline

The dissertation comprises 6 chapters, out of which one is submitted and one is prepared for submission to peer-reviewed ISCI journals. Each chapter is arranged as introduction, methodology, results and conclusion.

Chapter 1 offers a brief overview and objectives of the dissertation. It includes a general introduction and the motivation for the research work, the research scope and specific objectives, the applied approach and the thesis outline. **Chapter 2** presents the main theoretical concepts and methods used in this thesis. It begins with the concept of soil forming process as a foundation for quantitative soil modeling. The basic concepts of predictive statistical modeling, including the theory of linear regression and shrinkage regression methods are included. The main theoretical aspects of geostatistics that are followed by specific methods used in this study are also provided. At the end of this chapter, the extension of 2D geostatistical methods to a 3D space was presented. **Chapter 3** provides details on a case study and the data used for this research. **Chapter 4** presents the study of layer-specific mapping of arsenic concentration that was atmospherically deposited from the Bor Copper Mining Complex. The presented approach considers the effects of the prevailing climatic conditions and local topography on the terrain exposure to the dispersion of pollutants. Several exposure parameters were created and employed as spatial covariates within the 'Spline-Then-Krige' approach. **Chapter 5** presents the usage of the shrinkage regression method Lasso for building the 3D interaction linear trend models of soil properties. The obtained models were further used as a part of 3D regression kriging for the interpolation of soil properties over the whole 3D prediction domain. **Chapter 6** describes the R package PenInt3D, which is still under development, for the prediction of soil properties by penalized interaction models. **Chapter 7** gives a short summary of the most important conclusions.

Chapter 2

The main concepts and methods

This Chapter presents the main theoretical concepts and methods used in this thesis. The specific topics include: (1) A conceptual model of pedogenesis that provides the theoretical basis for quantitative analysis and mapping of soil properties; (2) Principles of predictive statistical modeling and linear regression techniques; (3) The theory of regionalized random variable and variography; (4) Basic geostatistical methods; (5) A universal model of soil variation and hybrid techniques; (6) An extension of geostatistical techniques in 3D space.

2.1 The Concept of Soil Formation-CLORPT model

Soil is a very complex system where a variety of physical, biological, and chemical processes interact. The understanding of how their joint influence affect the long-age process of pedogenesis has always been a challenge facing soil scientists. Initially, a number of conceptual models were formulated (Stockmann et al., 2011). The most well-known model of soil formation is Jenny's (Jenny, 1941) state-factor model, also known as *clorpt* model. It conceptualizes the state of soil as a resultant of joint influences of five main independent factors and a number of additional, unspecified factors:

$$S = f(cl, o, r, p, t, ...) \tag{2.1}$$

where cl is the climate, o are the organisms, r is the topography, p is the parent material, and t is the time, and ... stands for additional, unspecified factors. Such formulation has provided an intuitive framework for much of the subsequent work on solving the function f. Most efforts were spent not to formulate the overall equation f but rather to examine the individual contribution of each factor. In that regard, empirical methods have mostly been employed in literature. This approach involves the examination of soil behavior in situations where one factor is allowed to vary while others are kept constant. Such treatments led to the development of empirical models, known as climo-functions, bio-functions, topo-functions, litho-functions, and chrono-functions (Yaalon, 1975).

2.1.1 SCORPAN framework

The conceptual model published by Jenny has served as a foundation for further investigations on quantitative relations between soil and soil forming factors. A variety of international researchers have sought the way to construct mathematical solutions that would represent the closest approximation of joint influences of soil forming factors (McBratney et al., 2000; Minasny et al., 2008).

With the introduction of GIS and digital terrain analysis, new opportunities for soil scientists have arisen. Digital Elevation Model (DEM) along with other digital layers have provided a detailed quantitative description of the area, thus opening the possibility to extend the *clorpt* concept to the spatial domain. Consequently, soil scientists all over the world have begun to use increasingly mapped auxiliary variables to explain the specific spatial patterns of soil and hence to produce maps of specific soil properties. Standard multiple linear regression was used to model the relationship between soil data and terrain attributes (Moore et al., 1993; Gessler et al., 1995). This approach was later termed as the "environmental correlation" method (McKenzie and Ryan, 1999), or the spatial prediction by multiple regression with auxiliary variables (Odeha et al., 1994).

Following up on this trend, McBratney et al. (2003) utilized the Jenny's *clorpt* concept to propose a more generic framework called the *scorpan* model, which primarily aimed at providing empirical quantitative descriptions of relationships between soil and other spatially referenced factors. The *scorpan* model states that the soil type or soil attribute at an unvisited site can be predicted from a numerical function or model (f) of the environmental factors plus the locally varying, spatial dependent residuals (ε):

$$S_c = f(s,c,o,r,p,a,n) + \varepsilon \qquad or \qquad S_a = f(s,c,o,r,p,a,n) + \varepsilon \tag{2.2}$$

where S_c and S_a represent soil classes and soil attribute respectively. The environmental factors within the acronym *scorpan* are: s: soil, other properties of the soil at a point; c: climate, climatic properties of the environment at a point; o: organisms, vegetation or fauna or human activity; r: relief, topography or landscape attributes; p: parent material, lithology; a: age, the time factor; n: space, spatial position.

Mathematical model of f is the empirical quantitative function linking the soil variable (S) to the *scorpan* factors. Each factor can be represented by a set of one or more continuous or categorical variables. For example, r can be represented by DEM but also with the various DEM derivates such as *slope*, *curvature* etc. Various data layers can be used to describe the *scorpan* factors. Today, the creation of these layers is seen as an integral part of any digital soil mapping study.

2.2 Predictive soil mapping - linear regression approach

Scull et al. (2003) defined the ’predictive soil mapping" (PSM) as the *development of numerical or statistical model of a relationship among environmental variables and soil properties, which is then applied to a geographic data base to create a predictive map*. Today, so defined PSM is just an inevitable part of broader concept called Digital Soil Mapping (DSM) (McBratney et al., 2003; Minasny and McBratney, 2016). Digital soil mapping is defined as: *the creation and population of spatial soil information systems by the use of field and laboratory observational methods coupled with spatial and non-spatial soil inference systems* (Lagacherie and Mcbratney, 2007).

Jenny’s model and `scorpan` framework have formed the theoretical basis for using a variety of statistical methods in predicting soil properties based on auxiliary spatially referenced data. Advances in mathematical and statistical theory (including machine learning techniques) have created a great potential for improvements in predictive soil mapping. In statistical theory, the term "predictive modeling" refers to the data-driven process of building a statistical model, which ideally should provide the best possible prediction. It means that the statistical model must be determined in a way to minimize the *prediction error*. The prediction error refers to the average error that results from using a statistical model to predict the soil variable on data that has not been used in the model building process (test data). Prediction error is also known as *test error*. On the other hand, the *training error* can be calculated by applying the statistical method to observations used in its training (*training data*). Training error is often quite different from test error. Various statistical methods can be used for this purpose; however, no single method has been proven dominant when examined using all possible data sets. An exhaustive review of recent achievements using this approach was provided by McBratney et al. (2003); Malone et al. (2016).

In this thesis, the statistical approach used for solving the *scorpan* problem is based on linear regression methods. In linear regression, the model has a vector of parameters set up to minimize the training error. The potential disadvantage of linear regression models is that the obtained model usually does not match the true unknown form of f very well. Alternatively, the function f can be approximated by more flexible models (like tree-based or neural network models) that can fit many different possible functional forms for f. In linear regression, fitting a more flexible model requires estimating a greater number of parameters. These more complex models can lead to *overfitting*, which essentially means that they follow the observed data too closely.

The following sections provide a brief overview of the statistical methods and fundamental principles of predictive modeling that are used in this study. Considering that linear regression is adopted as a general modeling framework in this thesis, a brief concept of linear regression, their extensions and a review of modern approaches in linear modeling, will be presented. For more details, the interested reader shall be refered to Hastie et al. (2009); James et al. (2013); Perović (2005); Kuhn (2008) which are used as a guide when presenting statistical methods.

2.2.1 Quantitative Measures of Model Performance

In order to evaluate the performance of a statistical model, a measure must be defined that can quantify how well predictions match the observed data. In the regression setting, when the outcome is a number, the most commonly-used measure is the mean squared error (MSE) given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2 \tag{2.3}$$

where y_i is the $i-th$ observation and $\hat{f}(x_i)$ is the prediction that $\hat{f}$ gives for data point x_i. The MSE becomes a smaller value as the the predicted values approach the observations, i.e if the residuals tend to be small. By squaring the residuals, MSE becomes more sensitive to outliers, as the larger residuals contribute more to the final estimate than the smaller residuals. Often, a more suitable measure is the root mean squared error (RMSE, Equation 2.4) which is derived from MSE by taking the square root of the MSE so that it is in the same units as the original data. The RMSE can be interpreted as the average distance between the observed values and the model predictions.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2} \tag{2.4}$$

Another common measure is the coefficient of determination, which is commonly denoted as R^2. R^2 value is a number that indicates how the fit of proposed model is better than the fit of the simple mean model. The mean model gives the observed mean value for every predicted value and generally it would be used if there were not any useful predictors. R^2 value takes the form of a proportion and therefore assumes a value between 0 and 1:

$$
\begin{aligned}
R^2 &= \frac{TSS - RSS}{TSS} = 1 - \frac{RSS}{TSS} \\
&\text{where :} \\
TSS &= \sum_{i=1}^{n} (y_i - \bar{y})^2 \\
RSS &= \sum_{i=1}^{n} (y_i - \hat{y}_i)^2
\end{aligned}
\tag{2.5}
$$

As it can be seen from Figure 2.5, the estimation of R^2 is based on two sums of squares, TSS and RSS. TSS measures how far the observed data are from the mean value and can be thought of as the amount of variability that is left after fitting the mean model. On the other hand, the RSS reflects the amount of variability that is left after fitting the proposed model. Hence, the difference between TSS and RSS reflects the improvement in prediction reached by fitting the proposed model when compared to the mean model. Dividing that difference by RSS provides the R^2 value.

Model selection, which will be discussed later in this chapter, implies the consideration of several models with different subgroups of predictors. If we assess the quality of these models by comparing training error (i.e. training RMSE), it is very likely to be shown that the smallest error is provided by fitting the model with the largest number of predictors. For that reason, the training RMSE or training R^2 value can not be used to rank models that have different numbers of predictors. However, there are several measures that can penalize the model performance based on how many predictors are used in the model. For linear regression, a commonly used statistic is the Akaike Information Criterion (Akaike, 1974):

$$
AIC = n \log \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2 + 2P \tag{2.6}
$$

where P is the number of terms in the model. The first term of the Equation 2.6 decreases as more variables are added to the model, whereas the second term increases. In this way, AIC controls for overfitting by penalizing models that include too many variables.

The adjusted R^2 statistic is another popular measure for model selection. Since RSS in Equation 2.5 always decreases as more variables are added to the model, the R^2 always increases as more variables are added. For a model with d variables, the $Adjusted\ R^2$ statistic is calculated as:

$$Adjusted\ \ R^2 = 1 - \frac{RSS/(n-d-1)}{TSS/(n-1)} \tag{2.7}$$

Unlike the AIC, a large value of $adjusted\ R^2$ indicates a model with a small error.

2.2.2 Resampling methods

Today, the application of any modern regression techniques in predictive modeling cannot be imagined without the extensive use of resampling methods. Resampling methods involve repeatedly fitting the same statistical method using different subsets of training data in order to obtain additional information about the fitted model. This information may relate to the optimal subset of predictors, which aids in *model selection*, or eventually to the predictive accuracy of the model, which is referred to as *model assessment*. Resampling methods were devised to compensate for the lack of sufficiently large test sets that can be directly used for the estimation of test error.

2.2.2.1 Cross-validation

Cross-validation involves splitting the data up into a set of K parts (folds) of approximately equal size. In each step of the process, one fold is treated as the test data set (validation data set) while the remaining folds, joined in one group, are treated as the training data set. Also, in each step, cross-validation uses the training data set to fit the model and the test data to compute the prediction error (Figure 2.1). By doing so, the K estimates of prediction error can be combined to obtain the average prediction error. For example, if each step of cross-validation results in the test $MSE_k, \quad k = 1, 2, \ldots, K$, the average cross-validation MSE is:

$$MSE_{cv} = \frac{1}{K}\sum_{k=1}^{K} MSE_k \tag{2.8}$$

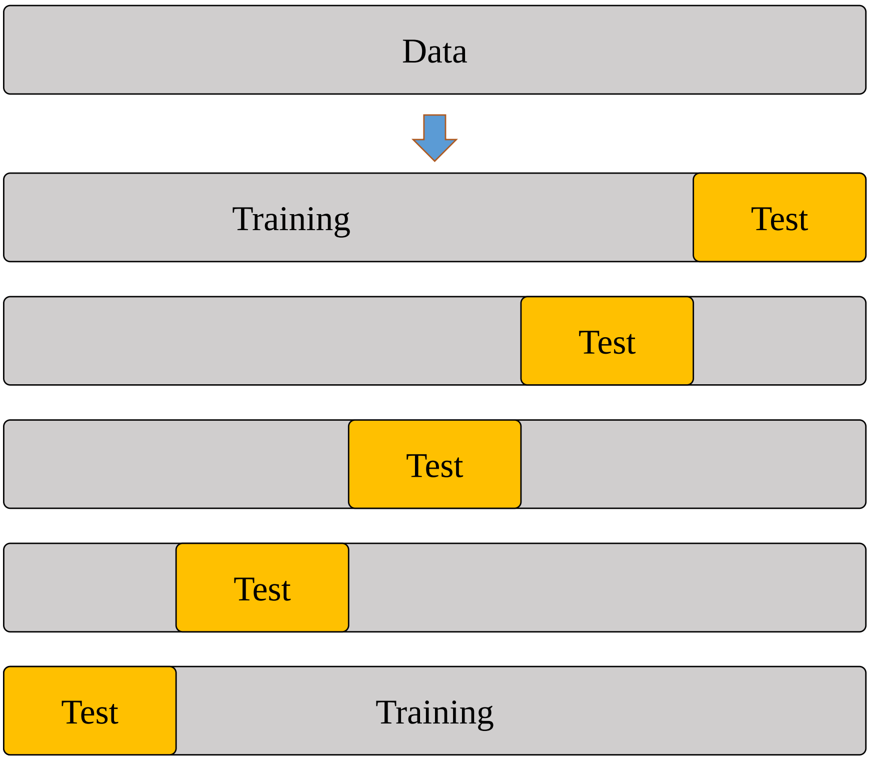


FIGURE 2.1: 5-fold data partitioning

In predictive modeling, for methods with meta-parameters (e.g. shrinkage parameter λ for lasso, see Section 2.2.7), cross-validation is often used as a tool for selecting the optimal meta-parameter. For example, for a model M that depends on meta-parameter θ, a cross-validation error can be computed for a whole set of meta-parameter values $\theta \in \Theta$, which are set previously. This results in cross-validation error curve which relates the cross-validation error to the values of θ. The optimal θ is:

$$\hat{\theta} = \underset{\theta \in \{\theta_1, \theta_2, \ldots, \theta_m\}}{\arg\min} \; CV(\theta) \tag{2.9}$$

Cross-validation procedure in model selection is given in Algorithm 1.

Algorithm 1 Selection of the best value for meta-parameters based on cross-validation procedure

```
Partition $D$ into stratified sets $D_i$, $k = 1,\ldots,K$ of approximately equal size
for $k = 1$ to $K$ do
    Let $D'$ be $D \setminus D_i$
    for each $\theta \in \Theta$ do
        Fit the model $M(\theta, D')$
        Make predictions by $M(\theta, D')$ on $D_i$
    end for
    For each parameter $\theta$ compute the average error $MSE_{cv}(\theta_i) = \frac{1}{K}\sum_{k=1}^{K} MSE_k$
    Let $\theta^*$ be $\arg\min_{\theta \in \Theta} MSE_{cv}(\theta)$
end for
```

When $K = n$, we call this *leave-one-out* cross-validation, because we leave out one data point at a time.

2.2.2.2 Nested cross-validation

The cross-validation procedure, as explained above, provides a biased estimate of accuracy parameters for methods that require the optimization of meta-parameters. Choosing meta-parameters is also a part of the training process, and, since the whole data set was used in cross-validation to select the best value of meta-parameters, the whole data set is used for the training (Krstajic et al., 2014). This procedure violates the fundamental requirement of predictive modeling in that the training and test data need to be separated. The use of the nested cross-validation technique can overcome the limitations described above (Krstajic et al., 2014). In short, the nested cross-validation consists of two nested cross-validation loops. The outer loop serves to assess the performance of the model, which was selected in the inner cross-validation loop. For each outer fold, the model is selected on each outer training set, using a standard cross-validation procedure in the inner loop and then applied to the outer test set. The process yields a prediction for each fold, obtained from a model which was not trained on that fold. Using these predictions, the overall accuracy measure is computed. The nested cross-validation procedure is presented in Algorithm 2 and in Figure 2.2.

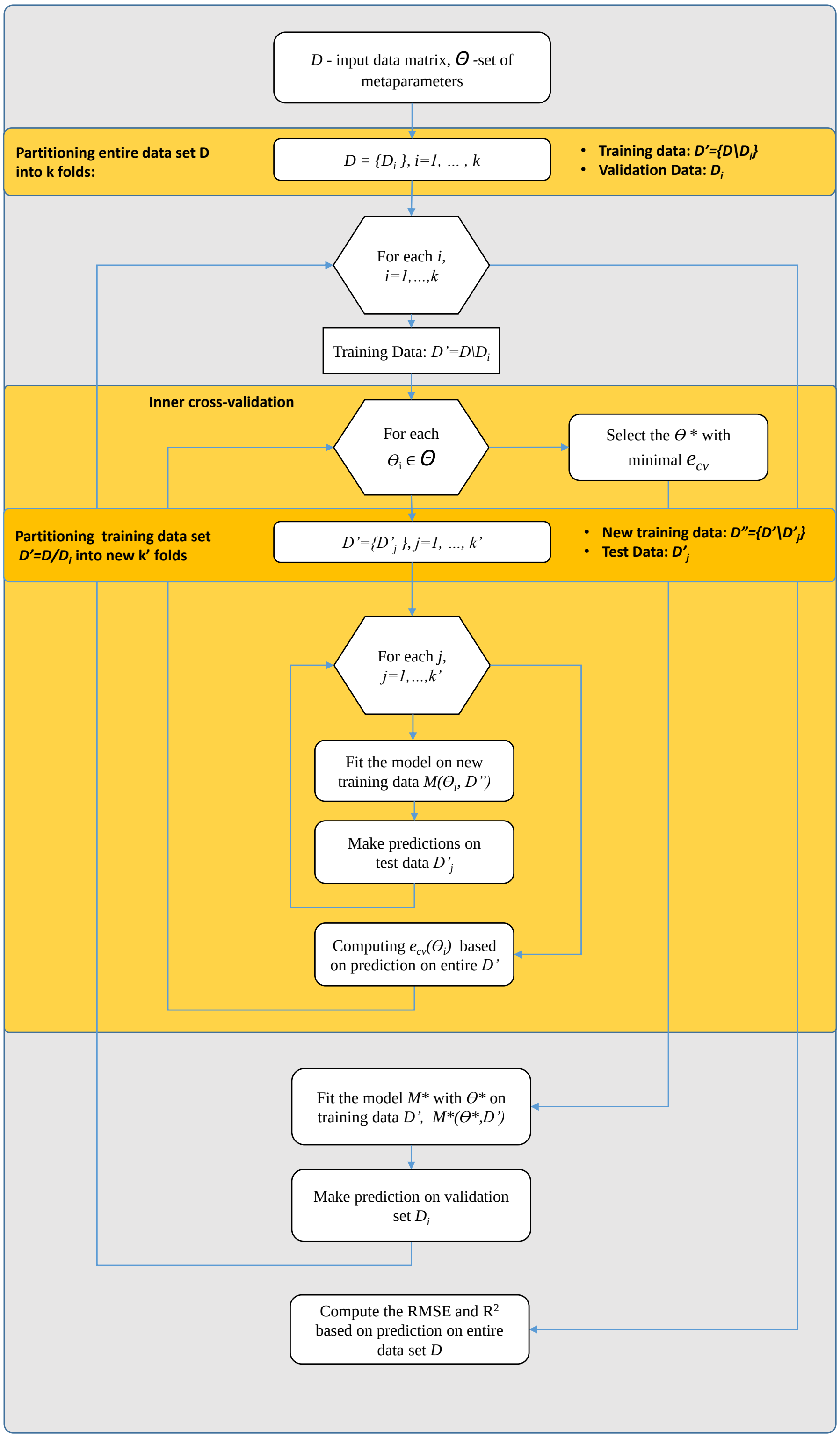

D - input data matrix, Θ -set of metaparameters
Partitioning entire data set D into k folds:
$D = \{D_i\}, i=1, \dots, k$
Training data: $D'=\{D \backslash D_i\}$
Validation Data: D_i
For each i, $i=1,\dots,k$
Training Data: $D'=D \backslash D_i$
Inner cross-validation
For each $\Theta_i \in \Theta$
Select the Θ * with minimal e_{cv}
Partitioning training data set $D'=D/D_i$ into new k' folds
$D'=\{D'_j\}, j=1, \dots, k'$
New training data: $D''=\{D' \backslash D'_j\}$
Test Data: D'_j
For each j, $j=1,\dots,k'$
Fit the model on new training data $M(\Theta_i, D'')$
Make predictions on test data D'_j
Computing $e_{cv}(\Theta_i)$ based on prediction on entire D'
Fit the model M^* with Θ^* on training data D', $M^*(\Theta^*,D')$
Make prediction on validation set D_i
Compute the RMSE and R^2 based on prediction on entire data set D

FIGURE 2.2: Model assessment based on nested cross-validation procedure.

Algorithm 2 Nested cross-validation

```
Partition $D$ into stratified sets $D_i$, $i = 1, \dots, k$ of approximately equal size
for $i = 1$ to $k$ do
    Let $D'$ be $D \setminus D_i$
    for each $\theta \in \Theta$ do
        Partition $D'$ into stratified sets $D'_j$, $j = 1, \dots, k'$ of approximately equal size
        for $j = 1$ to $k'$ do
            Fit the model $M(\theta, D' \setminus D'_j)$
            Make predictions by $M(\theta, D' \setminus D'_j)$ on $D'_j$
        end for
        Compute error $e_\theta$ based on predicted and real target values on $D'$
    end for
    Let $\theta^*$ be $\arg\min_{\theta \in \Theta} e_\theta$
    Fit the model $M(\theta^*, D')$
    Make predictions by $M(\theta^*, D')$ on $D_i$
end for
Report error computed on predicted and real target values on $D$
```

2.2.3 Bias-Variance trade off

The understanding of the concept of bias-variance trade off is particularly important for any type of statistical predictive modeling. Bias-variance explains how different sources of error influence the overall accuracy of the model. The expected test error for a particular test point x_0 can be decomposed into the sum of three fundamental quantities, the variance of $\hat{f}(x)$, the squared bias of $\hat{f}(x)$, and the variance of the error variance terms ε:

$$E(y_0 - \hat{f}(x_0))^2 = Var(\hat{f}(x_0)) + [Bias(\hat{f}(x_0)]^2 + Var(\varepsilon) \tag{2.10}$$

where $E(y_0 - \hat{f}(x_0))^2$ is the expected test error, and refers to the average test error that would be obtained if f is repeatedly estimated using a large number of training sets and tested each at test point x_0. The equation 2.10 shows that the expected test error reaches a minimum only if the predictive model achieves the lowest bias and variance possible.

The variance refers to the error caused by fitting the regression model by using many different data sets. Since different data sets are used to fit the model, the predictions for a given point vary between different realizations of the model. Ideally, the $\hat{f}$ should not vary significantly between different data sets. The bias refers to the error that is introduced by modeling the true f by a particular model $\hat{f}$. For example, if the true f is non-linear, and we are trying to fit the linear model, an irreducible error is introduced. This error is known as *bias*.

Generally, more flexible methods (like spline or tree-based methods) result in models with less bias, but with high variance. Figure 2.3 also shows the typical relationship between the training and test error, as the complexity of model varies. The training error monotonically decreases as the complexity of the model increases, whereas the test error decreases as the model reaches a certain complexity. As a result, the test error tends to increase due to the increasing variance. For example, linear models becomes more flexible as more variables are included in the model. Therefore, a key task in linear modeling is to determine which subset of variables should be included in order to provide balance between bias and variance. The relationship between bias, variance, and test error is referred to as the bias-variance trade-off. In reality, the true f is not known, so it is generally not possible to determine how much the adopted model deviates from the true f. Accordingly, bias-variance trade off is not a rule, but it is rather a kind of a problem which should always be kept in mind when modeling.

2.2.4 Linear regression

A linear regression model is a very straightforward approach for predicting a quantitative response Y on the basis of a group of predictor variables (predictors) X_j. It assumes that there is approximately a linear relationship between Y and X_j, i.e. each variable is linearly related to the modeled variable. Mathematically, the linear regression model can written in the form:

$$f(X) = \beta_0 + \sum_{j=1}^{p} X_j\beta_j + \varepsilon \tag{2.11}$$

where the β_j's are unknown parameters or coefficients, and the X_j can be (Hastie et al., 2009):

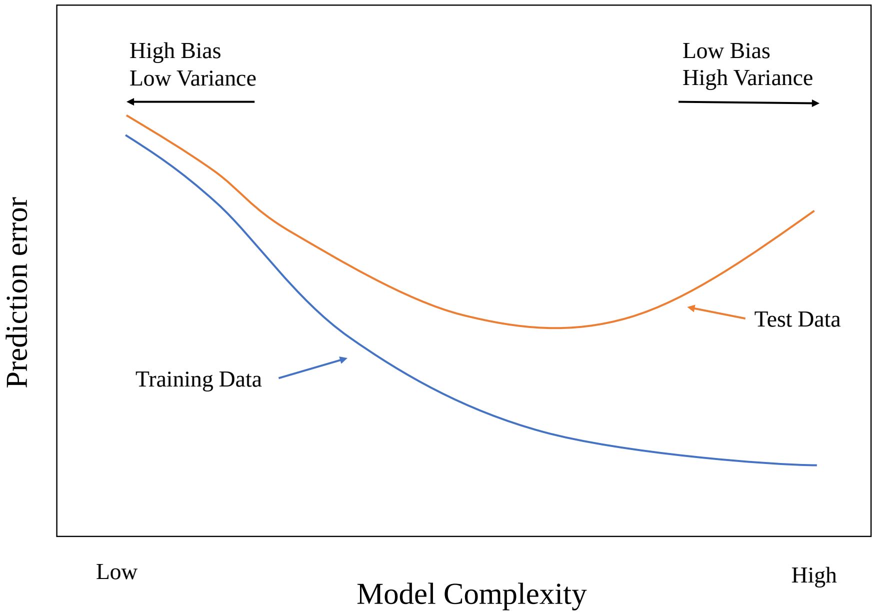


FIGURE 2.3: Bias-Variance trade-off, from (Hastie et al., 2009)

1. quantitive inputs
2. transformation of quantitive inputs, such as log, square-root
3. basis expansions of inputs, like polynomial function of particular inputs
4. numeric or 'dummy' coding of the levels of qualitative inputs.
5. interactions between variables, for example, $X_i = X_j X_k$

The observed data from which the coefficients β have to be estimated are typically given in the form of pairs: $(x_1, y_1, x_2, y_2, \ldots, x_N, y_N)$. Where each $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^T$ is a vector of $i-th$ predictor variable measurements. According to the least square estimation, which is the most popular method for estimation, the coefficients $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ are determined to minimize the residual sum-of-squares (RSS):

$$\begin{aligned} RSS(\beta) &= \sum_{i=1}^{N} (y_i - \hat{f}(x_i))^2 \\ &= \sum_{i=1}^{N} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \end{aligned} \tag{2.12}$$

If the $N \times (p+1)$ matrix, where each row represents the input vector, is denoted by $\mathbf{X}$, and if the N-vector of outputs in the data is denoted by $\mathbf{y}$, then the residual sum-of-squares can be written as:

$$RSS(\beta) = (\mathbf{y} - \mathbf{X}\beta)^{\mathbf{T}}(\mathbf{y} - \mathbf{X}\beta) \tag{2.13}$$

If $\mathbf{X}$ has full column rank, and hence $\mathbf{X^T X}$ is a positive definite, and the first derivative is set to zero $\mathbf{X^T}(\mathbf{y} - \mathbf{X}\beta) = \mathbf{0}$, the unique solution for $\beta = (\beta_{\mathbf{0}}, \beta_{\mathbf{1}}, \ldots, \beta_{\mathbf{p}})^{\mathbf{T}}$ will be given by:

$$(\hat{\beta}) = (\mathbf{X^T X})^{-\mathbf{1}}\mathbf{X^T y} \tag{2.14}$$

2.2.5 Extensions of linear models

The standard linear regression model provides an interpretable model form. However, it makes a set of restrictive assumptions that can be rarely encountered in practice. Two highly restrictive assumptions state that the relationship between the predictors and the response variable must be additive and linear.

2.2.5.1 Inclusion of interactions

One way to relax the additive assumption is to extend the linear model by allowing for *interaction* effects. Interactions exist when a change in the level of one variable has different effects on the response, depending on the value of the other variable. An interaction effect is an additional term in model setting that is constructed by computing the product of two variables. Accordingly, the linear model with interaction terms has the following form:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_1 X_2 + \varepsilon \tag{2.15}$$

The effects of interactions can be distinguished if we reformulate the Equation 2.15 as:

$$\begin{aligned} Y &= \beta_0 + (\beta_1 + \beta_3 X_2) X_1 + \beta_2 X_2 + \varepsilon \\ &= \beta_0 + \tilde{\beta_1} X_1 + \beta_2 X_2 \end{aligned} \tag{2.16}$$

where $\tilde{\beta_1} = \beta_1 + \beta_3 X_2$. Since $\tilde{\beta_1}$ changes concomitantly with X_2, the effect of X_1 on Y is no longer constant. Changing in X_2 will change the impact of X_1 on Y. Therefore, interaction models distinguishes two types of effects: *main effects*, which are the individual effect of single variable and the *interaction effects*, or the *synergy effects* of two linked variables.

2.2.5.2 Polynomial expansion

As mentioned previously, the linear regression model assumes a linear relationship between the response and its predictors. However, in reality, the true relationship between predictors and the response is often nonlinear. A simple way to address this issue with linear models is to use a *polynomial regression*. Polynomial regression involves the polynomial expansion of predictors, i.e. includes the polynomial functions of predictors within the linear regression model. In other words, polynomial regression extends the linear model by adding extra predictors, obtained by raising each of the original predictors to a power (James et al., 2013). For example, a cubic regression uses three variables, X, X^2, and X^3, as predictors, which results in the model:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_3^3 + \ldots + \beta_d x_i^d + \varepsilon_i \tag{2.17}$$

The coefficients in Figure 2.17 can be easily estimated using least squares linear regression because this is still a standard linear model. Even though this is a common linear regression model, the individual coefficients are not of particular interest. Instead, the attention should rather be payed on the entire fitted ’function’ that corresponds to the one variable.

2.2.6 Model selection

Model selection is one of the most frequently encountered problems in statistical data analysis. Generally, it involves the task of selecting an optimal statistical model from a set of candidate models. In predictive statistical modeling, model selection should provide a balance between model complexity and its ability to predict. Complex models fit training data better, but they are more prone to overfitting and lead to lower quality predictions. In linear regression modeling, with a large number of predictors, smaller subsets that exhibit the strongest effects are preferred.

2.2.6.1 Best subset selection

Best subset selection involves the separate-fitting of models consisting of each possible combination of p predictors, and choosing the one with the smallest test error. This is usually done through the following algorithm:

Algorithm 3 Best subset selection, from James et al. (2013)

```
Let $\mu_0$ denote the model which contains no predictors (null model).
for $k = 1, 2, \ldots, p$: do
    Fit all $\binom{p}{k}$ models that contain exactly $k$ predictors.
    Pick the best among these $\binom{p}{k}$ models, and call it $\mu_k$. Best is defined as having the smallest $RSS$, equivalently largest $R^2$.
end for
Select a single best model from among $\mu_0, \mu_1, \ldots, \mu_p$ using cross-validated prediction error, $AIC$ or $Adjusted R^2$.
```

Best subset selection is indeed simple and a very conceptually appealing approach but, on the other hand, it is very computationally demanding. For any subset of p predictors there are 2^p models that must be considered. Consequently, the number of possible models that must be considered increases dramatically as p increases.

2.2.6.2 Forward step-wise selection

Forward stepwise selection is a popular algorithm for considering a sequence of nested linear regression models. It begins with a model with no predictors, sequentially adding one predictor at a time until the model with all predictors is fitted. In each step, the predictor that gives the greatest additional improvement to the fit is added to the model. This approach appears to be a very appealing alternative to the best subset selection because it considers a much smaller set of models. The forward stepwise selection procedure is given in Algorithm 4.

Algorithm 4 Foreward step-wise selection, from James et al. (2013)

```
Let $\mu_0$ denote the model which contains no predictors (null model).
for $k = 1, 2, \dots, p-1$: do
    Consider all $p-k$ models that augment the predictors in $\mu_k$ with one additional predictor.
    Pick the best among these $p-k$ models, and call it $\mu_{k+1}$. Best is defined as having the smallest $RSS$, equivalently largest $R^2$.
end for
Select a single best model from among $\mu_0, \mu_1, \dots, \mu_p$ using cross-validated prediction error, $AIC$ or $Adjusted R^2$.
```

2.2.6.3 Backward step-wise selection

Backward step-wise selection is very similar to forward stepwise selection. Backward step-wise selection also provides an efficient alternative to the best subset selection. However, unlike forward step-wise selection, it begins with the full least squares model containing all p predictors and then iteratively removes the least useful predictor one-at-a-time James et al. (2013). The backward stepwise selection procedure is given in Algorithm 5.

Algorithm 5 Backward step-wise selection, from James et al. (2013)

```
Let $\mu_0$ denote the model which contains all $p$ predictors (full model).
for $k = p, p-1, \ldots, 1$: do
    Consider all $k$ models that contain all but one of the predictors in $\mu_k$ for a total of $k-1$ predictors.
    Pick the best among these $k$ models, and call it $\mu_{k-1}$. Best is defined as having the smallest $RSS$, equivalently largest $R^2$.
end for
Select a single best model from among $\mu_0, \mu_1, \ldots, \mu_p$ using cross-validated prediction error, $AIC$ or $Adjusted R^2$.
```

Backward selection has just one important requirement: the number of samples n must be larger than the number of variables p, considering that the algorithm starts from the full-model. In contrast, forward step-wise selection can be used even when $n < p$, and so it is one of the viable subset methods when p is very large.

2.2.7 Shrinkage Methods

According to Hastie et al. (2009, 2015) there are two main problems with the least squares estimation:

1. The first problem is *prediction accuracy*. Least squares estimates for a model with a large number of predictors often have low bias, but very large variance. In order to

reduce the variance, a little bit of bias must be introduced. This can be accomplished by shrinking or setting certain coefficients to zero.

2. *The second problem is interpretation. Least squares fitting yields models that retain all predictors of greater or smaller importance. The inclusion of irrelevant predictors introduces unnecessary complexity into the model. A large number of predictors make model interpretation difficult. It is not rare that a certain number of predictors are in fact not associated with the modeled variable.*

The possible solutions to overcome these issues are the subset selection or step-wise selection procedures described previously. The main task of these techniques is to provide a model with a limited subset of relevant predictors, which would result in a reduction of variance and also in simpler model interpretation. However, due to their repetitive nature, a large number of potentially useful predictors can make this task very computationally demanding. The easiest and the most effective solution for this problem is to use the Shrinkage methods. Shrinkage methods fit the model containing all p predictors, by using one of the common loss functions (e.g. square loss) extended with additional regularization penalties that shrink the coefficient estimates towards (or exactly to) zero. The two most popular shrinkage (penalized) methods are *ridge regression* and *lasso* (least absolute shrinkage and selection operator).

The rationale behind the efficiency of shrinkage methods lies in bias-variance trade off. By shrinking the coefficients towards zero, the flexibility of the model decreases, leading to an increased bias, but a decreased variance of the model. However, a small increase in bias may result in a large decrease in variance, which may lead to substantial improvements in prediction accuracy. The efficiency of shrinkage methods is particularly evident when the number of variables p is almost as large as the number of observations n, i.e. exactly in cases where the least squares solution has a high variance. Unlike the ridge regression, lasso performs the model selection within the fitting procedure, forcing some of the coefficient estimates to be exactly equal to zero. For this reason, *lasso* was used for the trend modeling of the soil variables in this research. The following sections discuss the lasso in more detail. An exhaustive review of lasso and its generalizations was recently published in a text by Hastie et al. (2015).

2.2.7.1 LASSO

Lasso (Least Absolute Shrinkage and Selection Operator) is the computationally attractive one-step approach for parameter estimation and variable selection for linear regression, proposed by Tibshirani (1996). Lasso combines the well-known least squares loss function with the bound on the $l1 = \sum_{j=1}^{p} |\beta_j|$ norm of coefficients, to create a sparse linear model, which is a unique global solution of the convex minimization problem. $l1$ norm is bounded by a pre-specified value t. Therefore, the coefficients of lasso regression are the solution for the following optimization problem:

$$\min_{\beta_0,\beta} \left\{ \frac{1}{N} \sum_{i=1}^{N} (y_i - \beta_0 - x_i^T \beta)^2 \right\} \text{ subject to } \sum_{j=1}^{p} |\beta_j| \leq t \tag{2.18}$$

where y_i represents the observed value of response, β represents the vector of model coefficients and x_i is a vector of predictor values for the $i-th$ case. The value of t can be understood as a *budget* which controls how large $\sum_{j=1}^{p} |\beta_j|$ can be. In this way, lasso controls the complexity of the model. For a small value t, more coefficients are forced to be exactly equal to zero, while for sufficiently large t, lasso coefficients are getting closer to their least squares estimates. In this way, lasso yields models that simultaneously use regularization to improve the model and to conduct the variable selection.

It is convenient, and more suitable for the optimization process to express Equation 2.18 in one-to-one corresponding Lagrangian matrix form:

$$\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{N} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \tag{2.19}$$

where parameter λ (*shrinkage* or *regularization* parameter) controls the strength of l_1 constraint, as 't' does in equation 2.18.

Typically, the use of penalized regression models implies the standardization of predictors prior to model fitting (Hastie et al., 2009). The reason for this lies in the dependence of the lasso solution on the variables' unit.

Selecting the optimal value of λ is the most important issue, considering that different values of λ can produce very different models. Lasso produces different coefficient estimates for each value of λ. The n-fold cross-validation procedure is a common way to select the best model, or, equivalently, the optimal λ parameter. By defining a *grid* of values of λ parameters and computing cross-validation errors e_{cv} for each value, the optimal λ value is the one which gives the lowest e_{cv}. Figure 2.4 shows the path of the coefficients over different values of λ, estimated for the SOM (Soil Organic Matter) data with *IntL* model, see Chapter 5. The upper x-axis refers to the number of predictors, while the lower x-axis refers to the *log* function of λ parameters. The y-axis refers to the values of estimated coefficients. Each line corresponds to a different model variable, which were centered and scaled prior to model fitting. As we scan from left to right on the graph, $\log(\lambda)$ increases and the the coefficient estimates move toward 0 at different rates. When the $\log(\lambda)$ is sufficiently large, many of the coefficients are set to 0. The optimal λ value can be selected by computing the cross-validation error for each value of λ. The dashed vertical line denotes the best λ parameter as calculated by the 5-fold cross-validation. In addition, lasso has one considerable advantage over the step-wise selection methods, or best subset selection. Within the cross-validation, for any value of λ, lasso fits only a single model, and the model-fitting procedure can be performed very efficiently.

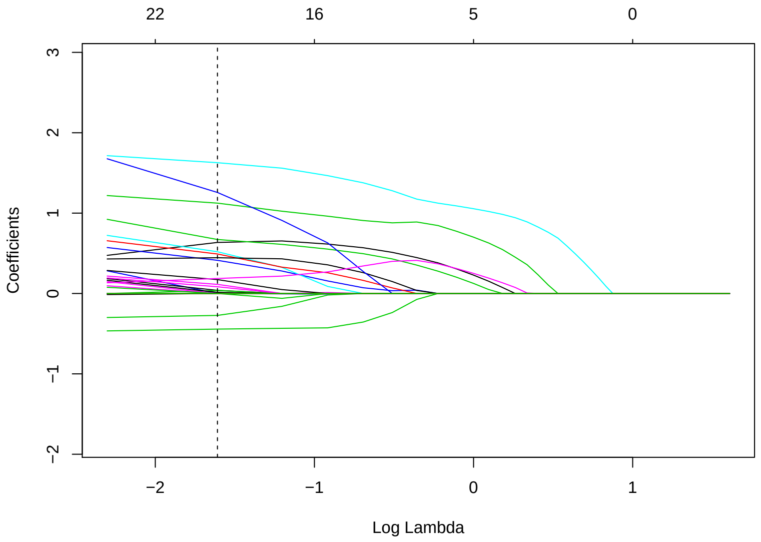


FIGURE 2.4: Coefficients path for different value of λ

2.2.7.2 LASSO for hierarchical interactions

Hierarchical interactions refers to the parameters setting in linear model, according to which the interaction terms are included in the model only if the associated 'main' terms are important or statistically significant for the prediction. The consideration of hierarchical interactions in this study is based on an approach proposed by (Bien et al., 2013). Their approach produces an interaction model that is guaranteed to be hierarchical. They consider a regression model for an outcome Y and the predictors $X_1, X_2, \dots, X_p$ with the pairwise interactions between these predictors:

$$\begin{gathered} Y = \beta_0 + \sum_j \beta_j X_j + \frac{1}{2} \sum_{j \neq k} \Theta_{jk} X_j X_k + \varepsilon \\ \text{where} \quad \varepsilon = N(0, \sigma^2) \end{gathered} \tag{2.20}$$

with the goal to estimate $\beta \in \mathbb{R}^p, \Theta \in \mathbb{R}^{p \times p}$, where $\Theta = \Theta^T$, $\Theta_{jj} = 0$. Additive terms are called *main effects*, while the multiplicative terms are called *interaction effects*. Two different types of hierarchy restrictions are defined as *strong* and *weak hierarchy*:

$$\begin{aligned} \text{Strong hierarcy :} \quad & \hat{\Theta}_{jk} \neq 0 \implies \hat{\beta}_j \neq 0 \quad \text{and} \quad \hat{\beta}_k \neq 0 \\ \text{Weak hierarcy :} \quad & \hat{\Theta}_{jk} \neq 0 \implies \hat{\beta}_j \neq 0 \quad \text{or} \quad \hat{\beta}_k \neq 0 \end{aligned} \tag{2.21}$$

They proposed a lasso procedure that produces sparse estimates of β and Θ, while satisfying either the strong or the weak hierarchy constraint. In contrast to other approaches, such as grouped lasso penalties (Yuan and Lin, 2006), their approach involves adding a set of convex constraints to the lasso:

$$\begin{aligned} \underset{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^p, \Theta \in \mathbb{R}^{p \times p}}{\text{Minimize}} \quad & q(\beta_0, \beta^+ - \beta^-, \Theta) + \lambda \|\beta\|_1 + \frac{\lambda}{2} \|\Theta\|_1 \\ \text{subject to} \quad & \Theta = \Theta^T \\ & \|\Theta\|_1 \leq \beta_j^+ + \beta_j^- \\ & \beta_j^+ \geq 0 \quad \text{for} \quad j = 1, \dots, p \\ & \beta_j^- \geq 0 \quad \text{for} \quad j = 1, \dots, p \end{aligned}$$

2.3 Geostatistical mapping - concept and methods

This section provides a concise description of geostatistical methods, which are described in more detail in Cressie (1993); Webster and Oliver (2007); Goovaerts et al. (1997); Oliver and Webster (2015); Goovaerts (1999a); Hengl et al. (2004); Goovaerts (1999b); Oliver and Webster (2014); Diggle (2011).

Geostatistics was introduced into soil science more than 30 years ago. Originally, geostatistics was developed for the mining industry (Krige, 1951), and today it is applied widely as a modeling tool in environmental sciences. As already mentioned, soil is a product of many interacting physical, chemical and biological processes. Although these processes are physically determined, their interactions are quite complex, whereas their mutual influences make the soil variation appear as if it was random (Oliver and Webster, 2014). For that reason, deterministic or any exact mathematical solution does not cover all the variations of soil property.

From the geostatistical point of view, the observation of a particular soil property at any place $z(\mathbf{x})$, where $\mathbf{x}$ represents geographic location, is considered to be just one of the infinite possible values that might be observed. Thus this value can be treated as a random variable, which is denoted by the capital Z. The set of such random variables at all of these places in one region constitutes a spatial random process (or random function), denoted as $Z(\mathbf{x})$. Random variables in the real space, such as the concentrations of elements in soil, are also called *'regionalized variables'* (Matheron, 1963).

Such a random process cannot have explicit mathematical descriptions, i.e. it cannot be expressed by a mathematical equation. On the other hand, it can be described by stochastic relations, such as spatial correlation. This means that levels of environmental variables at different places may be related to one another in a statistical sense. Intuitively, the levels of environmental variables appear to be more similar, if the spatial locations, where the values are taken, are closer to each other.

2.3.1 Stationarity and Variography

Considering that the set of actual (observed) values is a single realization of a spatial random process, it is theoretically impossible to determine any statistical parameter of the spatial random process, or even of a random process at a particular point in space. In order to overcome this limitation, geostatistical theory introduces one additional assumption, named *stationarity*. Stationarity implies that a spatial random process has the same degree of variation over a region of interest. Under the assumption of stationarity, a spatial random process can be represented as:

$$Z(x) = \mu + \varepsilon(x) \tag{2.22}$$

where μ is the mean of the process and $\varepsilon(x)$ is a random quantity with the mean of zero and the covariance, $C(\mathbf{h})$ where the **h** is the separation in space. The covariance can be expressed as:

$$C(h) = E[Z(x) - \mu Z(x+h) - \mu] = E[Z(x)Z(x+h) - \mu^2] \tag{2.23}$$

where $Z(x)$ and $Z(x+h)$ are the values of random variable Z at places x and $x+h$, and E denotes the expectation. In this way, the covariance depends only on h, which is a separation between samples, and not on their locations within the observed area. The stationarity assumption implies a constant mean over the whole area. Since this is rarely the case, Matheron (1963) has introduced a relaxed assumption called *intrinsic stationarity*, which implies that the expected differences between the values of a random variable Z at places x and $x+h$ is equal to zero. Therefore, the covariance is replaced by half the variance of the differences, referred to as the semivariance:

$$\gamma(h) = \frac{1}{2}var[Z(x) - Z(x+h)] = \frac{1}{2}E[Z(x) - Z(x+h)^2] \tag{2.24}$$

The semivariance, expressed as the function of h, is called the *variogram* $\gamma(h)$. Estimating the variogram values from the observed data, $z(x_1), z(x_2), ...z(x_n)$ by changing h, is usually the first step in any geostatistical analysis. The usual way to compute the variogram values for different h is the Matheron's method of moments (MoM):

$$\hat{\gamma}(h) = \frac{1}{2m(h)} \sum_{i=1}^{m(h)} [z(\mathbf{x}_i + h) - z(\mathbf{x}_i)]^2 \tag{2.25}$$

where $z(\mathbf{x_i})$ and $z(\mathbf{x_i} + \mathbf{h})$ are the observed values of z at places x_i and $x_i + h$,and $m(h)$ is the number of paired comparisons at lag $\mathbf{h}$. This set of values is called the experimental or sample variogram, because it is based on the observed data. The sample variogram can be modeled by fairly simple mathematical functions. The most used variogram models are: *Nugget*, *Exponential*, *Spherical*, *Gaussian*, *Linear*, and *Power* (Oliver and Webster, 2014). The obtained model of sample variogram leads to the geostatistical prediction technique known as kriging.

2.3.2 Ordinary Kriging

Kriging is a generic name for an entire family of geostatistical interpolation techniques. Kriging technques provide predictions on punctual or block supports that are unbiased and have minimum prediction errors. For this reason, kriging is often known as the Best Linear Unbiased Predictor (BLUP). Kriging predicts values at unsampled locations by weighting the neighboring measurements in a way that takes into account the structure of spatial dependence, as represented in the variogram or the covariance function.

Ordinary kriging is by far the most common type of kriging. Ordinary kriging is based on the assumptions that the variation is random and spatially dependent, and that the underlying random process is intrinsically stationary with a constant mean and a variance that depends only on separation distance, and not on absolute position within the observed area (Oliver and Webster, 2015). The whole computation can refer to one, two, or three dimensional space, as well as to the point or block support. The most common case is still two-dimensional, but later in this work an extension to the three-dimensional space will be presented.

If we denote the values of random variable Z that have been collected at locations $x_1, x_2, x_3 ... x_n$ as $z(x_i)$. Kriging prediction $\hat{Z}$ of a random variable Z at any new point $\mathbf{x_0}$ is given by:

$$\hat{Z}(x_0) = \sum_{i=1}^{N} \lambda_i z(x_i) \tag{2.26}$$

where λ_i are the weights. In order to ensure the unbiased estimate the weights are summed to one:

$$\sum_{i=1}^{N} \lambda_i = 1 \tag{2.27}$$

The prediction variance is given by:

$$var[\hat{Z}(x_0)] = E[\hat{Z}(x_0) - z((x_0)^2] = 2\sum_{i=1}^{N} \lambda_i \gamma(x_i - x_0) - \sum_{i=1}^{N}\sum_{j=1}^{N} \lambda_i \lambda_j \gamma(x_i - x_j) \tag{2.28}$$

where the quantity $\gamma(x_i - x_0)$ is the semivariance of Z between the sampling point x and the target point x_0. $\gamma(x_i - x_j)$ is the semivariance between the $i-th$ and $j-th$ sampling points. It is important to note here that the kriging variances are independent from the data values, and, as such, cannot be used as a measure of reliability of the kriging predictions.

The essential step in kriging prediction is to find the kriging weights that ensure the minimized kriging prediction error. These are found by solving the following system of equations:

$$\begin{gathered} \sum_{j=1}^{N} \lambda_i \gamma(x_i - x_j) + \psi(x_0) = \gamma(x_i - x_0) \quad \text{for all } j \\ \sum_{i=1}^{N} \lambda_i = 1 \end{gathered} \tag{2.29}$$

the $\psi(x_0)$ is the Lagrange multiplier introduced to achieve minimization.

In matrix form, this system can be expressed as:

$$\begin{bmatrix} \hat{\lambda}_{\mathbf{0}} \\ \mu \end{bmatrix} = \begin{bmatrix} \gamma(x_1, x_1) & \cdots & \gamma(1_n, x_n) & 1 \\ \vdots & \ddots & \vdots & \vdots \\ \gamma(x_n, x_1) & \cdots & \gamma(x_n, x_n) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma(x_n, x_0) \\ \vdots \\ \gamma(x_n, x_0) \\ 1 \end{bmatrix} \tag{2.30}$$

the additional parameter μ is a Lagrange multiplier, see details in Isaaks and Srivastava (1990).

2.4 Universal Model of Soil Variation and Hybrid Techniques

Despite the assumption that soil variation is a realization of a spatially random process, it may turn out that a significant part of variation cannot be treated in this way (Lark et al., 2006). For example, soil properties that are influenced by topography may show a pronounced trend across an explored area, which is not consistent with the constant mean model from Equation 2.22. For this reason, it is convenient to extend this model with a more generic *universal model of soil variation*:

$$Z(x) = u(x) + \varepsilon(x) + \varepsilon \tag{2.31}$$

The universal model of soil variation distinguishes three major components: (1) the deterministic-trend component $u(x)$, (2) the spatially correlated component (stochastic residuals) $\varepsilon(x)$ and (3) pure noise ε. The deterministic component refers to a systematic part of variation caused by the strong impact of other environmental factors, and can be materialized through a deterministic function of coordinates, or available spatial covariates (*scorpan* factors). This part of variation is also known as the 'trend component'. The second component covers spatially correlated small-scale variations, described by the variogram function. Accordingly, the variogram is no longer estimated based on the observed data but is rather visualized using the residuals $\varepsilon(x) = Z(x) - u(x)$. The third component includes the part of the spatial variation which cannot be described by the means of the previous two components.

Numerous mapping techniques have been developed to accommodate the varying mean by combining the information from auxiliary sources with observations. All these techniques are known as 'Hybrid Techniques' (McBratney et al., 2000). The first proposed and probably the most commonly used hybrid technique is the Universal Kriging method (Matheron, 1969).

2.4.1 Universal Kriging

Universal kriging (UK) uses an integral computing procedure for the estimation of trend and residual interpolation by kriging. The original version, proposed by (Matheron, 1969), models the trend as a linear function of spatial coordinates:

$$u(x) = \sum_{k=0}^{K} \beta_k f_k(x) \tag{2.32}$$

where $\beta_k, k = 0, 1, \ldots, K$ are unknown coefficients, and the $f_k(x)$ are known functions of x (i.e. functions of spatial coordinates).

If a variogram model $\gamma(h)$ is given, the prediction of Z at any x_0 can be obtained by:

$$\hat{Z}(x_0) = \sum_{i=1}^{n} \lambda_i f_k(x_i) \tag{2.33}$$

where $\lambda_i, i = 1, 2, ..., N$ are the UK weights. The estimator is unbiased if:

$$\sum_{i=1}^{N} \lambda_i f_k(x_i) = f_k(x_0) \tag{2.34}$$

The UK can be expressed as an extended ordinary kriging, taking into account the fixed effects of the trend in addition to the spatially correlated component (Webster and Oliver, 2007):

$$\sum_{j=1}^{N} \lambda_i \gamma(x_i,x_j) + \psi_0 + \sum_{k=0}^{K} \psi_k f_k(x_j) = \gamma(x_0,x_j) \tag{2.35}$$

$$\sum_{i=1}^{N} \lambda_i = 1 \tag{2.36}$$

$$\sum_{i=1}^{N} \lambda_i f_k(x_i) = f_k(x_0) \tag{2.37}$$

The values of $\gamma(x_i,x_j)$ are the semivariances of the residuals between the data points x_i and x_j, and the $\gamma(x_0,x_j)$ are the semivariances between the target point and the data points. Moreover, there are additional Lagrange multipliers ψ_k for each term of the trend model. The universal kriging, like ordinary kriging, is a set of linear equations which can be represented in matrix notation by:

$$\begin{bmatrix} \hat{\lambda}_1 \\ \hat{\lambda}_2 \\ \vdots \\ \hat{\lambda}_N \\ \psi_0 \\ \psi_1 \\ \psi_2 \\ \vdots \\ \psi_K \end{bmatrix} = \begin{bmatrix} \gamma(x_1,x_1) & \cdots & \gamma(x_1,x_N) & 1 & f_1(x_1) & \cdots & f_K(x_1) \\ \gamma(x_2,x_1) & \cdots & \gamma(x_2,x_N) & 1 & f_1(x_2) & \cdots & f_K(x_2) \\ \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \\ \gamma(x_N,x_1) & \cdots & \gamma(x_N,x_N) & 1 & f_1(x_N) & \cdots & f_K(x)_N \\ 1 & \cdots & 1 & 0 & 0 & \cdots & 0 \\ f_1(x_1) & \cdots & f_1(x_N) & 0 & 0 & \cdots & 0 \\ f_2(x_1) & \cdots & f_2(x_N) & 0 & 0 & \cdots & 0 \\ \vdots & \cdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ f_K(x_1) & \cdots & f_K(x_N) & 0 & 0 & \cdots & 0 \end{bmatrix}^{-1} \begin{bmatrix} \gamma(x_1,x_0) \\ \gamma(x_3,x_0) \\ \vdots \\ \gamma(x_N,x_0) \\ 1 \\ f_1(x_0) \\ f_2(x_0) \\ \vdots \\ f_K(x_0) \end{bmatrix} \tag{2.38}$$

The major limiting factor in using the UK is that it requires the knowledge of a residual variogram, prior to estimating the regression coefficients. This creates a circular problem, since the computation of the residual variogram, which is needed for the UK, requires the estimated trend coefficients.

2.4.2 Regression Kriging

Regression kriging (RK) assumes that deterministic and stochastic components of spatial variation can be modeled separately. It is mathematically equivalent to the previously explained UK, where auxiliary predictors are used to solve the kriging weights directly. However, RK combines two conceptually different techniques, regression for trend estimation and ordinary or simple kriging, to interpolate stochastic residuals (Hengl et al., 2007; Bajat et al., 2013).

The regression kriging prediction for variable Z at new location $\mathbf{x}_0$ is:

$$\hat{z}(x_0) = \hat{m}(x_0) + \hat{e}(x_0) = \sum_{k=0}^{p} \hat{\beta}_k \cdot f_k(x_0) + \sum_{i=1}^{n} \lambda_i \cdot \varepsilon(x_i) \tag{2.39}$$

where the $\hat{\beta}_i$ are estimated regression coefficients, the f_k is the known function of $k-th$ covariate, that must be exhaustively known over the spatial domain, and p is the number of covariates. The λ_i are the kriging weights determined by the spatial dependence structure, and $e(x_i)$ is the regression residual at location x_i.

In practice, trend coefficients are mainly obtained by ordinary least squares (OLS). However, this can cause bias in the estimates of the residual variogram (Cressie, 1993). One solution to reduce the bias is to estimate the residuals taking into account the spatial correlation between the observations Hengl et al. (2007). For that reason, the usage of Generalized Least Squares (GLS) is recommended instead of the commonly used OLS. However, GLS implies an iterative procedure for the variogram estimation. In the first step, the trend model is estimated by using the OLS. The given OLS residuals are then used to construct the covariance function needed to obtain the GLS estimates. In the next step, the GLS residuals are used to update the covariance function in order to recalculate the GLS residuals, from which an updated covariance function is computed. This procedure should be repeated until the trend coefficients no longer change. The final residual variogram is then estimated from the final GLS residuals, and then modeled as a continuous function of lag distance.

If the covariance matrix of the residuals is denoted as **C**, the matrix of covariate values at the sampling locations as **q**, and, the vector of measured values of the target variable as **z**, the vector of trend coefficients obtained by GLS ($\hat{\beta}_{\mathtt{GLS}}$), is:

$$\hat{\beta}_{\mathtt{GLS}} = \left(\mathbf{q}^{\mathbf{T}} \cdot \mathbf{C}^{-\mathbf{1}} \cdot \mathbf{q}\right)^{-\mathbf{1}} \cdot \mathbf{q}^{\mathbf{T}} \cdot \mathbf{C}^{-\mathbf{1}} \cdot \mathbf{z} \tag{2.40}$$

The Equation 2.39 can be rewritten in matrix form and the kriging prediction at new location x_0 is:

$$\hat{z}(x_0) = \mathbf{q}_{\mathbf{0}}^{\mathbf{T}} \cdot \hat{\beta}_{\mathtt{GLS}} + \lambda_{\mathbf{0}}^{\mathbf{T}} \cdot (\mathbf{z} - \mathbf{q} \cdot \hat{\beta}_{\mathtt{GLS}}) \tag{2.41}$$

where $\hat{\lambda}_{\mathbf{0}}$ is the estimated vector of weights for the location x_0. Prediction variance is defined as:

$$\begin{aligned} \hat{\sigma}^2(x_0) = {} & (C_0 + C_1) - \mathbf{c}_{\mathbf{0}}^{\mathbf{T}} \cdot \mathbf{C}^{\mathbf{1}} \cdot \mathbf{c_0} \\ & + \left(\mathbf{q_0} - \mathbf{q}^{\mathbf{T}} \cdot \mathbf{C}^{-\mathbf{1}} \cdot \mathbf{c_0}\right)^{\mathbf{T}} \cdot \left(\mathbf{q}^{\mathbf{T}} \cdot \mathbf{C}^{-\mathbf{1}} \cdot \mathbf{q}\right)^{-\mathbf{1}} \cdot \left(\mathbf{q_0} - \mathbf{q}^{\mathbf{T}} \cdot \mathbf{C}^{-\mathbf{1}} \cdot \mathbf{c_0}\right) \end{aligned} \tag{2.42}$$

where $C_0 + C_1$ is the sill variation and $\mathbf{c}_0$ is the vector of covariances of residuals at the unvisited location, **C** is the covariance matrix of the residuals, **q** is a matrix of covariate values at the sampling locations, $\mathbf{q_0}$ is a matrix of covariate values at the unvisited location.

2.5 Mapping in 3D

2.5.1 Modeling Soil Variation with depth

Soil sampling is often based on taking a bulked sample of soil from each horizon within the soil profile. Accordingly, measurements of particular soil properties are assumed to reflect the mean values for the soil horizons from which the samples were taken. If we assign the lower and the upper bound of each horizon to each observation, vertical variation of profile data can be expressed as a step-wise function of depth. However, this concept is often too restrictive, because it assumes that the soil horizons are perfectly homogeneous. For that reason, soil scientists were interested to find more realistic representations of vertical soil variation. Such realizations were achieved by fitting continuous functions, such as exponential decay, log-log functions, polynomials or even piece-wise polynomials, through the mid-depth of horizon data (Moore et al., 1972). Ponce-Hernandez et al. (1986) proposed the specific depth function, called *equal-area* spline or *mass-preserving* spline, which fit the piece-wise spline function through the horizon averages, maintaining that the areas above and below the fitted spline in any horizon are equal. However, different functions yield various predictions of soil properties along soil profile. Figure 2.5 depicts a vertical variation of soil carbon modeled by using a logarithmic function (left) and an equal-area spline (right).

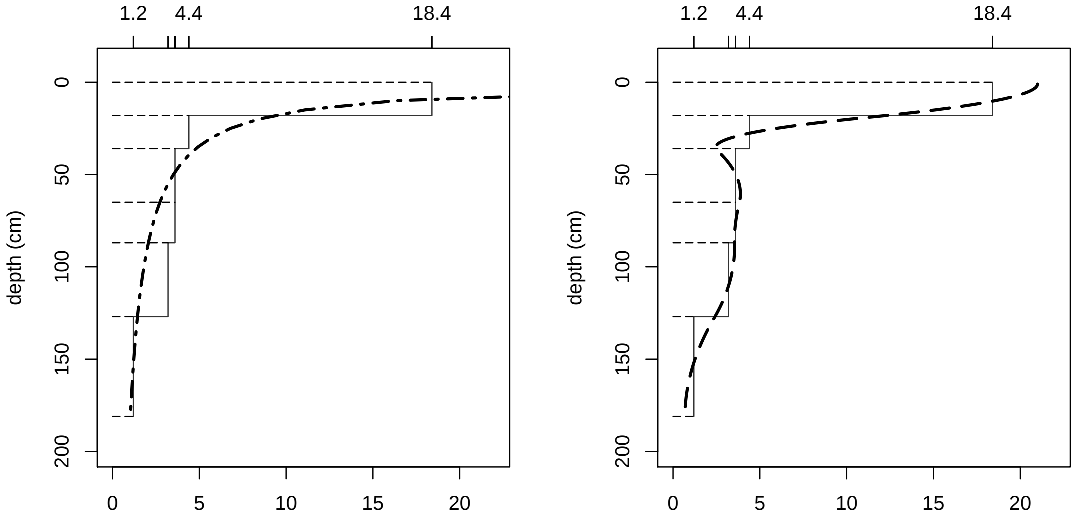


FIGURE 2.5: Log-log depth function (left) and equal-area spline depth function (right), from (Hengl and Heuvelink, 2013)

Equal-area spline is a continuous function of depth which must be estimated by using the profile data. The assumptions behind the equal-area spline imply that the $f(x)$ and its first derivative $f'(x)$ are continuous, and also that the $f'(x)$ is square integrable. The depth is denoted by x, and the depth function describing soil attribute values by $f(x)$. Further, if the depths of the boundaries of the n horizons are denoted by $x_0 < x_1, \ldots, < x_n$, where x_0 is the soil surface, so that $x_0 = 0$, then the measurement from the horizon i is assumed to reflect the mean level of soil properties at this depth (horizon). Thus, the equal-area spline models the measurements y_i as:

$$y_i = \bar{f}_i + \varepsilon_i \tag{2.43}$$

where $\bar{f}_i = \int_{x_{i-1}}^{x_i} f(x)dx/(x_i - x_{i-1})$ is the mean value of $f(x)$ over the interval $(x_i - x_{i-1})$. The errors ε_i are assumed to be independent, with mean 0 and the variance σ^2. The spline function that models $barf$ requires choosing the $f(x)$ that minimizes:

$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{f}_i)^2 + \lambda \int_{x_0}^{x_n} [f'(x)]^2 \tag{2.44}$$

The first term of Equation 2.44 represents a fit to the data, while the second term measures the roughness of function $f(x)$ represented by its first derivative $f(x)$. The parameter λ, known as the spline-smoothing parameter, controls the trade-off between the fit and the roughness penalty. The quality of the fit for the equal-area spline function largely depends on the λ value. Previous results, obtained in the studies of Bishop et al. (1999); Adhikari et al. (2012); Odgers et al. (2012) show that the value 0.1 for λ parameter provides the best fitting results.

Bishop et al. (1999) compared the predictive performance of equal-area spline with the exponential decay functions, and 1st and 2nd degree polynomial depth functions in predicting a number of soil properties including soil pH, electrical conductivity (EC), clay content and organic carbon content. The obtained results indicated the superiority of equal-area quadratic splines. Malone et al. (2009) made a minor modification to the their work and proposed a more general method based on equal-area spline, so that input data segments do not have to be contiguous with depth.

2.5.2 Spline Than Krige

Soil samples may relate to different depth intervals between sampling locations. This may cause a problem in assessing the spatial distribution of a soil property at a particular depth interval which may not correspond to the sampled intervals. A common way to map the soil property at any depth interval is the so-called Spline-Than-Krige (STK) approach. The spline-Then-Krige refers to a 2D geostatiscal approach for producing a suite of digital maps for soil properties at different soil depth intervals, first proposed by Malone et al. (2009). Methodologically, this approach combines the use of depth functions and geostatistical hybrid techniques to provide the estimate of soil property at unsampled locations and sepcific depth intervals. This approach was successfully used in many studies for the mapping of various soil properties (Adhikari et al., 2012, 2014; Lacoste et al., 2014; Mulder et al., 2016). Generally, STK approach can be conceptualized through the following procedural steps:

1. Fitting equal-area spline functions to soil profiles data and selecting the best λ;
2. Deriving mean values of the 'best' spline function, within the pre-specified depth intervals;
3. Modeling the relationship between the mean values and the environmental covariates;
4. Applying the given model onto the wider study area where soil observations do not exist;
5. Kriging the residuals at each depth interval;
6. Adding the kriging prediction of residuals to the 'trend' prediction to obtain final predictions;
7. Reconstructing the spline function at each predicted point with the same λ.

A major disadvantage of converting the soil profile observations to the continuous form by exact mathematical function is that these values are only estimates with associated estimation errors. If these values [1] are used as observations for spatial prediction at these depths, then an important source of error is disregarded, which may jeopardize the quality of the final soil prediction (Hengl and Heuvelink, 2013).

[1] values of averaged spline predictions over the depth increments

2.5.3 Model based 3D modeling

Model based 3D modeling refers to the methodology of soil mapping in which the variation of a soil property in three dimensions is described by a single model. 3D soil modeling is a natural extension of purely 2D approaches, such as Spline-Than-Krige.

The universal model of soil variation (Equation 2.31) from Section 2.4 can be extended to cover the variation of soil properties in 3D (horizontal + depth). The extension of the universal model of soil variation is based on the fact that soil varies in both horizontal and vertical directions, as well as that the soil properties are auto-correlated in both directions. Therefore, the universal model of soil variation can be formulated as follows:

$$Z(x,d) = \mu(x,d) + \varepsilon'(x,d) + \varepsilon \tag{2.45}$$

The trend component $\mu(x,d)$ is now a function of spatial covariates and the depth, measured from the terrain surface. It may be further decomposed into additive consisting of purely spatial and purely depth-related components (Hengl et al., 2014). The spatially correlated component is typically characterized by the 3D variogram model.

2.5.3.1 3D Variogram modeling

Spatial continuity of soil variables is particularly characterized by the strong anisotropy between horizontal and vertical directions. Spatial continuity observed in the depths of a few centimeters may correspond to several kilometers, or more, in horizontal direction (Hengl et al., 2015). The levels of continuity in both directions can be quantified and compared by calculating variograms in those directions. However, to incorporate the anisotropy into a 3D geostatistical model, an anisotropic 3D variogram model must be provided. Generally, the anisotropy is defined by major direction of continuity and the anisotropy ratio. The major direction of continuity is the direction in which the greatest continuity is observed. The continuity in a particular direction is greater if the range of directional variograms is larger than in any other direction. *Geometric* anisotropy occures if the two variograms reach the same sill, but at different ranges. In addition, the *zonal* anisotropy can also occur, whereby the sill varies as the variogram direction is changed.

The anisotropy ratio represents the magnitude of anisotropy. In the two dimensional setting, anisotropy ratio is typically reported as the ratio between the ranges of variograms calculated in two principal directions of spatial continuity; the direction of the greatest continuity (major direction) and the direction perpendicular to it, which is typically considered as the direction of the minimum continuity (minor direction). Therefore, the anisotropy ratio is used to quantify how much larger the continuity is in a major direction compared with the minor direction:

$$anisotropy\ ratio = \frac{range\ in\ minor\ direction}{range\ in\ major\ direction} \tag{2.46}$$

Alternatively, the anisotropy ratio can be expressed as a relative ratio, where the larger number represents the relative range in the major direction, and the smaller number represents the relative range in the minor direction.

In three-dimensional settings, in which the soil data are actually collected, it is common to distinguish three principal directions: major, minor and vertical (depth). Major and minor directions represent the two principal directions of spatial continuity in horizontal space. If we assume that the spatial continuity is isotropic in horizontal space, it remains to determine the anisotropy ratio between the the vertical and any horizontal direction. This is exactly the case with soil data, where the two principal directions of spatial continuity are almost always known *a priori*, considering the fact that the largest anisotropy ratio occurs between the vertical and the horizontal directions.

Anisotropy can be incorporated in the 3D anisotropic variogram model, once the principal direction of spatial continuity and the anisotropy ratio is determined. In traditional geostatistics, it is common to calculate the *effective anisotropic distances* (EAD) for this purpose. The effective anisotropic distance is a unitless scalar distance that is calculated as Euclidean norm of lag (h) components h_{major}, h_{minor} and h_{depth}, each divided by the corresponding range of spatial continuity a_{major}, a_{minor} and a_{depth}:

$$h_{EAD} = \sqrt{(\frac{h_{major}}{a_{major}})^2 + (\frac{h_{minor}}{a_{minor}})^2 + (\frac{h_{depth}}{a_{depth}})^2)} \tag{2.47}$$

Each directional variogram is reduced to one common model with a standardized range equal to 1, i.e. a variogram with range 1 and a variogram with range a yield equal value for the same lag (Isaaks and Srivastava, 1990):

$$\gamma_a(\frac{h}{a}) = \gamma_1(h) \tag{2.48}$$

Directional model with range a can be reduced to a standardized model with range 1 simply by replacing the separation distance, h, by a reduced distance h/a. Therefore, the corresponding 3D anisotropic variogram is given by:

$$\gamma(h) = \gamma(h_{major}, h_{minor}, h_{depth}) = \gamma_1(h_{ead}) \tag{2.49}$$

In space-time geostatistics, equivalent model is known as the *metric* model:

$$\gamma(h,d) = \gamma(\sqrt{h^2 + (\alpha \times d)^2}) \tag{2.50}$$

where the distances in the third dimension d are simply rescaled by anisotropy parameter α in order to be comparable with the distances h in other dimensions.

A more general model is known as the separable (product) covariance model. It was used in the study by Orton et al. (2016) as a part of their comprehensive approach to three-dimensional modeling of soil variables. Separable covariance model can be expressed as:

$$\gamma(h,d) = \text{nug} \times 1_{h>0,d>0} + \text{sill} \times (\gamma_s(h) + \gamma_d(d) - \gamma_s(h) \times \gamma_d(d)) \tag{2.51}$$

The most comprehensive model was proposed by Heuvelink and Griffith (2010). This model is known as the *sum-metric* model:

$$C(h,u) = C_1(h_1) + C_v(h_v) + C_{lv}(\sqrt{h_l^2 + (\alpha h_v)^2}) \tag{2.52}$$

The corresponding variogram model is:

$$\gamma(h,d) = \gamma_1(h) + \gamma_d(d) + \gamma_{hd}(\sqrt{h^2 + (\alpha d)^2}) \tag{2.53}$$

Brus et al. (2016) made an analogy with the space-time analysis from Heuvelink and Griffith (2010) and used *sum-metric* covariance structure to model the spatial structure of soil organic carbon in 3D. Sum-metric model distinguishes three components of variation: $C_1(h_1)$, the covariance in horizontal direction at the distance of h_1; $C_v(h_v)$, the covariance in vertical direction at the distance of h_v; and $C_{lv}(h_{lv})$, the covariance in any direction, and α, the geometric anisotropy ratio. By modeling the covariance as a sum of a covariance in horizontal direction and a covariance in vertical direction, we can account for different residual variances in these two directions (zonal anisotropy). The geometric anisotropy ratio α in the third covariance term is needed because one distance unit in the vertical direction is not equivalent to one distance unit in the horizontal direction.

2.5.3.2 Spatial prediction in 3D

Common method for spatial prediction in 3D is 3D regression kriging. It can be expressed as:

$$\hat{z}(x_0,d_0) = \sum_{j=0}^{p} \hat{\beta}_j X_j(x_0,d_0) + \hat{g}(d_0) + \sum_{i=1}^{n} \hat{\lambda}_i(x_0,d_0)\varepsilon(x_i,d_i) \tag{2.54}$$

where $\hat{z}$ is the predicted soil property, x_i are geographical locations and d_i is depth, measured downward from the land surface. $\sum_{j=0}^{p} \hat{\beta}_j X_j(x_0,d_0)$ and the $\hat{g}(d_0)$ are the predictions of two trend components, horizontal and vertical. Horizontal component is expressed as a standard multiple linear regression model, whilst the vertical component is expressed as any function g of depth. In the study of Hengl et al. (2014), the vertical component is modeled by spline function. The $\hat{\lambda}_i(x_0,d_0)$ are kriging weights derived from spatial covariance structure and $\varepsilon(x_i,d_i)$ are the residuals interpolated by using 3D kriging.

Chapter 3

Data and Case Study

3.1 Case Study Area

The study area is situated in the central part of Eastern Serbia, a approximately 10 kilometers in north-east direction from the town of Bor (Figure 3.1). Bor is a small town, widely known as one of the main centers of mining and metallurgical industry in this part of Europe. The Municipality of Bor covers an area of 856 km^2. The town contains a total of 35.000 inhabitants and an additional 20.000 people are settled in the surrounding settlements.

The north-south transect of the survey area is about 20 km, while the east-west transect is about 10 km. Study area occupies the territory the three districts of Bor municipality, called Čoka Kuruga, Čoka Kupjatra i Tilva Njagra. More precisely, the area is located between the Zlot limestone massif on the west, the village of Zlot on the south, Bor lake on the southeast, Žagubica district on the north, and the Krivelj limestone massif on the northeast.

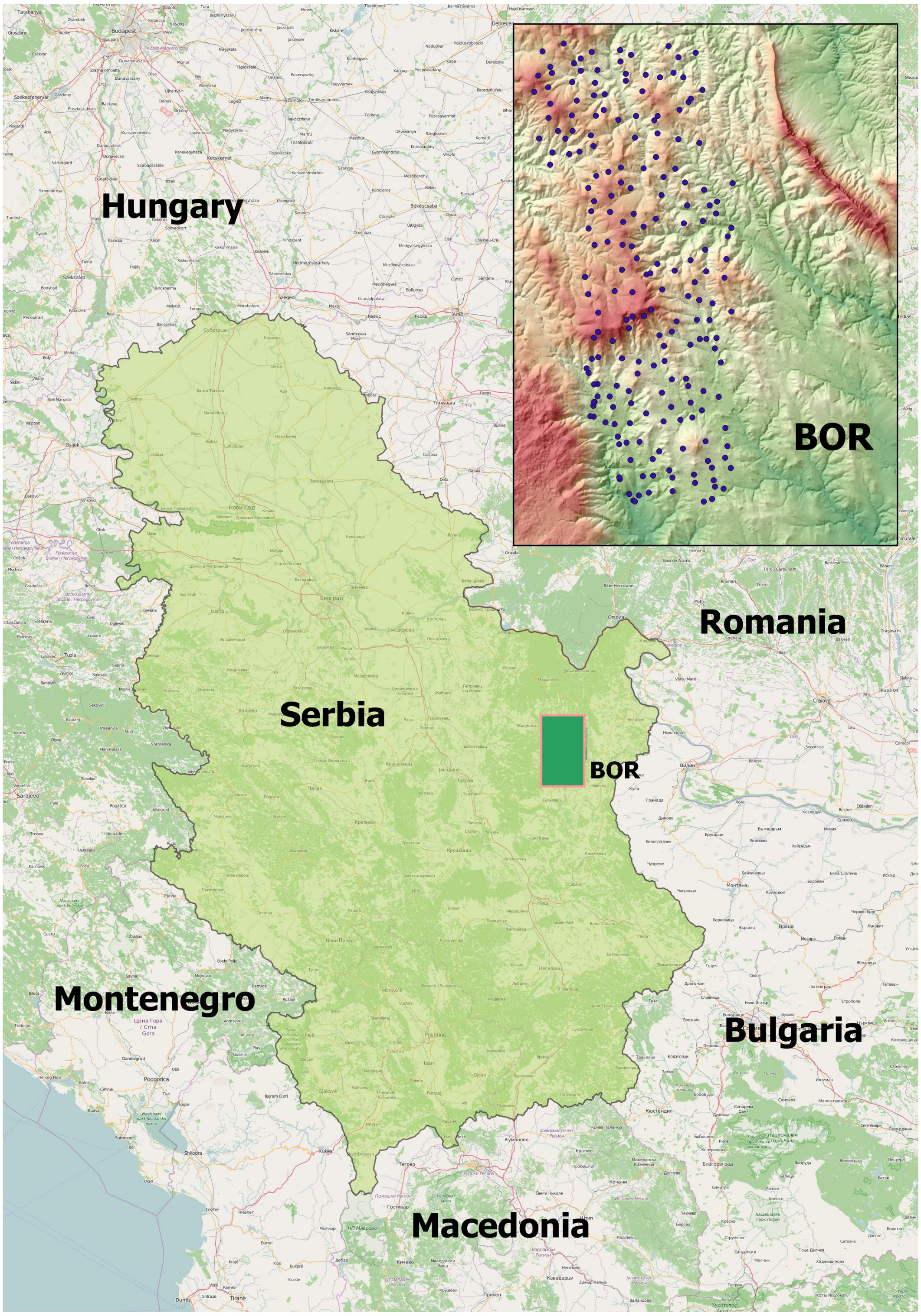


FIGURE 3.1: Location of case study area

Topographically, this is a predominantly hilly and mountainous area with terrain heights varying from 387 m to 1243 m. Mountain Crni vrh, Tilva Njagra and the Zlot limestone massif dominate in the relief structure of the sampling area. The surrounding area is covered with deciduous forests and agricultural lands. The landscape around the Bor Lake, where, agricultural crops have been produced for a long time, is predominantly hilly. Oak is the dominant type of tree in these forests. These forests are well preserved by a dense upper storey, which is very important for the protection of the forest soils from erosion, even on very steep slopes. Areas with burned forest occur sporadically.

There are several streams located in the study area. In the southern area there are the Zlot and Brestovacka Rivers, in the eastern area the river Krivelj, in the northern area the rivers Lipa and Velika Tisnica with two big tributaries, the Varfa Strz and Crna. Due to the fact that the rivers are short and located in narrow valleys, they do not influence the soil formation processes.

The climate of the Bor region is characterized by long severe winters and cool short summers with moderate precipitation. Mean annual temperature at Bor is 10.1$^{\circ}C$ and at Crni Vrh is 8.0$^{\circ}C$. Temperature range is large with the absolute minimum air temperature being -27.0$^{\circ}C$ in January and the absolute maximum air temperature reaching +41$^{\circ}C$ in July and August, with the mean summer temperature being +20.0$^{\circ}C$. Mean annual precipitation is 707 mm at Zlot, while at Crni Vrh it is 850 mm. Mean monthly precipitation is uneven, with most of the precipitation occurring in May and June rather then in October and November. The precipitation amount during the growing period in Bor area is 354 mm. Air circulation is mainly controlled by prevailing northwest and eastern winds. Winds from the northwest prevail during warmer months, whereas eastern and southeastern winds prevail during colder periods of the year. Table 3.1 depicts the average wind speed and wind directions in Bor for period 1998-2009.

The development of mining and metallurgy in Bor has caused a serious effects on the environment over more than hundred years of production. The copper smelter, which is a part of the Mining-Metallurgical Complex Bor is recognized as the major pollution source in this region. The Smelter plant which processes copper concentrate, emits high quantities of SO_2 (20,000 tons/year), arsenic (300 tons/year) and heavy metals (including 150 kg mercury/year) into the atmosphere which has caused erosion, high acidity of soils and destruction of vegetation in the this area. It is estimated that over 25,500 hectares of

TABLE 3.1: Average wind speed and wind direction (%) in Bor, 1998-2009, from (Kovačević et al., 2010)

Year	Calm	N	NNE	NE	ENE	E	ESE	SE	SSE	S	SSW	SW	WSW	W	WNW	NW	NNW
1998	56.6	0.2	0.2	0.8	7.5	3.9	0.3	0.3	0.2	3.3	0.7	0.1	1	7	6.9	10	1
1999	61.2	2.2	0	0.1	5.2	3.2	0.5	0.3	0.1	2.7	0.7	0.3	0.5	3.4	6.4	9.7	1.4
2000	75.7	0.5	0.1	0.1	3.1	2	0	0.1	0.2	2	0.7	0.2	0.5	2.4	6.2	5.5	0.8
2001	66.1	0.2	0.1	0.4	3.3	2.6	0.2	0.2	0.2	3.3	0.1	0.4	0.2	3.1	6.4	0	0
2002	58	0.8	0.7	0.6	3.1	8.5	0.5	0.2	0.4	4	0.4	0.6	1.7	7.4	8.4	4.4	0.3
2003	62.3	0.2	0.2	0.1	2.3	7	0.4	0.3	0.3	1.8	0.3	0.3	0.8	6.5	9.4	6.5	1.3
2004	51.7	0.9	0.2	0.3	1.5	7.6	0.9	0.4	0.4	4.7	0.8	0.4	1.2	6.1	11.2	10.7	1
2005	54.3	1.5	0.2	0.3	1.5	8.1	1.2	0.3	0.4	3.9	0.3	0.1	1.4	7.7	9.4	7.1	0.7
2006	53.6	0.7	0.1	0.3	1.4	6.8	1.3	0.4	0.6	3.9	0.3	0.2	1.4	8.5	9.6	8.2	0.8
2007	49.8	0.4	0.7	0.2	2.3	7.9	1.3	0.5	0.6	5.4	1.5	0.4	1.4	8.6	10.7	7.8	1.1
2008	50.9	0.6	0.2	0.1	3	7.6	1.3	0.6	0.6	4.1	2.2	0.5	1.4	10.4	9.2	5.5	1.8
2009	58.2	0.4	0.3	0.6	3.2	7.8	1.7	0.4	0.7	0.7	3.4	0.9	0.2	1.2	9.3	6.4	3.9
Average	58.2	0.7	0.2	0.3	3.1	6.1	0.8	0.3	0.4	3.3	0.7	0.4	1	6	8.6	7.4	1.3

soil are damaged, which accounts for 60% of the agricultural soil in the Bor municipality (LEAP, 2003). It is known that the distribution of air pollutants emitted from the copper smelter is strongly influenced by the smelter operation mode and meteorological parameters such as wind speed and direction. There are several studies conducted in urban and sub-urban areas surrounding the copper smelter in Bor which prove the seriousness of the problem of environmental pollution caused by copper production in Bor (Serbula et al., 2013, 2014)

3.2 Data

The survey of the area was carried out in June 2006 with the aim to document the existing conditions of soil prior to mining investigation. The preliminary survey of the land was performed to obtain the data concerning the natural characteristics of the area and to approximate the selection of the soil units and soil types.

The in-depth field study has involved the opening of 205 soil profiles and 382 boreholes, with recording their coordinates via GPS. The boreholes were used for the establishment of boundaries between different soil types and soil sub-types, as well as for the determination of the basic morphological characteristics. In the soil profiles the morphological characteristics were described and samples were taken from the horizons. Profiles depth varies considerably, from 24 cm to 1.2 m. Consequently, the number of samples per soil profile varies from 1 to 5, according to soil horizons. Samples were taken from four soil horizons including: O (organic soil horizon), A horizon, B (if it existed) and C. The depth to the top of the C-horizon varied between 10 and 123 cm. Therefore each sample corresponds to the different soil depth increments. Table 3.2 summarizes the number of soil samples according to standard soil depth increments. In total, 450 soil samples were collected and analyzed for comprehensive physical, chemical, and microbiological properties. Among other soil properties, As concentration expressed in `mg/kg`, SOM content expressed in %, and pH (measured in H_2O) were selected as target soil properties for this research.

TABLE 3.2: Number of soil samples per soil depths

0-5 cm	5-15 cm	15-30 cm	30-60 cm	60-100 cm	100-200 cm
204	204	185	134	52	6

Figure 3.2 illustrates the three different samples of 20 soil profiles. Colors illustrate the observed values for As concentration, SOM content and pH. Figure 3.3 depicts the spatial distribution of soil profiles along with the soil type and the depth classes, where the size and the color of the circle identifies which profiles belongs to which soil type or reaches particular depth. It is important to note that there exist a number of soil types with the relatively small number of soil profiles (Figure 3.4) which might hamper a subsequent statistical modeling of relationship between the soil property and soil type.

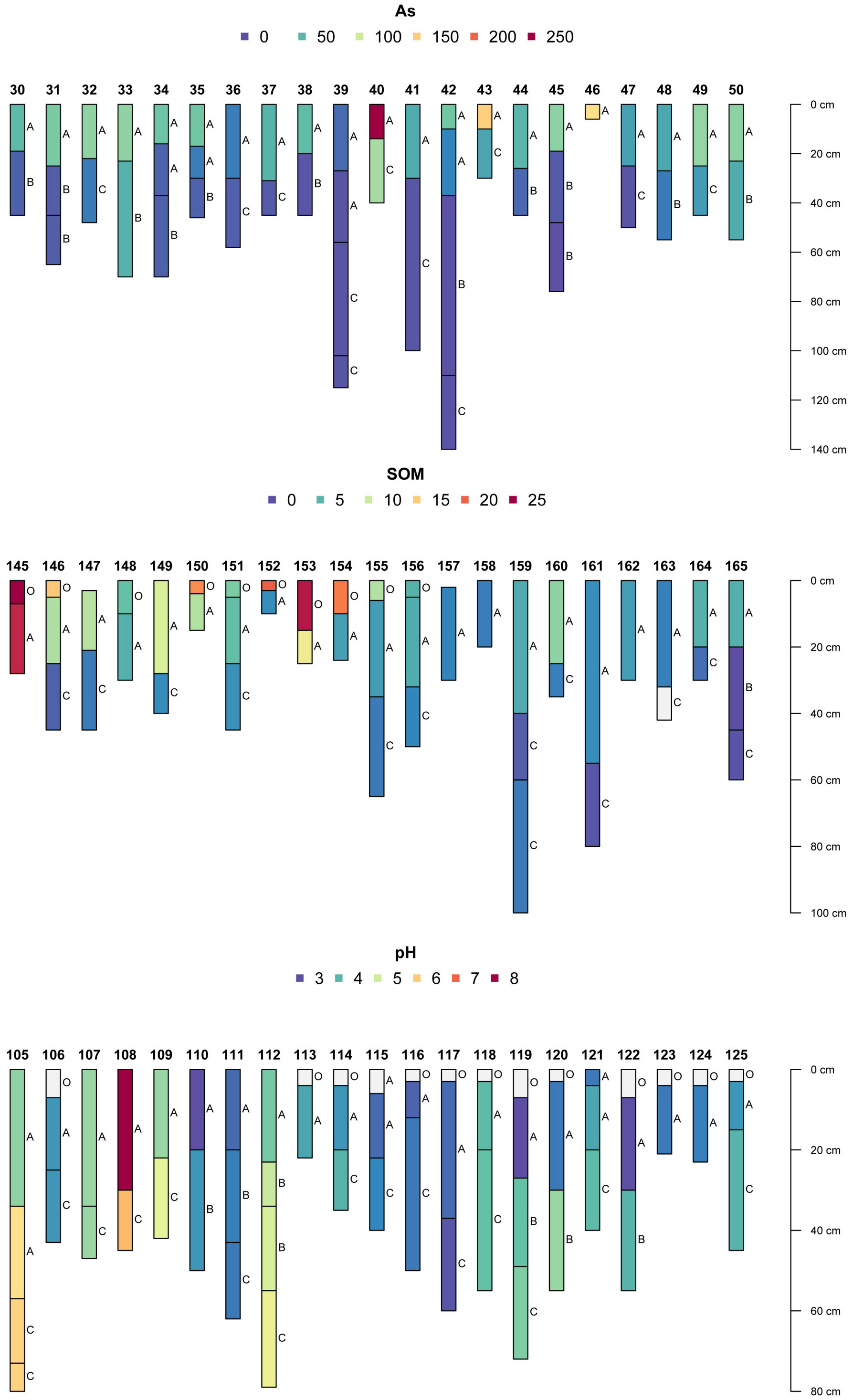


FIGURE 3.2: Soil profiles

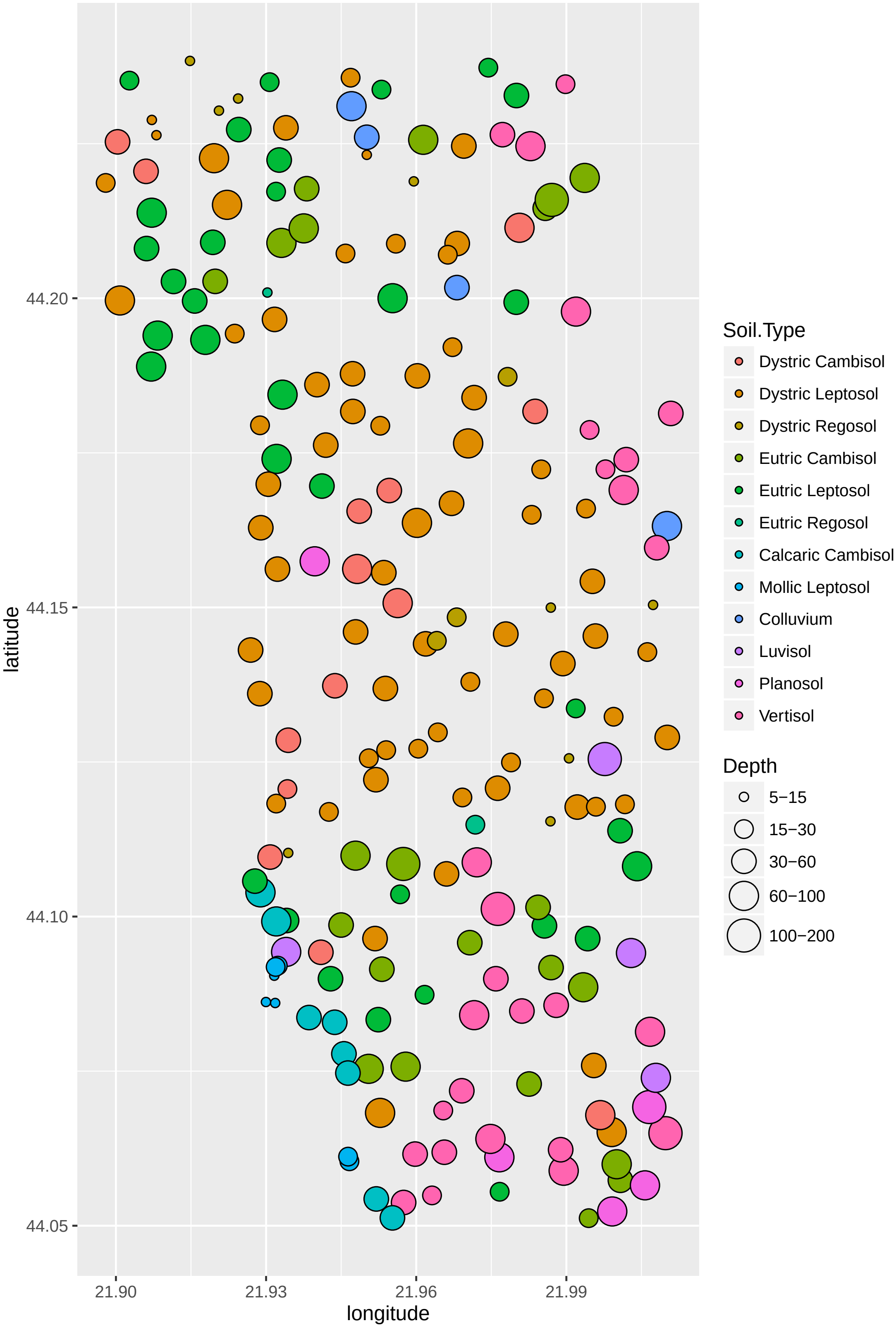


FIGURE 3.3: Profiles plotted against the soil type and depth

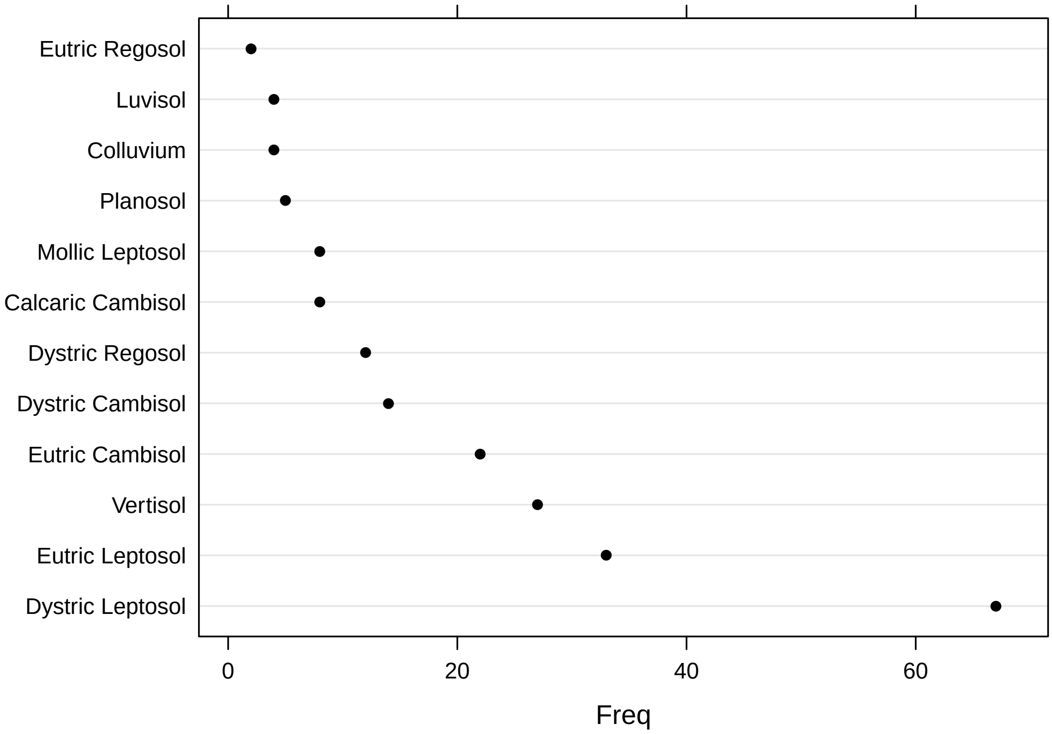


FIGURE 3.4: Number of profiles per soil type classes

Preliminary data analysis revealed some interesting distributional patterns in the data, typical for soil that has been exposed to the pronounced human influence for a long time. Figure 3.5 illustrates the depth-wise distribution of As, SOM and pH observations created by aqp R package (Beaudette et al., 2013b). The observations from all profiles were aggregated and summarized over 5 cm depth increments. As it is apparent, the As and SOM data are characterized by clear decreasing trend in mean with depth as well as by significantly higher variation in the upper soil layers which is displayed with the interquartile (blue-shaded) area. There can also be spotted the distinct breaking point at 30cm depth from which the variations appear to be more stable. On the other hand, pH appears to have varying mean followed by nearly constant variation along depth.

Higher variation of As and SOM in the upper soil layers indicate the strong influence of external factors on soil in this region. As it is generally known, the high SOM variation in the surface layers can be attributed to the complex influences of many environmental factors, such as climate conditions, topography, soil texture, land use, and other micro-scale factors that affect the surface soil layers (Parton et al., 1987; Burke et al., 1989). Similarly, the higher variation of As in the upper soil layers is most probably connected with long term smelting activity (Kovačević et al., 2010; Serbula et al., 2013, 2014).

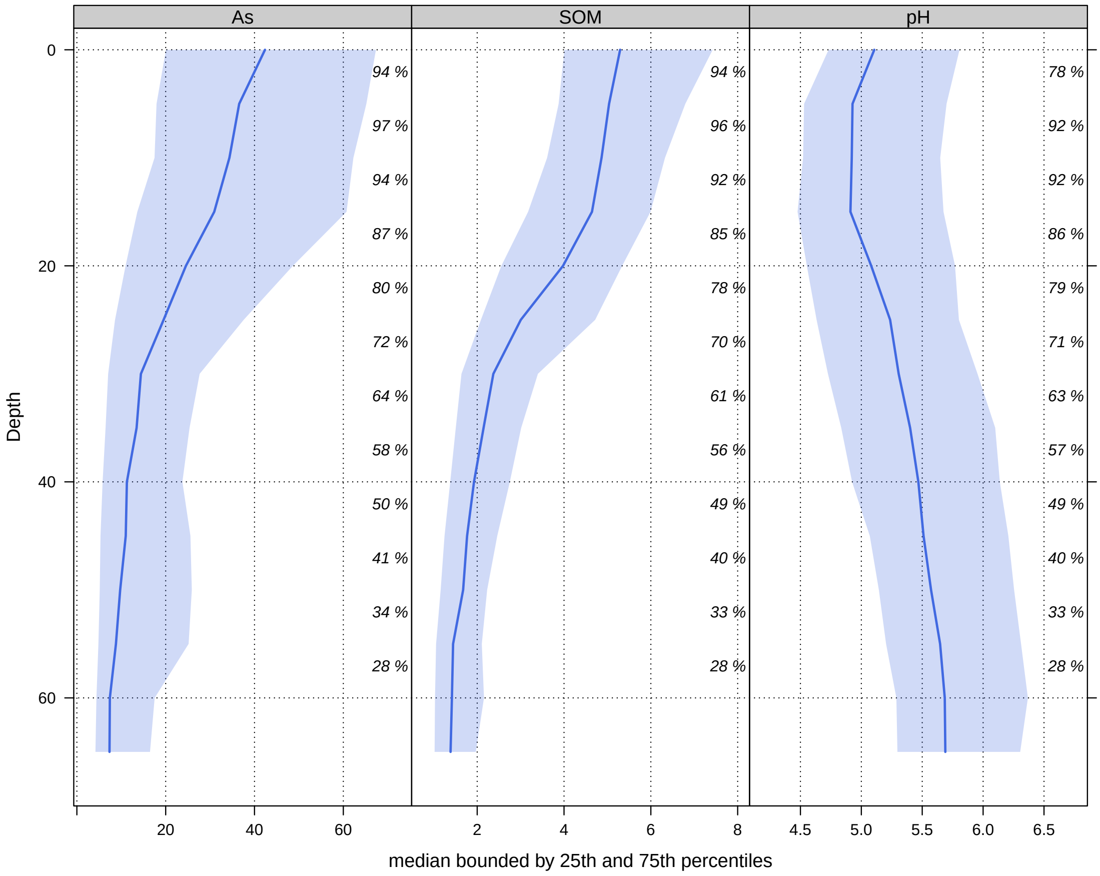


FIGURE 3.5: Depth-wise distribution of profile observation of As (left), SOM (middle) and pH (right)

Chapter 4

Layer-specific mapping of arsenic concentration by considering terrain exposure

This chapter constitutes a large excerpt from my manuscript entitled *Layer-specific As concentration modeling by considering terrain exposure* that has been submitted for publication in *Journal of Geochemical Exploration*.

4.1 Introduction

Without a doubt, industrial mining has significant consequences on the environment and human health (Unit, 2013). Spatial extension and the magnitude of soil pollution in mining areas are conditioned by many environmental factors such as climatic conditions, relief, human or mining activity, the soil type, and land use. In geostatistics, environmental factors are approximated by spatial covariates. These are mainly maps in raster format, but could also be the output of some existing models. For example, Goovaerts et al. (2008a) used an EPA Industrial Source Complex (ISC3) dispersion model EPA (1995) in combination with kriging and geostatistical simulation to delineate areas with high levels of dioxin TEQDF WHO98 in soil around an incineration plant. Their dispersion model explained 47.3% of the variance found in the soil TEQ data, leaving the residuals suitable for geostatistical analysis. Dispersion models like ISC3 can take a wide range of parameters into account that pertain to meteorological conditions, the local topography, and the characteristics of the source (e.g., emission rate, stack height and diameter, particle diameter etc.) (De Visscher, 2014). These parameters are often inaccessible for long-term pollution processes; therefore soil scientists have to deal with only a few known parameters that are often related to relative distances from the source of pollution, terrain topography or common meteorological parameters including prevailing wind direction and wind speed. Žibret and Šajn (2008) presented successful implementation of the power function with negative exponent to model how the level of heavy metal concentrations in the air and soil decreases in relation to incremental increases of the distance from the source of pollution. Saito and Goovaerts (2001) incorporated the knowledge of the position of a pollution source and deviations from major wind direction into a kriging system to map the spread of pollutants from a known source.

In mountainous or hilly areas, the spatial variation of wind-deposited materials is highly affected by terrain topography. It is generally known that the amounts of wind-deposited materials tend to be greater on areas that are more directly exposed to wind flux. This fact has inspired researchers to develop many topographic indices with the aim to quantify topographic exposure to wind (Antonić and Legović, 1999; Lindsay and Rothwell, 2008; Winstral et al., 2002; Winstral and Marks, 2002) . Generally, all topographic exposure indices are based on Digital Elevation Model (DEM) analysis and tend to determine whether a particular area is sheltered by a distant topographic obstacle or not. There are several studies where topographic exposure indices were successfully used to model the spatial patterns of snow depths (Erickson et al., 2005; Plattner et al., 2004; Winstral and Marks, 2002).

Antonić and Legović (1999) introduced the aspect of topographic exposure to wind in their exploration of environmental pollution studies. They proposed the new comprehensive index, referred to as the Exposure toward the Wind Flux (EWF). EWF can be conceptualized as the angle between a plane orthogonal to the wind and a plane that represents the local topography at a grid cell. They utilized EWF to estimate the direction of an unknown air pollution source.

In this study, different aspects of terrain exposure are considered in order to explain the complex spatial trend of Arsenic (As) concentration that was atmospherically-deposited from one of the largest Copper Mining and Smelting Complexes in Europe, Bor in Serbia. Several exposure parameters were created and employed as covariates within the 'Spline-Then-Krige' (STK) approach (Malone et al., 2009; Orton et al., 2016) for producing maps of As concentration at three standard soil depth layers (0-5cm, 5-15cm and 15-30cm). The created exposure parameters were grouped as follows: geometrical (proximity) exposure parameters and topographical exposure parameters. The distances to the source of pollution and angular deviations from prevailing wind direction were utilized to create geometrical (proximity) exposure parameters. Furthermore, topographical exposure was quantified by using DEM and two DEM derivates: modified EWF index and the Morphometric Protection Index (MPI). A modification of EWF was performed to account for the location of the pollution source with the aim to emphasize the effects of topographical exposure to the known source, and not just limiting the index to wind direction. This study primarily aims to evaluate the effectiveness of using different exposure parameters for mapping atmospherically-deposited Arsenic at different soil depth layers. Relative importance analysis was performed to access the individual contribution of each exposure parameter in the trend model for each depth layer. By analyzing the role of exposure parameters in As variation at different soil depth layers, the limit of significant influence of copper smelting in soil depth direction was assessed. This is the first study of its kind that evaluates the usage of different terrain exposure indices for mapping atmospherically-deposited pollutants from a known source, so far.

4.2 Data

The data used in this study are described in Section 3.2. The target variable is Arsenic concentration expressed in `mg/kg`. Generally, the data consist of 196 soil profiles that are randomly distributed over the entire study area (Figure 3.1). The soil samples were digested with concentrated HNO3 and then analyzed for As concentration using the iCAP 6300 ICP optical emission spectrometer (Thermo Electron Corporation, Cambridge, UK).

As shown in Figure 3.5, the distribution of arsenic concentration in soil is characterized by pronounced decreasing trend in median with depth, as well as with considerable higher variation in the upper soil layers. The abrupt change in the trend of median and inter-quartile range occurs at about 30 cm depth. The numbers of profiles that contribute to the estimated median values are shown in percentages on the right vertical axes. It can be seen that less than 50% of available profiles contribute to the estimated value for layers below 30 cm of depth. Due to the fact that the number of observations sharply decreases below the depth of 30 cm, the analysis was confined to the first three standard soil layers above this depth: 0-5 cm, 5-15 cm and 15-30 cm.

The presence of extreme values is an important characteristic of this data set. Figure 4.1 depicts the spatial pattern of observations from the first soil layer, allocated within the $4-th$ quartile (red circles: 80-326 mg/kg), with respect to the smelter location. The circle size depicted in the figure is proportional to the observed value. Terrain colors represent possible spatial coverage of plume dispersion from copper smelter.

High concentration and high variability in the As data at the upper soil layers combined with distinct differences between the upper and lower soil layers are generally considered to be the indicators of external factors that have a pronounced influence on the soil. As a result, a hypothesis can be established that the upper soil layers were indeed affected by a long term pollution process.

4.3 Terrain exposure

In this study, terrain exposure parameters aim to provide the numerical quantification of terrain exposure with regard to the location of the source of pollution, wind direction and topography. As mentioned above, the considered terrain exposure parameters have been divided into two groups: topographical exposure and geometrical (proximity) exposure. Table 4.1 summarizes the exposure parameters used in this study.

4.3.1 Topographic Exposure

There are many existing parameters that are suitable for explaining the topographic exposures to wind. An exhaustive review of existing topographic wind related parameters was

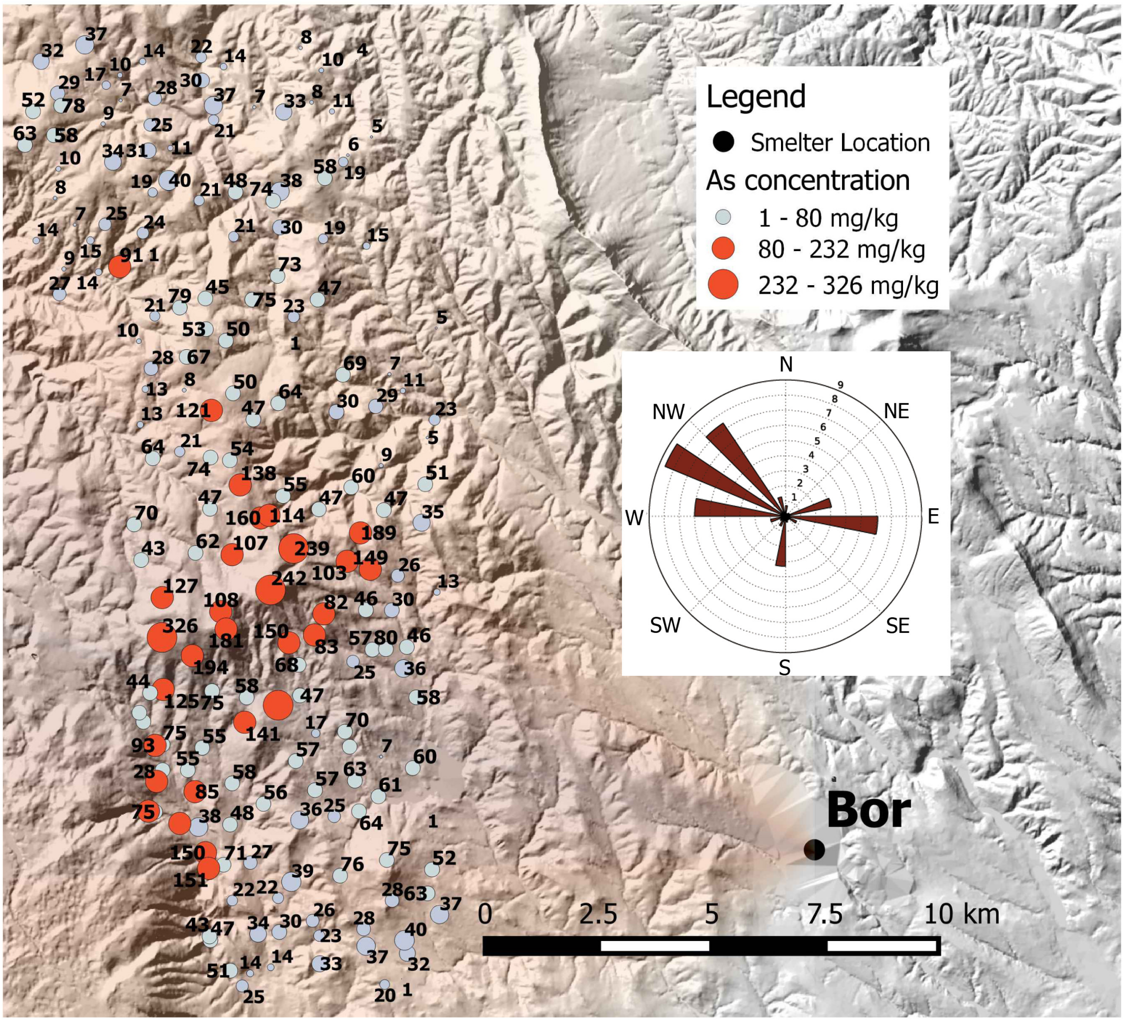


FIGURE 4.1: Spatial disposition of extreme values of observations relative to the location of smelter. Bor is in the lower-right corner; red circles represent the observations that belong to the fourth quartile.

TABLE 4.1: Exposure parameters used in this study

	Name	Abbrevation	Group	Range
1	Digital Elevation Model	DEM	Topographical	300-1045
2	Exposure toward the Source	ES	Topographical	0.75-1.33
3	Morphometric Protection Index	MPI	Topographical	0-0.70
4	Down-wind dilution	DD	Geometric	0.20-0.64
5	Cross-wind dilution	CD	Geometric	0.38-1

outlined in studies reported by Lindsay and Rothwell (2008); Winstral et al. (2002); Winstral and Marks (2002). In this study, a topographic exposure analysis was confined to the following three parameters: Digital Elevation Model (DEM), Morphometric Protection Index (MPI) and Exposure toward the Source of pollution (ES).

4.3.1.1 DEM

Considering the assumptions that areas on higher altitudes are more exposed than lowlands, elevation was selected as the first topographic exposure parameter. A high resolution DEM with a grid size of 20 m was created by digitizing contours from 1:25.000 scale topographic map sheets (Figure 4.2(a)). All other exposure parameters were computed based on this grid system.

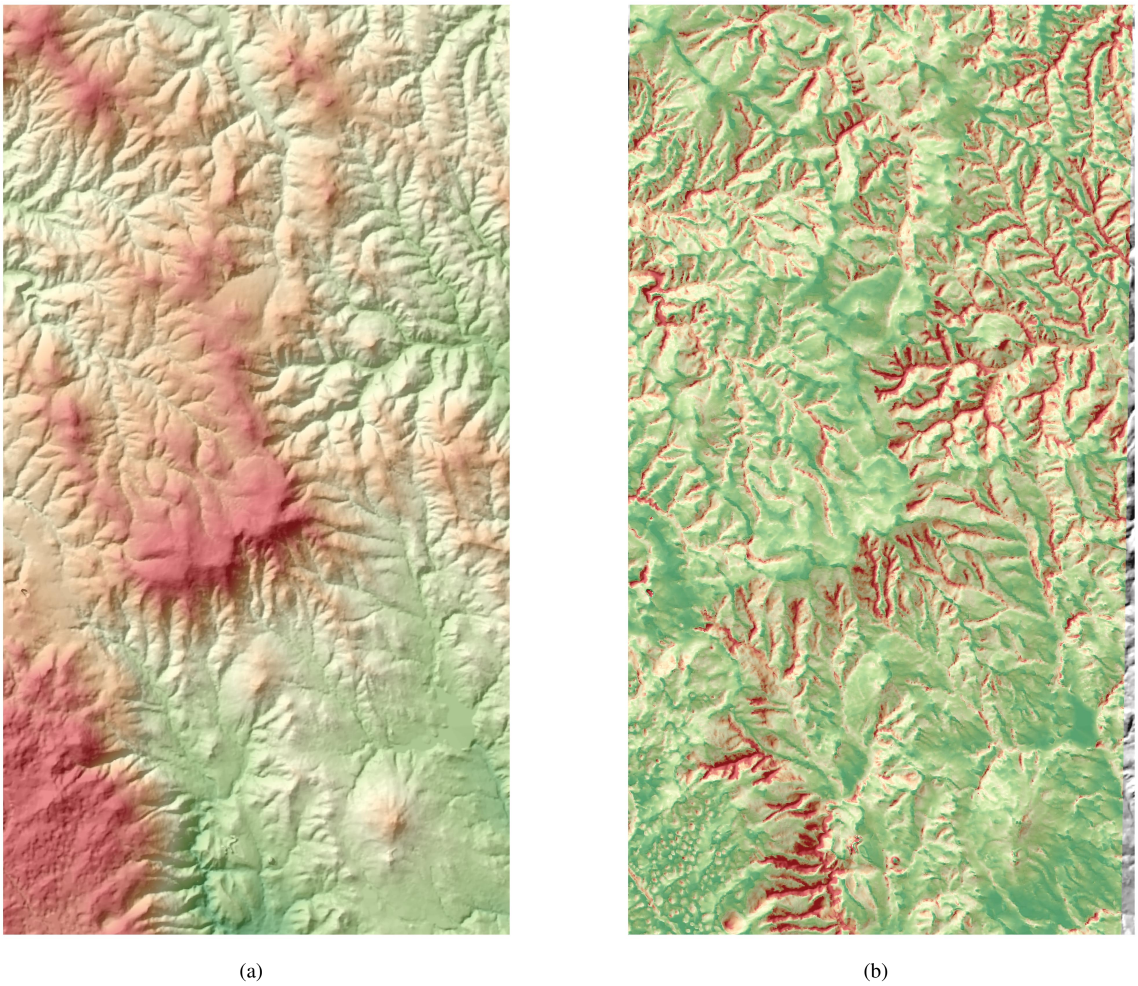
4.3.1.2 Morphometric Protection Index

The influence of local (neighboring) topography was considered by the Morphometric Protection Index (MPI) calculated for each grid cell. The calculation of MPI is equivalent to the "positive openness" described by Yokoyama et al. (2002). It considers neighboring grid cells of DEM in eight directions (cardinal and diagonal) up to a given distance (with 200 m radius), while searching for the maximum horizon angle in each direction. The final MPI for one cell represents the average value of eight maximum horizon angles and quantifies how the neighboring relief protects that cell. The map of MPI is given in Figure 4.2(a).

4.3.1.3 Exposure toward the Source of pollution

The effects of topography along wind direction were considered through the modified EWF measure. By definition, the EWF index combines two simple exposure parameters to quantify topographic exposure to the wind flux. These two parameters are the relative terrain aspect and the horizon angle:

$$EWF = \cos(\mu)\sin(\beta) + \sin(\mu)\cos(\beta)\cos(\delta - \omega) \tag{4.1}$$



(a)

(b)

FIGURE 4.2: a) DEM and b) MPI computed for the whole area

where μ represents the terrain slope, γ is the terrain aspect, δ is the azimuth of the dominant wind direction, and β is the horizon angle in the wind direction (Figure 4.3,a)).

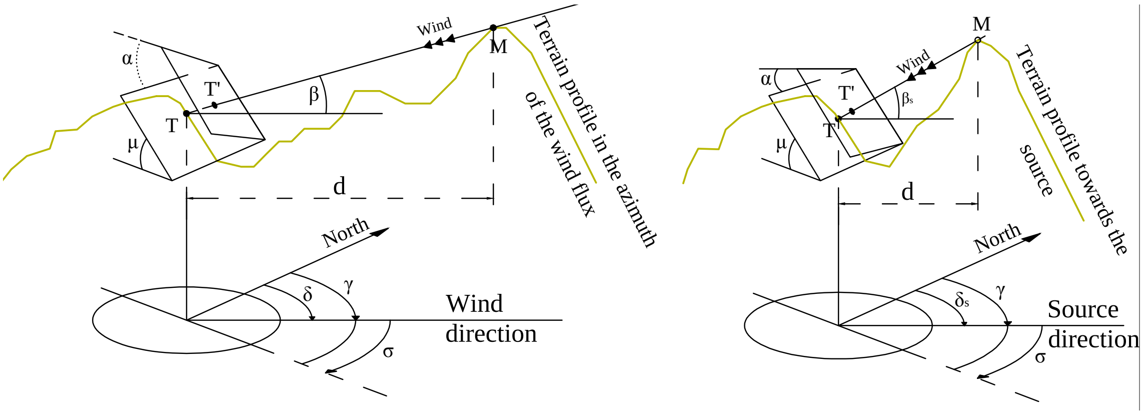
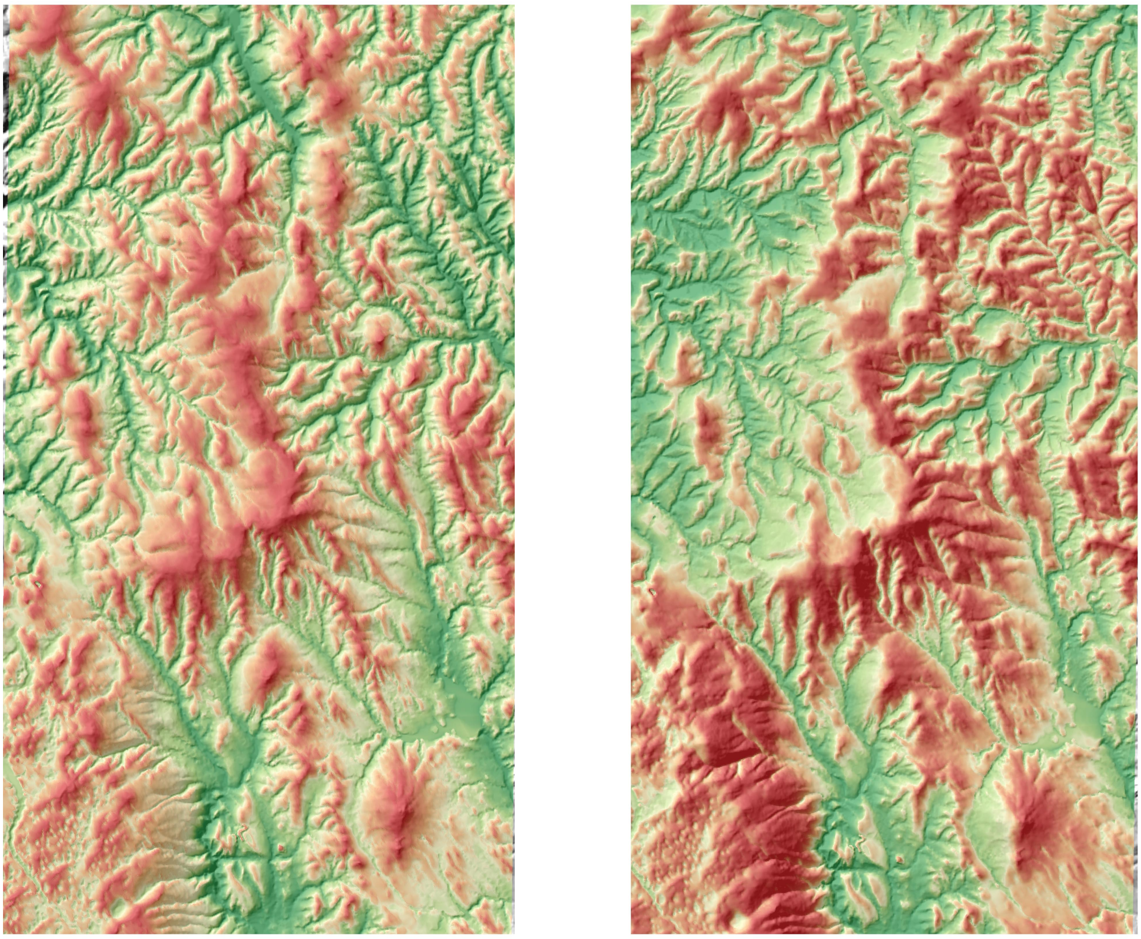

Wind
M
α
T'
β
μ
T
d
Terrain profile in the azimuth of the wind flux
North
δ
γ
Wind direction
σ
M
α
Wind
T'
βs
μ
T
d
Terrain profile towards the source
North
δs
γ
Source direction
σ

FIGURE 4.3: a) Graphical representation of EWF components. Terrain at the point T has the maximum slope μ and terrain aspect γ. σ represents relative terrain aspect at point T for a given azimuth of the wind flux δ . β is the horizon angle of the point T for a search distance d. α is the angle between regression plane through the terrain point T and the plane orthogonal to the wind. b) Equivalent graphical representation of the ES components. Index 's' denotes the source of pollution.

The relative terrain aspect represents the orientation of the local terrain plane in relation to the selected wind direction. This is the angle between the land-surface aspect and the wind direction bounded between 0°, indicating an exposed location, and 180°, indicating sheltered location. The horizon angle quantifies the effects of upwind topography searching for the maximum elevation angle along the direction of the prevailing wind flux. The search distance has a crucial effect on the horizon angle estimation. According to the definition of the horizon angle, a more exposed area is characterized by a negative horizon angle, whereas a sheltered area is characterized by a positive horizon angle. Horizon angle has been used as the basis for many subsequently devised parameters (Erickson et al., 2005; Winstral et al., 2002). Depending on the extent of the horizon angle search distance, EWF has been referred to the horizontal wind flux (zero search distance), or to the slope wind flux (search distance differs from zero).

The standard EWF index presumes a constant direction of wind flux, which participates in two terms of its equation: the relative aspect and the horizon angle. Taking into account that the contaminated air flux starts from the one copper smelter stack and expands towards the explored region, it is assumed that: (1) the local terrain plane facing the source is more exposed to pollution than planes that are not; (2) the topographic obstacles founded within the direction of the source have a greater effect on the redistribution of pollutants than the obstacles founded strictly in the upwind direction. Based on these assumptions, the EWF index was calculated for each grid cell with the adjustable wind direction. More specifically, the wind direction was defined as the azimuth between each grid cell and the source of pollution. In this regard, the relative aspect becomes the angular distance between the land-surface aspect and direction to the source. At the same time, the horizon angle search path is also directed towards the source (Figure 4.3,b). This new parameter was denoted as Exposure toward the Source (ES). Figure 4.4 shows the maps of EWF and ES parameters values for the whole area of interest.



(a) (b)

FIGURE 4.4: a) EWF and b) ES indices computed for the whole area

4.3.2 Geometric (Proximity) exposure

The creation of geometric (proximity) exposure parameters was inspired by a dilution mechanism considered in the Gaussian dispersion model (Gaussian plume model). It assumes that the dilution of plume emitted in the atmosphere could be considered in three directions: downwind, crosswind and vertical (De Visscher, 2014). The downwind plume dilution is the result of mixing the plume with the ambient air, while the dilution in the cross-wind direction is a result of a large number of negligible effects related to atmospheric motions. Taking into account all of the assumptions mentioned before, we presumed that areas are more geometrically exposed if they are closer to the copper smelter and/or to the prevailing wind direction. Therefore, the effects of dilution in downwind and crosswind directions in this study were approximated by Downwind Dilution (DD) and Crosswind Dilution (CD) parameters computed for each grid cell. These were modeled using a negative-exponential function, where the exponents were the distance to the smelter for DD and the directional departure from dominant wind direction for CD. The wind rose (Figure 4.1) shows that the prevailing winds blow from east and northwest directions. However, in order to found the CD which is most correlated with the observed As data, wind direction was determined based on the correlation analysis between the first soil layer data and the CD parameter computed for several major wind directions, in the range of 90° ± 30°, along with increments of 5°. Finally, the wind direction of 105° was found to be the most correlated with the observed data. Figure 4.5 depicts the graphical representation of CD and DD parameters.

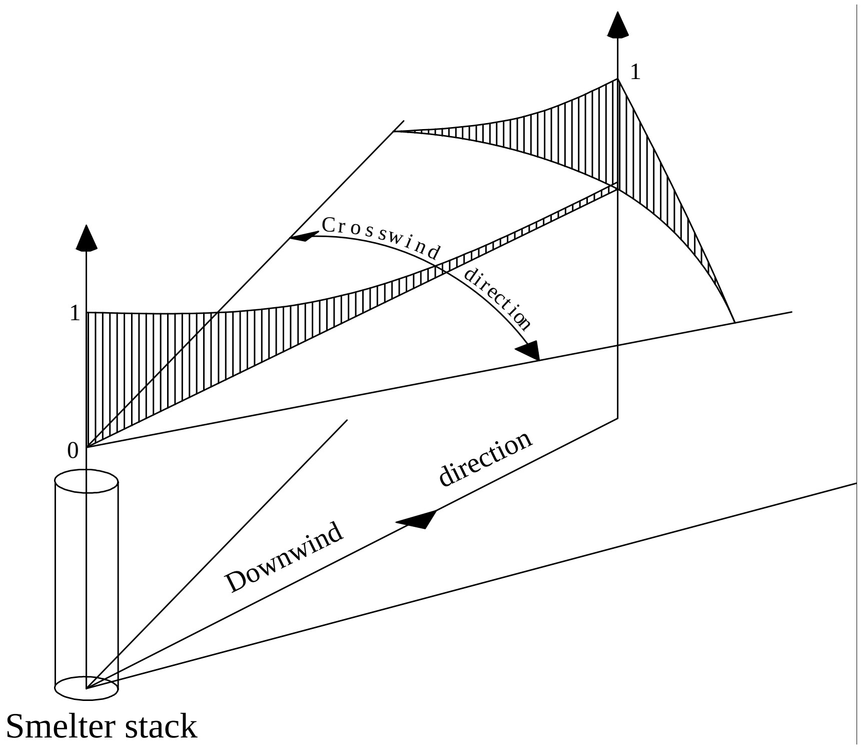


FIGURE 4.5: Graphical representation of geometrical (proximity) measures.

4.4 Geostatistical Mapping

In this study, Spline-Then-Krige (STK) (Section 2.5.2) method was used for mapping As concentration at three different depths. Generally, the STK involves conversion of profile data into a continuous form by particular depth function, computing mean value for specific depth interval and, finally, interpolating interval-specific mean values over the entire area.

4.4.1 Vertical variation modeling

Variation in the soil profile was modeled by equal-area spline function, (see Section 2.5.1. Equal-area spline function implies continuous vertical variation. This can be expected, considering the fact that over one hundred years of copper production in Bor, vertical leaching of deposited toxic materials in soil has certainly occurred. This function provides that, for each sampling layer (soil horizon), the average of the spline function equals the measured value for the horizon, i.e. the area above and below the fitted spline in any horizon are equal. Figure 4.6 depicts an example of a fitted spline to the measured As data from profile No. 119. The colored horizontal bars represent the measured As concentration at different horizons (each bar corresponds to one horizon), while the vertical curve represents the equal-area spline depth function fitted to these data. In order to obtain the As concentration related to the selected fixed depth intervals (0-5 cm, 5-15 cm and 15-30 cm), the spline function was averaged within these intervals. These intervals correspond to the standard soil depth intervals specified in *GlobalSoilMap* specifications (Arrouays et al., 2014). Soil profiles containing only one sample layer were not modeled. Instead, they were considered as profiles with constant As concentration up to the depth of the sampling horizon. The equal-area spline function was fitted via the *mp.spline* function implemented in the GSIF R package Hengl (2015).

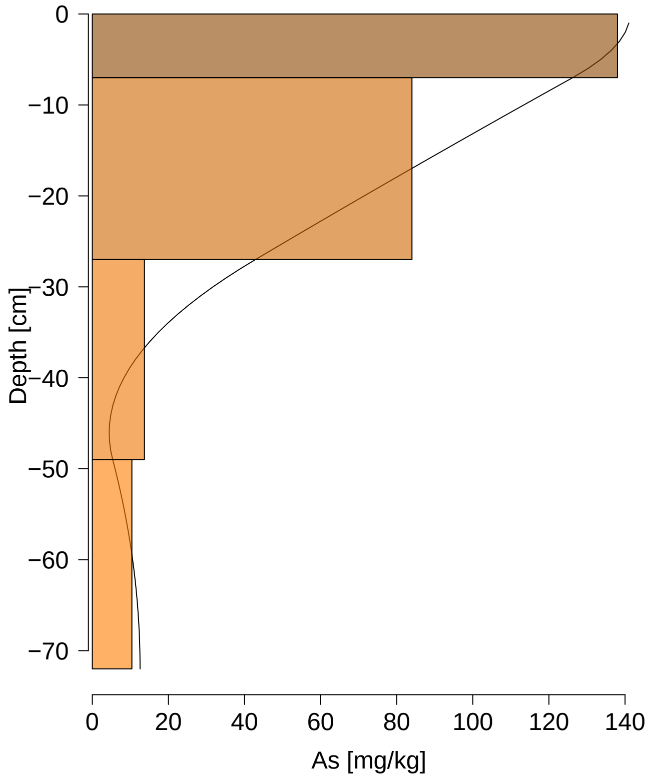


FIGURE 4.6: Equal-Area spline depth function fitted to the data from profile No. 119.

4.4.2 Trend Analysis and Spatial Prediction

The first task in trend analysis was to identify the type of relationship between the As data and exposure parameters. It is convenient to represent the relationship between the target variable and the covariates using a linear model (Pebesma, 2006; Hengl et al., 2007; Kilibarda et al., 2014). The adequacy of this specification was checked by examining the residual plots. Prior to model fitting, the exposure parameter values at each profile location were extracted and joined to the spline-predicted As values for each depth increment. The interaction effects between each pair of exposure parameters were also considered to be included in the model. By doing this, it was enabled that the effects of one exposure parameter depends on the value of the other exposure parameter.

Model selection was conducted by performing stepwise linear regression analysis using the Akaike information criterion (AIC) (Akaike, 1974) as a selection criterion, (see Section 2.2.1). The complete process of model selection was conducted on data from the first soil layer, considering the fact that the effects of atmospheric pollution are the most pronounced near the terrain surface.

An integral part of this trend analysis was to determine the contribution of each individual exposure parameter, as well as the interactions between them, to the overall prediction accuracy. This was achieved by computing the measures for the relative importance of predictors. It is important to note that the term "predictor" is associated to the independent model variable, which could refer to the main effect or interaction effect as well. There are several measures for the relative importance of predictors in linear modeling theory that are all available in the *relaimpo* R package (Grömping and Others, 2006). These measures provide the information about the individual contribution of each predictor to the portion of the explained variance (R^2). The most comprehensive and recommended measure, called LMG, was used. This measure was first proposed by Lindeman et al. (1980).

4.4.3 Spatial Prediction

Regression Kriging (RK) was adopted as a general statistical framework for spatial prediction. Regression kriging combines two conceptually different techniques, regression for trend estimation, and simple or ordinary kriging for the interpolation of stochastic residuals (Hengl et al., 2007), (see Section 2.4.2).

For a given trend model and residual variogram, the prediction of a target variable at an unsampled location s_0 is obtained by:

$$\hat{z}(s_0) = \sum_{k=0}^{p} \hat{\beta} \cdot x_k(s_0) + \sum_{i=1}^{n} \lambda_i \left[z(s_i) - \sum_{k=0}^{p} \hat{\beta} \cdot x_k(s_0) \right] \tag{4.2}$$

where $z(s_i)$ represents the observed values at the neighboring locations s_i, $\hat{\beta}$ represents the estimated trend model coefficients, $x_k(s_0)$ are the known value of covariates at the predicted location, $x_k(s_i)$ are the known value of covariates at the location s_i, and λ_i are the kriging weights.

Prediction accuracy was evaluated based on the leave-one-out cross-validation procedure. The following common statistical measures were calculated to evaluate the prediction accuracy: Mean Error (ME), Root Mean Squared Error (RMSE) and R^2 (see Section 2.2.1).

4.5 Results and Discussion

Table 4.2 shows the common descriptive statistics measures computed for aggregated profile data divided into a total of six standard depth increments. It is obvious that the measures of central tendency (mean, median) systematically decrease by depth. The mean values in the upper layers are almost double the mean value from the layers below the 30 cm depth. This trend is even more pronounced when comparing median values. Decreases in mean (median) values are accompanied with decreases in variation (IQR and standard deviation), which result in small changes in coefficients of variation. The presence of extreme observed values, even in the deeper layers, causes considerable differences between calculated mean and median values. The same statistical quantities computed with the data predicted by the equal-area spline function, and averaged over the same depth increments, reveal that overall distribution remains almost unchanged after the transformation (Table 4.3).

TABLE 4.2: Depth-wise summary of observations

depth	min	1^{st} quartile	mean	mean.sd	median	3^{rd} quartile	max	IQR	sd	CV	obs
$0-5\,cm$	4.00	18.80	51.22	1.64	39.00	65.70	328.00	46.90	49.26	0.96	195
$5-15\,cm$	4.00	17.20	45.21	1.03	33.20	57.90	311.00	40.70	45.04	1.00	195
$15-30\,cm$	3.10	10.70	35.53	0.80	23.70	46.50	311.00	35.80	40.02	1.13	181
$30-60\,cm$	2.40	5.80	20.90	0.50	11.20	24.30	246.00	18.50	26.85	1.28	135
$60-100\,cm$	1.80	3.90	19.68	1.28	6.00	11.90	228.00	8.00	44.14	2.24	52
$100-120\,cm$	2.70	4.50	6.99	0.30	5.80	9.60	10.10	5.10	2.84	0.41	6

TABLE 4.3: Depth-wise summary of spline-predicted data

depth	min	1^{st} quartile	mean	mean.sd	median	3^{rd} quartile	max	IQR	sd	CV	obs
$0-5\,cm$	1.00	20.96	52.53	3.50	41.94	67.62	326.15	46.66	48.93	0.93	195
$5-15\,cm$	2.19	18.40	46.81	3.24	35.36	60.40	313.87	42.00	45.23	0.97	195
$15-30\,cm$	1.46	12.76	36.49	3.03	25.54	44.06	305.86	31.31	40.81	1.12	181
$30-60\,cm$	1.63	5.74	21.13	2.62	10.88	23.78	243.10	18.04	30.42	1.44	135
$60-100\,cm$	1.00	3.43	17.30	4.94	6.91	13.16	225.65	9.73	35.65	2.06	52
$100-120\,cm$	3.89	4.46	7.20	1.26	7.10	9.68	11.01	5.22	3.08	0.43	6

4.5.1 Trend analysis and spatial prediction

In order to examine if the assumptions which justify the usage of linear regression are met, the residual plots were created (Figure 4.7). For this purpose, only the model fitted to the first layer data (0-5 cm) was selected. First five graphs (except graph in lower-right corner) depict the relation between residuals and each exposure parameters separately, while the last graph shows the residuals against the fitted values. The lack of systematic curvatures in the first five graphs confirms the linearity of the relationship between As data and exposure parameters. The presence of the increasing variation of residuals, with the level of fitted values depicted in the last graph, indicates the moderate violation of the assumption of constant error variance. The final sub-group of predictors was selected by combining the backward and the forward step-wise regression procedure. According to the step-wise regression analysis, all exposure parameters were included in the final model. In addition, the interaction effects between ES and DEM, as well as between DEM and CD, were also found to be useful for the As prediction. Considering this, the final model for each soil layer has the following formulation:

$$\begin{aligned} m_{As}(s_i) &= \beta_1 \cdot DEM(s_i) + \beta_2 \cdot ES(s_i) + \beta_3 \cdot MPI(s_i) + \beta_4 \cdot DD(s_i) + \beta_5 \cdot CD(s_i) \\ &\quad + \beta_6 \cdot CD(s_i) \cdot DEM(s_i) + \beta_7 \cdot DEM(s_i) \cdot CD(s_i) \end{aligned} \tag{4.3}$$

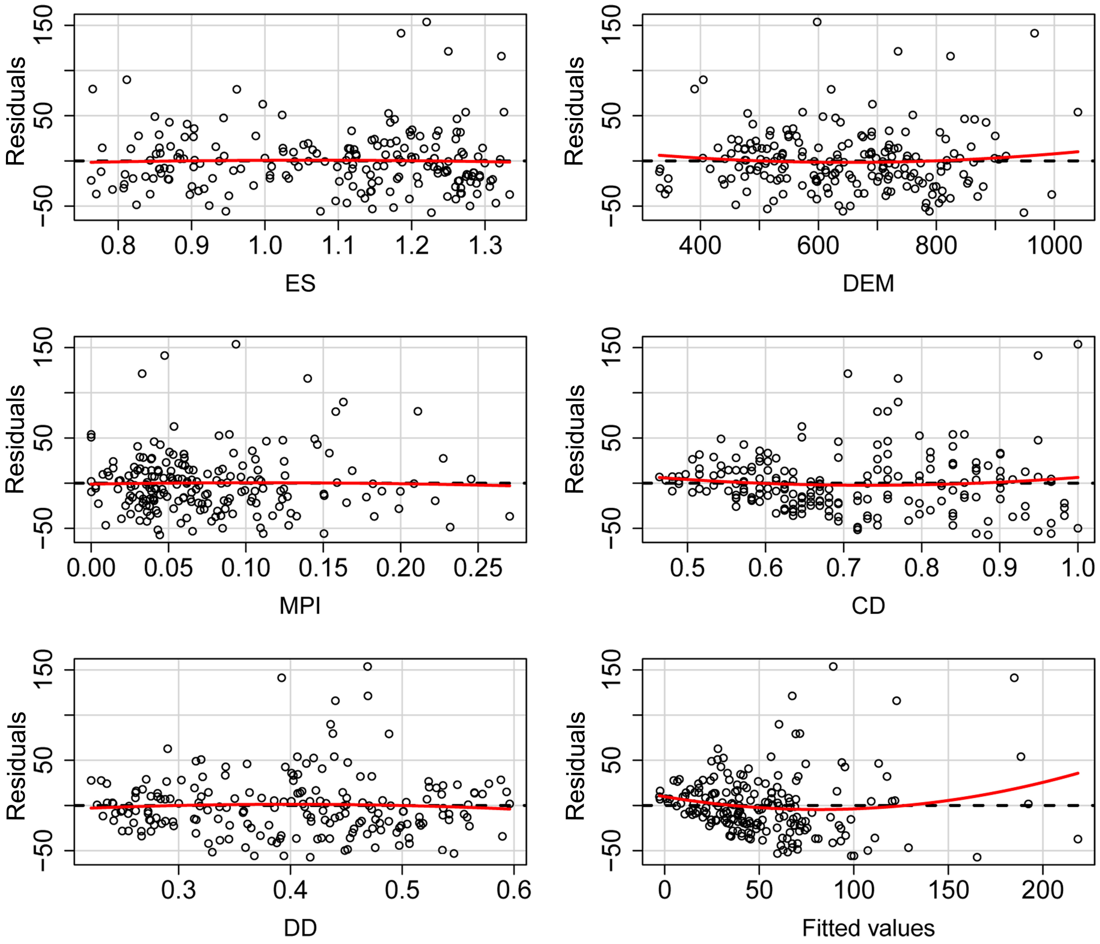


FIGURE 4.7: Residual Plots for initial regression model fitted on data from the first soil layer (0-5 cm).

The final model parameters for all three layers were obtained by using the GLS method within the RK algorithm given above. The estimated model coefficients, together with accuracy measures for each soil layer, are given in Table 4.4. The values shown in brackets represent the corresponding OLS coefficient estimates. Considerable differences between OLS and GLS estimates indicate the existence of significant spatial clustering between the observations in each layer. The asterisks following the estimated coefficients indicate the level of statistical significance according to the Wald test. Statistical significance for each predictor, except for DD, was confirmed for each soil layer. As expected, the variance in As data explained by trend models decreased with depth. R^2 values ranged from 0.52, for the first layer, to the 0.49 and 0.35 for the second and third layer. This is not so evident in RMSE which gains a marginally smaller value for deeper soil layers. This is not surprising, considering that the R^2 represents the relative measure, whereas the RMSE value represents the absolute measure of fit.

TABLE 4.4: The trend models coefficients and associated statistics

	Trend models:		
	$0-5\ cm$	$5-15\ cm$	$15-30\ cm$
	(1)	(2)	(3)
Constant	440.18***	394.96***	328.93***
	(369.88)	(385.96)	(327.38)
ES	$-$285.37***	$-$235.21***	$-$192.61**
	(274.13)	(251.37)	(224.06)
DEM	$-$0.85***	$-$0.79***	$-$0.64***
	(0.76)	(0.78)	(0.63)
MPI	115.63**	142.68***	144.30***
	(147.05)	(119.28)	(112.00)
CD	$-$308.27**	$-$269.65**	$-$251.64**
	(209.49)	(220.80)	(204.37)
DD	80.41	51.05	74.03
	(76.53)	(47.46)	(79.58)
ES:DEM	0.44***	0.39***	0.32***
	(0.45)	(0.42)	(0.36)
DEM:CD	0.73***	0.65***	0.52***
	(0.58)	(0.58)	(0.45)
Observations	195	195	181
R^2	0.52	0.49	0.35
AIC	1,940.59	1,919.34	1,793.65
RMSE	33.7	32.2	32.7

Note: *p<0.1; **p<0.05; ***p<0.01

As the DEM participates in each interaction effect, trend model coefficients for ES and CD can vary according to the level of altitudes. To illustrate this effect, Table 4.5 lists the trend model coefficients from the first model, estimated for four different levels of altitude: 400 m, 600 m, 800 m and 1000 m.

TABLE 4.5: Changing ES and CD coefficients according to the altitude level

Elevation	ES	CD
400 m	-109.37	-16.27
600 m	-21.37	129.73
800 m	66.63	275.73
1000 m	154.63	421.73

Results for Relative Importance analysis are depicted in Figure 4.8. It can be noted that the CD appeared as the dominant predictor in each model. For the first two layers, it participated in R^2 with the portion greater than 40%. It was followed by DEM and two interaction terms, while ES, MPI and DD showed considerably poorer predictive contribution.

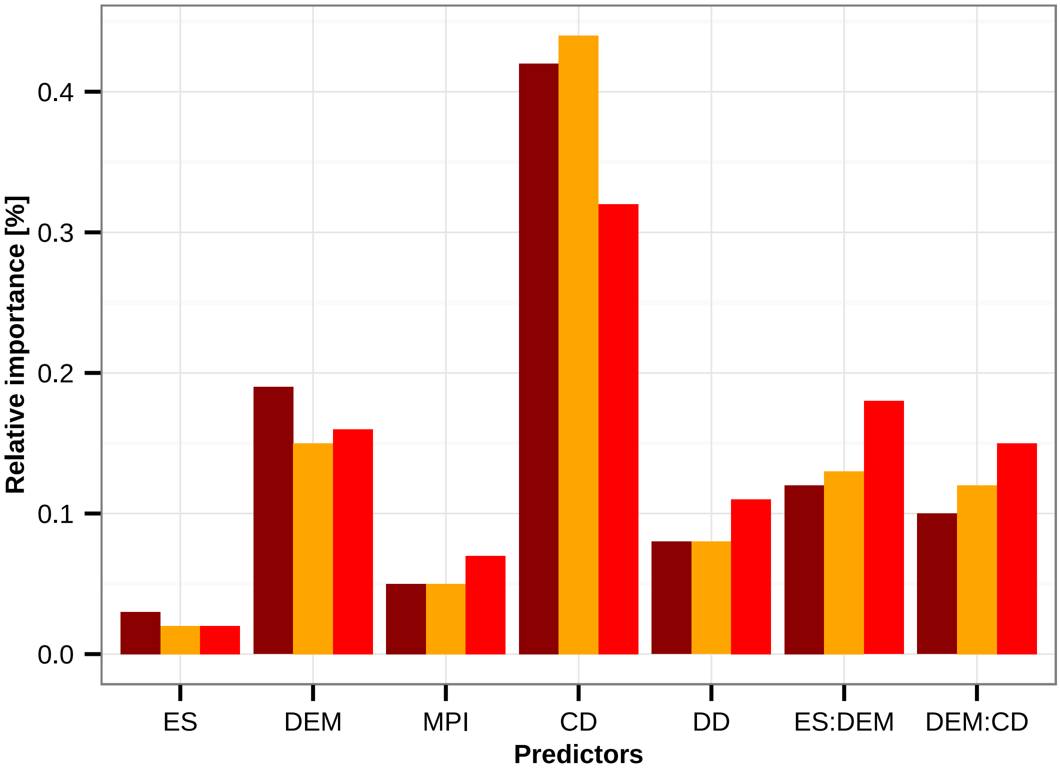
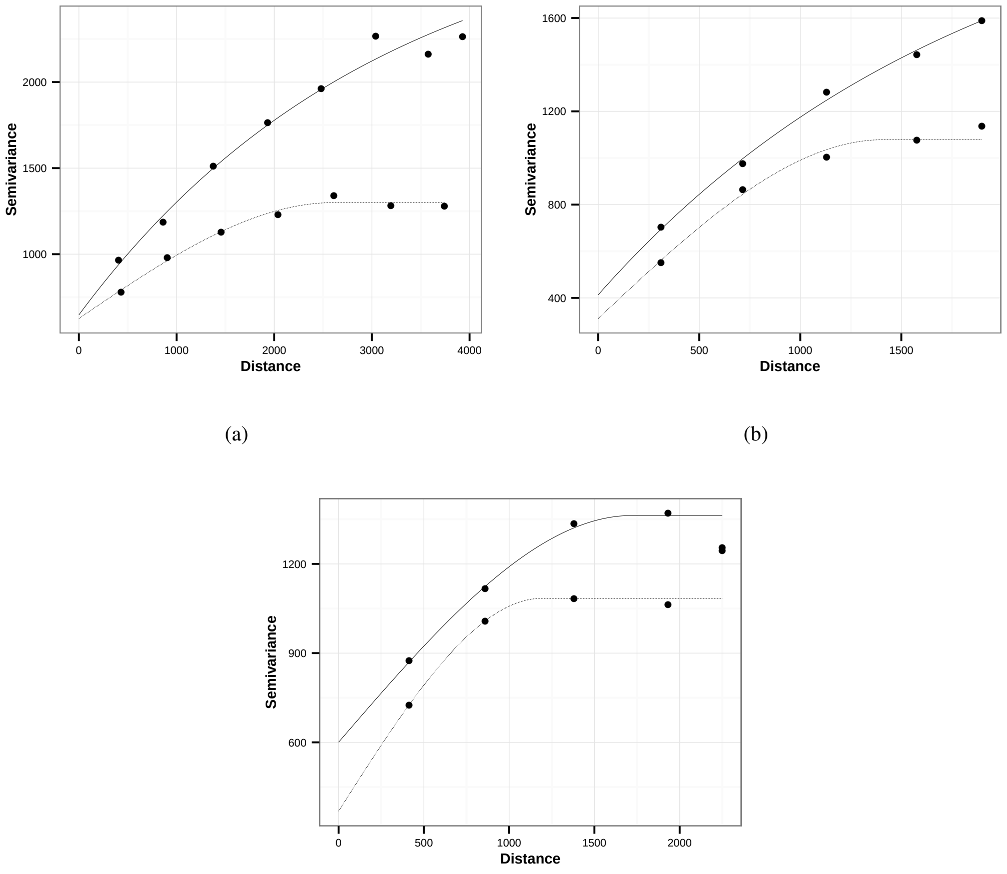


FIGURE 4.8: Relative importance of predictors for each soil layer (0-5 cm - dark red, 5-15 cm - orange, 15-30 cm - red).

Once the trend model was defined for all soil layers, the obtained residuals were then analyzed for spatial dependence. The presence of spatial dependence in residuals justifies the usage of kriging to improve the accuracy of the prediction. The lag increment was set to 550 m, which provided a sufficient number of point pairs for a reliable variogram estimation. The effects of trend removal on the spatial dependence structure are shown in Figure 4.9. Typically, the presence of a spatial trend in the observed data causes the monotonically increasing differences in the data as the separation increases, which is reflected in the experimental variogram that never reaches the sill. On the other hand, the residual variograms reflect the spatially correlated random effects more accurately, reaching the sill at a particular distance. This is more pronounced in the first soil layer, where the trend removal has the greatest effect (Figure 4.9(a)). This result confirms the fact that the atmospherically deposited spatial trend is more expressed in the soil surface layer. Generally, the sill variance and the nugget variances are substantially reduced in each soil layer. The differences between the two variograms are smaller in the deeper soil layers. It is also important to note a considerable decrease of the range parameter in the next two layers, indicating abrupt changes in spatial correlation over the soil depth (Figure 4.9(b) and Figure 4.9(c)). Similarly, the nugget variance for the first soil layer is also substantially higher than in the second two layers. This might be due to a higher variation in As data in the surface soil.



(a)

(b)

(c)

FIGURE 4.9: Omnidirectional variogram models for observed data (black line) and residuals (gray line) for all soil layers: (a) 0-5 cm; b) 5-15 cm; c) 15-30 cm

The final prediction accuracy measures, together with residual variogram parameters are reported in Table 4.6. Kriging interpolation slightly improved the trend model performance in each soil layer, while the accuracy between soil layers remained almost unchanged. The absolute prediction accuracy (RMSE) still remains almost equal for each soil layer. A decrease in accuracy was replicated in R^2 values ranging from 0.55 for the surface soil layer to 0.36 for the deepest layer. This is also supported with the coefficients of variation (CV) of predicted values, which take the values 0.6, 0.7 and 0.9 for first, second and third soil layers respectively. The mean errors indicate a negatively biased prediction for all layers. Considerable lower accuracy obtained for the third layer could also be noted. This might be due to the fact that a significant part of systematic variation still remains unexplained by the trend model. The obtained results are comparable with results outlined in similar studies previously reported by Adhikari et al. (2012, 2014); Goovaerts et al. (2008a); Saito and Goovaerts (2001); Lacoste et al. (2014) Moreover, the obtained results are in line with the statement given in Beckett and Webster (1971) that values the of R^2 higher than 0.7 are unusual, and the values of R^2<0.5 are quite common in soil attribute predictions.

TABLE 4.6: Variogram parameters and final prediction accuracy indices.

	Nugget	Sill	Range	Nugget/Sill	R^2	RMSE	ME	CV
$0-5\ cm$	626.35	1,299.88	2,613.21	48.18	0.55	32.75	−0.23	0.62
$5-15\ cm$	311.65	1,078.30	1,410.76	28.90	0.51	31.70	−0.26	0.67
$15-30\ cm$	368.49	1,084.21	1,192.19	33.99	0.36	32.51	−0.23	0.89

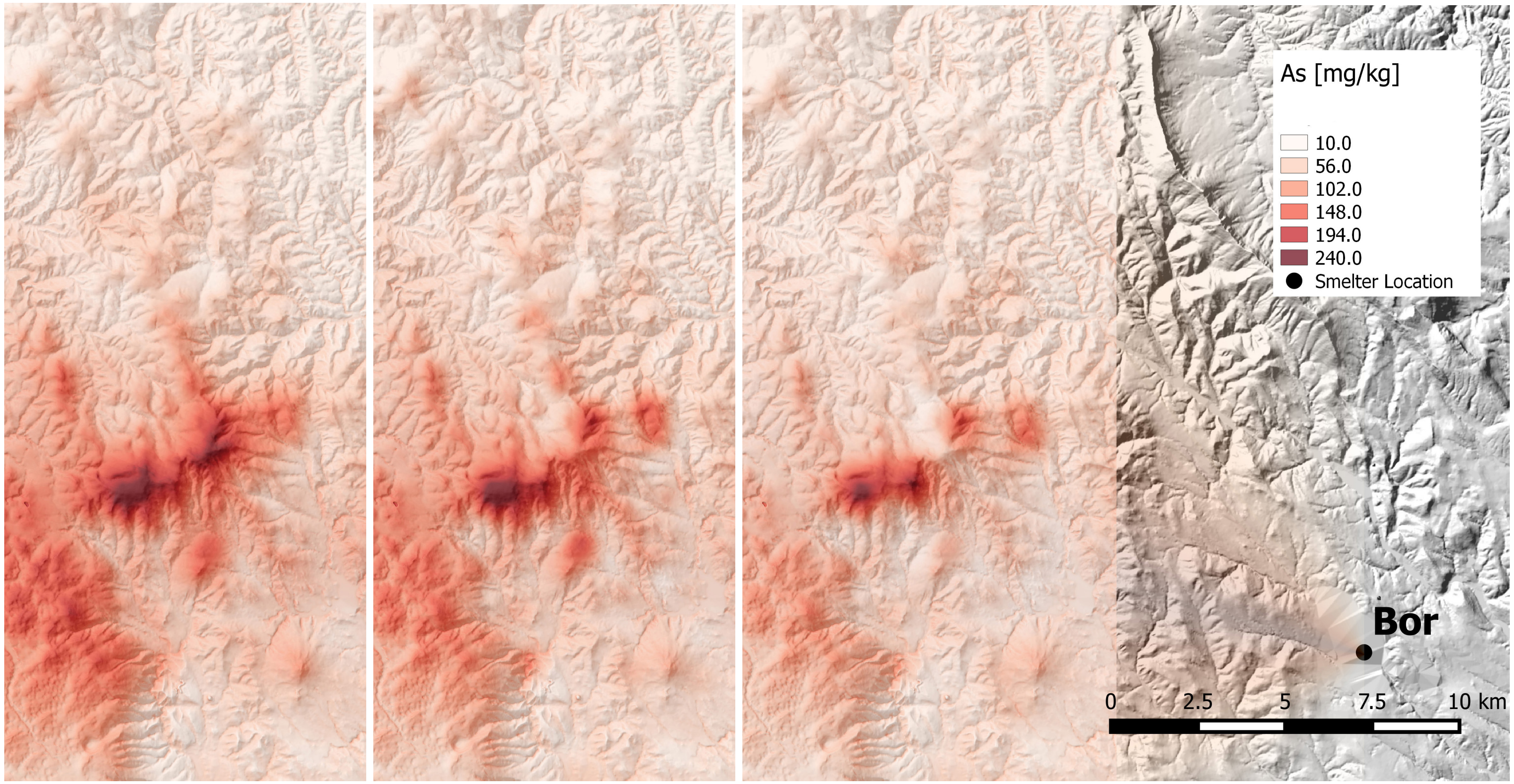
The maps of final prediction for all layers are displayed in Figure 4.10. The exposed area with high As concentration, in the central part, dominates in all soil layers. The mean predicted value ranged from 58.1 mg/kg for the first soil layer to the 51.8 mg/kg and 41.6 mg/kg for the second and third layer, respectively. The far Southwestern part of the mapped area is also regarded as a highly contaminated area. Topographically, this area is characterized by a downhill front that is directly exposed to the smelter. However, the lack of observations in this area makes predictions unverifiable. Due to its location and topographic configuration, this area could be suitable for additional sampling and validation for these models.

Interactive Web-based maps were also created in order to obtain a better insight into the predicted spatial distribution of As concentration. The R package plotGoogleMaps (Kilibarda and Bajat, 2012) was used for creating these maps. They are available as interactive maps in HTML format at the web page http://osgl.grf.bg.ac.rs/materials/Bor. In addition, the same maps in KML format, interactive point based maps, as well as background data are also available at the same web page.

4.6 Conclusion

This paper reviews a method for geostatistical mapping of atmospherically-deposited pollutants from a known source, by considering terrain exposure. The methodology applied is based on the so-called Spline-Then-Krige approach, which enables the production of a suite of maps for different soil depths. The exposure parameters were created to explain two different aspects of terrain exposure: geometrical (proximity) and topographical exposure. Based on the obtained results, the main conclusion can be drawn as follows:

1. The equal-area spline depth function provided a reliable estimate of continuous vertical distribution of As data.
2. Stepwise regression analysis confirmed the predictive capability of all of the designed exposure parameters, as well as of the two interactions: ES:DEM and DEM:CD. This confirmed the hypothesis that there is an association between spatial spreading of Arsenic from the copper smelter in Bor and terrain exposure parameters.
3. The trend model showed good overall accuracy for all soil layers. The highest accuracy was obtained for the surface soil layer, where the model explained 52% of data variation. The trend model explained 49% of variations for the second layer, and 35% for the third layer.
4. The relative importance analysis showed that the trend models at each depth were highly controlled by the CD and DEM. Significant influences of interaction effects between ES and DEM, as well as between DEM and CD, at each depth, indicate the importance of considering a more general model that includes interactions between exposure parameters.
5. The residual spatial dependence showed significant differences in structure between surface and other soil layers, indicating different effects of trend removal.



(a) 0-5 cm

(b) 5-15 cm

(c) 15-30 cm

FIGURE 4.10: Maps of final predictions of As concentration.

6. The kriging interpolation improved the regression accuracy for all three layers with R^2 ranging from 0.36 for the deepest layer to the 0.55 for the surface soil layer.

7. The relatively high RMSE values that follow the prediction at each soil layer indicates that a great portion of Arsenic data variation remains unexplained by the trend models, which implies that other factors, in addition to the wind-driven process, affect the Arsenic spatial distribution. However, in a situation when the wind indeed has an important role on spatial distribution of soil pollutants, the integration of topographic exposure parameters could be useful for prediction, even at a deeper soil layers.

8. The prediction maps show that approximately 78% of the mapped area is above the allowable concentration limits for agricultural soils in Serbia (As<25 mg/kg, Regulations of the Ministry of the Republic of Serbia, 1994). This percentage, to some extent, decreases with depth (75% for 5-15cm and 69% for 15-30 cm), suggesting that long term smelting activity has significant consequences for soil, even at deeper unexposed layers. It is important to note that the average distance between the explored area and the copper smelter is approximately 10 km, which also indicates that the smelting activity significantly affects the large area around.

9. The obtained results were consistent with those reported by Goovaerts et al. (2008b), suggesting that such an approach could be a promising alternative for complex air dispersion models. The direct comparison of these two approaches was not possible in this study due to the missing data necessary for the characterization of the dispersion model, but it will certainly be the focus of a future study.

Chapter 5

Modeling soil properties in 3D by using penalized interaction models

5.1 Introduction

Soil mapping in 3D space (2D+depth) has been recognized as one of the main methodological challenges facing soil scientists for the last 20 years (Arrouays et al., 2014). Spline-Than-Krige approach is the most widely used approach for the prediction of soil property at different depths, see Section 2.5.2. The first step towards the real 3D prediction model was presented in the GSIF framework for digital soil mapping (Hengl, 2015; Hengl et al., 2014). The GSIF framework extends the traditional 2D regression-kriging method to 3D space. With approach it is possible to obtain the prediction of soil properties at any 3D location, and not only on pre-specified soil depths. The trend component was modeled as a sum of horizontal and vertical components, which were fitted simultaneously. The horizontal component relates spatial covariates to soil properties, while the vertical component is modeled as a linear or non-linear (spline) function of soil depth. Considering the fact that all spatial covariates are surface related, these two components had two very distinct roles in the trend model. In some cases, this could be a limitation since the effects of spatial covariates are not allowed to vary with depth. In addition, the effects of vertical component terms cannot vary in 2D space.

One way to overcome these issues is to fit the trend model by using one of the advanced machine learning techniques, like Random Forest (Hengl et al., 2015), while taking into account the soil depth, together with spatial covariates. However, besides many useful tools and methods that allow "looking inside" such models (Welling, 2016; Jones and Linder, 2015), their interpretability remains low.

Another approach is to extend the linear model by allowing for the interactions between spatial covariates and depth, as proposed by Orton et al. (2016). In their study, the trend is modeled as a multiple linear regression model, extended by linear and quadratic interaction terms that obey the hierarchy principle. This implies that, during the process of model selection, the particular *main* term cannot be excluded from the model, as long as the related interaction term has proven to be a statistically significant predictor. Therefore, the important interaction terms were selected according to the importance of the associated main effect. If the main effect had been proven to be insignificant, according to the Wald test, it was then excluded from the model, together with the corresponding interaction effects. This process was repeated after each exclusion or retention of particular main and interaction effects.

The presence of interaction effects in the trend model could contribute to a deeper understanding of the relationships between spatial covariates and their impact on response soil properties. However, even a moderate number of covariates p entails the consideration of $\frac{p(p-1)}{2}$ two-way interactions to be included in the model. This could be more demanding if the categorical variables, which should be coded prior to model fitting, are present. At this point, determining exactly which predictors (including the interactions) should be included into a model has become a crucial issue, especially if the hierarchy principle is to be obeyed. There are a number of arguments in favor of enforcing the hierarchy principle in creating the interaction models (Cox, 1984; Bien et al., 2013). All these arguments are especially important in cases when the model interpretation is of primary interest. Otherwise, the inclusion of interactions might be considered only for the purpose of improving the prediction accuracy.

The present research discusses two approaches, hierarchical and non-hierarchical, that are based on the penalized regression method, *lasso*, proposed by Tibshirani (1996). The lasso uses a specific regularization penalty in a fitting procedure to enable the efficient parameter estimation and variable selection (including interaction terms) at the same time.

In this study, rather than looking at all the possible two-way interactions, we examined only the interactions between spatial covariates and depth. The glmnet R package (Friedman et al., 2010) was used to fit a non-hierarchical model. This is a widely popular implementation of the lasso method, because of its extremely efficient fitting procedure. The hierarchy principle can be obeyed while using lasso, by adding a set of convex constraints to the lasso estimator, as proposed by Bien et al. (2013). The entire implementation of this approach is provided via the hierNet R package (Bien and Tibshirani, 2014).

The presented approach was tested on profile observations of As concentration (expressed in mg/kg), SOM content (%), and pH (measured in H_2O), sampled on the 10×20 km area in central Serbia in the vicinity of the Bor Copper Mining and Smelter Complex (see Chapter 3).

In addition, this study aims to examine whether, and to what extent, the inclusion of interactions between spatial covariates and depth improves the overall model accuracy. In order to illustrate this, a total of six trend models was compared for each soil variable. The models were divided into three groups, depending on whether they included interactions, and on whether the interaction effects obeyed the hierarchy principle. Further, each group was divided according to the level of flexibility of vertical components: linear or polynomial. By making a comparison between these models, it can be distinguished how and to what extent each extension (or hierarchy restriction) improves or impairs the model, and whether the inclusion of interactions always makes sense.

Final models for each variable were selected based on the stratified 5-fold cross validation procedure (Krstajic et al., 2014). Residuals obtained from best fit models were further analyzed for horizontal and vertical dependency, by computing variograms in both directions. The presence of spatial dependency, in both horizontal and vertical sense, was utilized for 3D variogram modeling, which was subsequently used for the 3D kriging of residuals, with the aim to improve the prediction accuracy.

In order to provide a reliable accuracy assessment, the accuracy parameters (R^2 and RMSE) were calculated through the 5-fold nested cross-validation procedure (Krstajic et al., 2014). The nested cross-validation enables the computation of accuracy parameters separately from the modeling procedure. Methodologically, this research roughly follows the generic approach proposed by Kanevski (2013) and, in our case, implies following these steps: (1) explanatory analysis and data preparation, (2) data pre-processing and partitioning, (3) regression model selection, (4) regression model assessment, (5) residual analysis and spatial modeling, (6) 3D regression kriging accuracy assessment, and (7) final prediction.

5.2 Materials and Methods

5.2.1 Environmental covariates

Based on available information and the existing knowledge, seventeen environmental covariates were selected to make up the initial set of predictors. These include fifteen continual and two categorical predictors. A 20-m resolution Digital Elevation Model (DEM), with the extent of 545 × 1146 cells, was used as the main source, as well as the basic grid system for the creation of all other covariates. The DEM was derived from 1:25,000 scale topographic map, produced by the Serbian Military Geographical Institute.

The first nine continual predictors comprise the terrain attributes commonly used in soil mapping: aspect, topographic wetness index, slope, curvature-planar and cross-sectional, channel network base level, convergence index, and vertical distance to channel network.

Bearing in mind that the large amount of toxic materials carried by the wind from the copper smelter is probably the main contributor to the elevated arsenic concentrations in certain parts of the area, several terrain parameters related to topographic exposure of terrain to wind were created. These include: topographic openness - positive and negative, explaining the topographic protection of a particular point by the surrounding topography (Yokoyama et al., 2002), and wind effect which quantifies the topographic exposure of a particular point towards the selected wind direction. The original name of this measure is Exposure toward the wind flux (Antonić and Legović, 1999), see Section 4.3.1.3. It was calculated for two major wind directions: 105° (eastern) and 315° (North-western) (Figure 4.1). All these covariates were created by SAGA GIS software (SAGA, 2014).

Besides the predictors mentioned above, two additional predictors were created with the aim to approximate the dispersion mechanism for the spread of toxic materials from the copper smelter, as used in Chapter 4. As a reminder, these are modeled as the Euclidean distances between each grid cell and the copper smelter location, and as angular differences between two azimuths: the azimuth between each grid cell and the smelter location, and the azimuth of the prevailing wind direction. Then, for each grid value, a negative exponential function was calculated by taking these quantities as exponents. These predictors were referred to as Downwind Dilution (DD) and Crosswind Dilution (CD), and were exclusively used in the of modeling As concentrations.

The soil type map with eight classes, along with the Corine Land Cover map (CLC) with five classes, were used as categorical predictors. The soil type map was generated from field observation through the use of the *spmultinom* function from GSIF R package (Hengl, 2015). The soil types are classified according to the World Reference Base (WRB) for Soil Resources (Michéli et al., 2006). The Corine Land Cover (Nestorov et al., 2007), which was originally created as a vector data set, was transformed to a raster based on a defined grid system. Table 5.1 offers an overview of the covariates, including the classes of categorical variables.

5.2.2 Trend Models

Measurements of soil property commonly reflect the average value that corresponds to the specific depth interval at a particular location $\mathbf{s}$. Accordingly, each soil observation comes with several common meta-data that describe their position in 3D space, including: 2D coordinates of sample location, upper and lower bounds of soil layer (u,l), and the observed values of soil properties. If the values of environmental covariates are assigned to each observation according to a profile spatial location, soil observations could be regressed against spatial covariates and depth by a full 3D regression model. The common form of such a model is a linear two-component model that models a 3D variation as a sum of horizontal and depth components:

$$BaseL : \mu(\mathbf{s},d) = \beta_0 + \sum_{i=1}^{n} \beta_i \mathbf{x}_i(\mathbf{s}) + \beta_{n+1} d \tag{5.1}$$

TABLE 5.1: Spatial covariates

	Name	Predictor Name	Range	Type
	TERRAIN ATTRIBUTES			
1	Digital Elevation Model	DEM	300-1045	C
2	Aspect	Aspect	0-6.283	C
3	Slope	Slope	1-1.027	C
4	Topographic Wetness Index	TWI	2.077-21.751	C
5	Convergence Index	ConvInd	-97.5-94.4	C
6	Cross Sectional Curvature	CrSectCurv	-0.038-0.04	C
7	Longitudinal Curvature	LongCurv	-0.028-0.04	C
8	Channel Network Base Level	ChNetBLevel	301.2-974.8	C
9	Vertical Distance to Channel Network	VDistChNet	0-281.86	C
10	Negative Openness	NegOp	0.796-1.835	C
11	Positive Openness	PosOp	0.809-1.726	C
12	Wind Effect (East)	WEeast	0.756-1.323	C
13	Wind Effect (North-West)	WEnw	0.749-1.323	C
14	Down-wind Dilution	DD	0.202-0.646	C
15	Cross-wind Dilution	CD	0.389-1	C
16	CORINE LAND COVER 2006			
	Pastures	clc.231	0-1	F
	Complex cultivation patterns	clc.242	0-1	F
	Land principally occupied by agriculture	clc.243	0-1	F
	Broad-leaved forest	clc.311	0-1	F
	Transitional woodland-shrub	clc.324	0-1	F
17	SOIL TYPE			
	Dystric Leptosol	LPdy	0-1	F
	Eutric Leptosol	LPeu	0-1	F
	Mollic Leptosol	LPmo	0-1	F
	Dystric Cambisol	CMdy	0-1	F
	Eutric Cambisol	CMeu	0-1	F
	Calcaric Cambisol	CMca	0-1	F
	Dystric Regosol	RGdy	0-1	F
	Vertisol	VR	0-1	F
18	Depth	d	0-1.25	C
	C-continual; F-factor (categorical)			

where $\mathbf{x}(\mathbf{s})$ represents a vector of covariate values at sampled location $\mathbf{s}$, d represents depth, and β represents the vector of model coefficients. The depth component can be replaced by appropriate higher-order additive function of depth $g(d)$, such as polynomial or even piece-wise polynomial function (e.g. spline), in order to allow a higher degree of flexibility in the depth component.

In this study, two such two-component models, that are referred to as *Base* models, were considered: the first model uses a linear depth function (*BaseL*, Equation 5.1), and the second model uses a third degree polynomial depth function (*BaseP*, Equation 5.2):

$$BaseP : \mu(\mathbf{s}, d) = \beta_0 + \sum_{i=1}^{n} \beta_i \mathbf{x}_i(\mathbf{s}) + \sum_{j=1}^{3} \beta_{n+j} d^j \tag{5.2}$$

The *Base* models were used as benchmarks in comparison to the models extended by interactions between horizontal and depth model components. Two extended models were created. The first extended model was derived from the *BaseL* model, referred to as *IntL*. The second extended model was derived from the *BaseP* model, containing the interactions between spatial covariates and all polynomial terms of depth component, referred to as *IntP*:

$$IntL : \mu(\mathbf{s}, d) = \beta_0 + \sum_{i=1}^{n} \beta_i \mathbf{x}_i(\mathbf{s}) + \beta_{n+1} d + \sum_{i=1}^{n} \theta_i \mathbf{x}_i(\mathbf{s}) d \tag{5.3}$$

$$IntP : \mu(\mathbf{s}, d) = \beta_0 + \sum_{i=1}^{n} \beta_i \mathbf{x}_i(\mathbf{s}) + \sum_{j=1}^{3} \sum_{i=1}^{n} \theta_{ji} \mathbf{x}_i(\mathbf{s}) d^j \tag{5.4}$$

where β denotes the coefficients of the main effect, while θ denotes the coefficients of the interaction effects.

A common practice in dealing with interaction models is obeying the hierarchy (heredity) principle. It states that the interaction effect should have non-zero coefficient value only if both (*strong hierarchy*), or at least one (*weak hierarchy*) of the main effects has a non-zero coefficient value. Bien et al. (2013) pointed out that, practically, violations of hierarchy occur only in special situations. Therefore, the hierarchy principle should be enforced when fitting interaction models. Since that issue belongs to the domain of variable selection, the models used in this study, which obey the hierarchy principle, have the same formulation as (Equation 5.3, Equation 5.4), and are referred to as *intHL* and *intHP*, respectively.

5.2.3 Model selection and parameter estimation

A large number of potentially useful covariates that arise from the inclusion of interactions or expansion of some covariates in polynomial form in the model, makes the model selection one of the most important tasks in this study. A simple way to meet this task is to iteratively consider a wide range of models that include different subgroups of predictors, as proposed in the subset selection, or stepwise regression procedures. However, such procedures are prone to combinatorial explosion, due to a large number of possible subsets of predictors. To address this issue, a shrinkage regression method, lasso, and its extension that was specifically developed to produce sparse interaction models that honor the hierarchy restriction, were used (see Section 2.2.7. In short, lasso uses the $l1$ regularization penalty, in combination with square-loss function in model fitting, which leads to a sparse model solution.[1] Computationally, lasso is a very attractive technique, because the model resulting from lasso's regularization is a unique global solution for a convex minimization problem.

5.2.4 Modeling procedure

The proposed approach can be presented as follows:

1. **Explanatory analysis and data preparation**. This step involves inspecting the horizontal and vertical data distribution, horizontal and vertical coverage and the magnitude of data variation. It also involves defining the extent of spatial end vertical prediction domain, preparing the covariates, spatial overlapping, and creating the *data matrix* of n observations with p covariates (including spatial covariates and depth). For polynomial depth function models (BaseP, IntP, and IntHP), the matrix is extended by columns with quadratic or higher order depth terms. In the case of models with interactions (IntL, IntP, IntHL, and IntHP), the matrix is extended by columns obtained through the element-wise multiplication of covariates columns with depth related columns.

2. **Data pre-processing**. Typically, the use of penalized regression models implies the standardization of continual covariates, prior to model fitting (Hastie et al., 2009). The reason for this lies in the dependence of the lasso solution on the variable's unit. According to this, each continual variable (including the depth terms) was scaled and standardized to have the zero mean and standard deviation set to 1. The

[1] The sparse model is a model with only a subset of non-zero coefficients.

standardization parameters (mean and standard deviation) were derived from the data matrix and stored to be used for the preparation of covariates for the final prediction. In the case of categorical variables, a full dummy coding approach was used to convert categorical variables into binary variables.

3. **Data partitioning** It is important to find a good data partitioning strategy, especially considering that the model selection and accuracy assessment were performed through the process of cross-validation. The strategy used in this study had to ensure that each sample was representative according to three criteria: 1) spatial distribution of profiles; 2) profiles depth and 3) range of observed target values. In order to approximately fulfill the first criterion, the 3-means clustering was performed, according to spatial location. Five cross-validation folds were created by the partitioning of profiles of each cluster into 5 parts, and merging the corresponding parts from each cluster to form a fold. Profiles were kept undivided during the partitioning to approximately fulfill the second criterion. Cluster partitioning was stratified with respect to the weighted mean of the target variable in each profile, while taking the length of the observed soil horizon as weight. This ensures that the last criterion is approximately fulfilled. Table 5.7 depicts the summary statistics of SOM per each fold, where values in brackets refer to the soil depth. Figure 5.1 shows the spatial distribution of profiles per fold.

4. **Regression model selection** In the case of lasso regression, model selection implies the selection of the optimal shrinkage parameter. The n-fold cross-validation procedure is a common way to select the best model, or equivalently the optimal shrinkage parameter. By defining a *grid* of values for the λ parameter and computing the cross-validation error e_{cv} for each value, the optimal λ value is the one which gives the lowest e_{cv}. In this study, the 5-fold cross-validation, based on previously described sampling strategy and the sequence of λ between the 0 and 5 with a step of 0.1, was used. This sequence of λ is exclusively used for the glment models, while the hierNet algorithm defines the grid of λ automatically, for the IntHL(P) models.

5. **Regression model assessment**. Considering that the regression models were fitted by lasso, the accuracy assessment was based on the nested cross-validation procedure, see Section 2.2.2.2. The same data partitioning strategy, as used for model

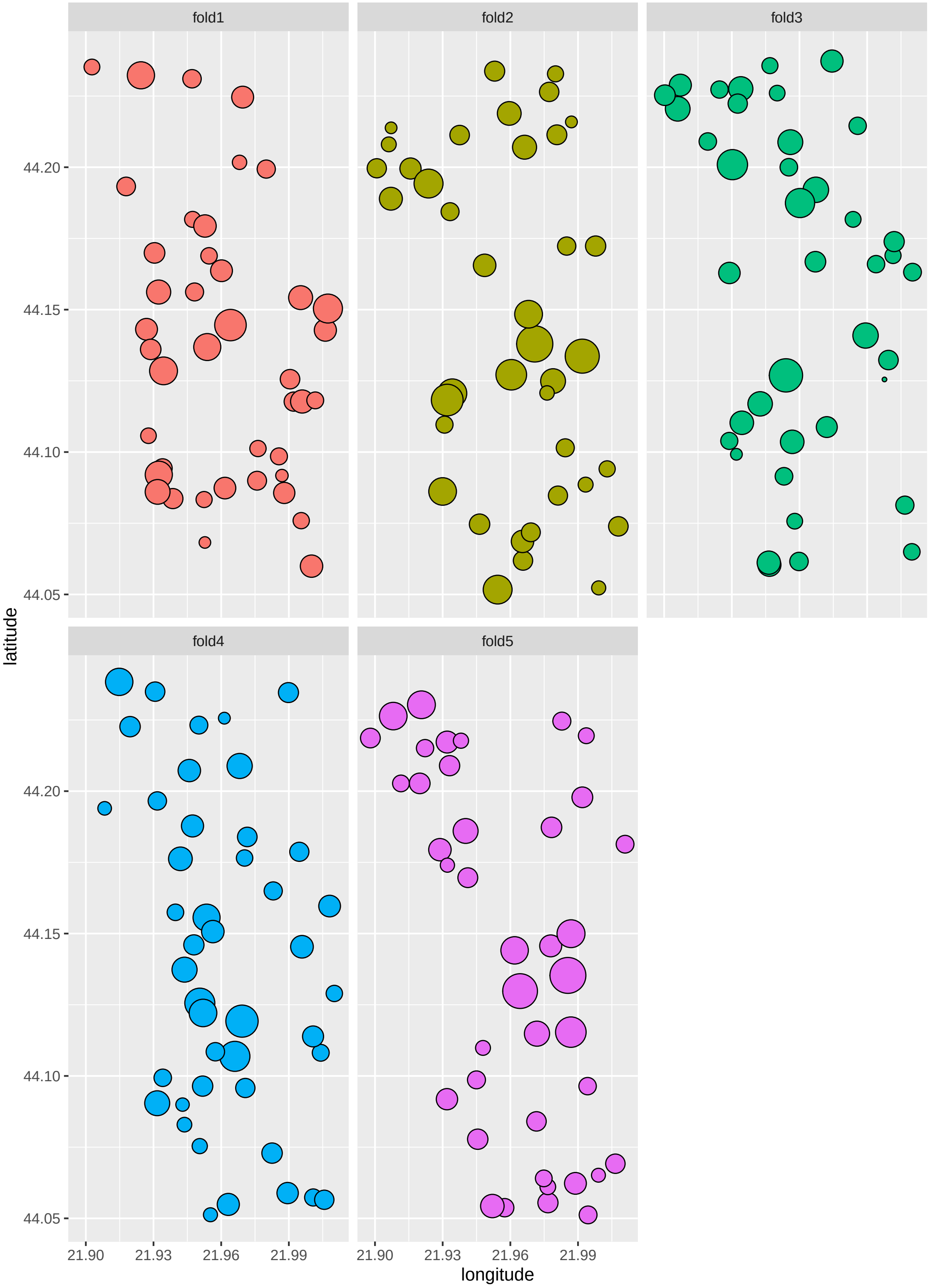


FIGURE 5.1: Spatial distribution of SOM content observations in each fold

selection, is used for data partitioning in both the outer and the inner loop of nested cross-validation.

6. **Residual analysis and spatial modeling** This step is of particular interest in this study. The residual analysis aims to examine the presence of spatial correlation patterns in residuals. The presence of any non-random structure in residuals justifies their further modeling by geostatistical tools. The analysis started by computing vertical and horizontal residual variograms. The measurements, and thereby the residuals, relate to the specific depth intervals within the soil profile (mostly soil horizons). This has been a limiting factor for vertical variogam calculation, because the distances were being calculated only between the mid-points of two horizons. In order to overcome this problem, we modeled the residuals by using the mass-preserving (equal-area) spline function (Bishop et al., 1999) which is implemented in the mpspline function from GSIF R package (Hengl, 2015). By using this approach, the interval-based residuals were transformed into a continuous form, enabling the computation of the semi-variances between any two points along the soil profile. Due to the continuous form, sample variograms in the vertical direction were typically modeled with Gaussian theoretical model, with zero nugget. The sample variogram in the horizontal (2D) space was computed using only the data from the surface horizons. The fitted variogram ranges were then used for defining the anisotropic distance according to Equation 2.47 and hence the 3D anisotropic variogram model was generated. By this approach, a single-structure 3D anisotropic variogram model, that incorporates the geometric anisotropy between horizontal and vertical directions, was modeled. Once the 3D variogram has been modeled, it was used for the 3D kriging prediction, as well as, for the assessment of accuracy of the kriging prediction via cross-validation.

7. **Accuracy assessment of 3D regression kriging** In order to ensure the consistency with the assessment of the regression part, the residual variograms were modeled exclusively based on the training data residuals. Final accuracy measures were computed based on the final estimates, which were obtained as sums of regression estimates and kriging residual estimates on the test data (in outer loop of nested cross-validation), within the nested cross-validation procedure. In other words, the residuals obtained by running the trend model on the training data were used for 3D

variogram modeling, which was then used to interpolate the test residuals by 3D kriging.

8. **Final prediction** The final predictions were produced using the sum of the regression and the 3D kriging predictions for the entire area, at three different depths: 10 cm, 20 cm and 30 cm. But first, this step requires applying the same preprocessing steps on gridded covariates to make them consistent with the input data.

The entire computation procedure is additionally explained in Chapter 6 through the description of software functions, specifically developed for this research.

5.3 Results and Discussion

It is important to clarify that the results obtained via glmnet package (BaseL,BaseP, IntL, and IntP models) and hierNet package (IntHL and IntHP models) are not completely comparable due to different implementations of the lasso method. Therefore, the difference between interaction and non-interaction models should be sought in comparison between BaseL(P) versus IntL(P), while the results for the IntHL(P) should provide the information about the consequences of hierarchy constraints, i.e. should show whether the hierarchy constraints cause any consequences for model selection, or even for overall prediction performance. Significant differences that exist in training time between two lasso implementations is also worth noting. For example, in our case, 5-fold nested cross-validation, that is run on a computer with a $4-th$ generation `i7` processor and `16 Gb RAM`, takes 5 seconds for glmnet implementation, while for hierNet implementation it takes about 40 minutes. The reason for such processing time differences most probably lies in the optimization procedure implemented in hierNet package, which becomes dramatically complicated by additional constraints added to ensure the hierarchical parameter settings.

Considering that the penalty parameter (λ_{cv}) obtained by cross-validation varies depending on the random partitioning of observations, data were split into 5 stratified folds according to sampling strategy explained in Section 5.2.4, prior to running model selection or model assessment procedures.

5.3.1 3D modeling and spatial prediction of arsenic concentration

Table 5.2 shows the summary statistics for 5 folds of As concentration data. Values in brackets refer to the same statistical parameters calculated for the values of depth in cm [2]. It is important to emphasize here that the same data partitioning strategy was used for the model selection and the model assessment.

TABLE 5.2: Basic statistical parameters for stratified 5-fold data splitting of As concentration data

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
fold1	2.7(−0.86)	10.1(−0.38)	23.0(−0.18)	42.7(−0.25)	55.3(−0.11)	328.0(−0.02)
fold2	1.8(−1.25)	9.0(−0.40)	25.4(−0.20)	34.3(−0.29)	52.8(−0.12)	174.0(−0.02)
fold3	2.4(−1.08)	8.7(−0.39)	25.6(−0.18)	38.1(−0.27)	56.0(−0.10)	228.0(−0.01)
fold4	2.4(−0.88)	8.4(−0.36)	23.8(−0.17)	49.7(−0.24)	50.3(−0.10)	392.0(−0.02)
fold5	2.3(−0.96)	9.4(−0.35)	24.9(−0.16)	41.5(−0.26)	47.8(−0.10)	255.0(−0.02)

Table 5.3 shows the RMSE and R^2 values obtained by stratified 5-fold nested cross-validation for BaseP(L), IntP(L) and IntHP(L). It is apparent that models that use the interactions between the spatial covariates and depth terms perform considerably better than the benchmark Base models. It is also apparent that differences are more pronounced in terms of the R^2 measures than in RMSE measures, which is to some extent a result of the inclusion of a larger number of predictors. On the other hand, results obtained for models with or without respect for the hierarchy, do not differ meaningfully in terms of RMSE; whereas differences are not negligible in terms of R^2 measures. Moreover, it is noticeable that the inclusion of a polynomial depth component to the Base model do not contribute much to a better predictive performance. This indicates that the vertical variation is not equal over the entire area. Hence, the spatially independent vertical trend is not a good solution. This result underlines the importance of considering the inclusion of interactions between spatial covariates and depth in the model. This actually means that a more flexible depth function is not a promising alternative for the inclusion of interaction effects.

[2] Negative sign indicates the measurements bellow the terrain surface

TABLE 5.3: Results of nested 5-fold cross-validation for As models

Model	Base		Int		IntH	
	L	P	L	P	L	P
RMSE	42.46	42.54	39.87	39.95	40.83	40.39
R^2	0.36	0.35	0.43	0.43	0.40	0.42

The coefficients for the final 3D regresion models of BaseP, IntP and IntHP regression settings for As concentration are shown in Table 5.4. [3] Considering that all predictors are scaled and standardized prior to model fitting, their importance can be compared to a certain extent simply by observing the magnitude of their regression coefficients.

In the case of the BaseP model, lasso retained all the variables, while setting the penalty parameter to 0. This resulted in unbiased coefficient estimates that are equivalent to ordinary least squares estimates. The "depth" terms, wind related variables (CD, DD and WEno), as well as classes clc321, clc324, CMca, RGdy and LPmo, appear to have the largest effects on prediction, whereas the other predictors have small to moderate effects. It is worth knowing that some of the important predictors are almost constant across observations, which are mostly dummy variables, such as Calcaric Cambisol and Mollic Leptosol that appear in <20 profiles (see Figure 3.4). Typically, such variables would be considered as non-informative (near zero variance variables) and as such removed from further analysis. However, in this case, lasso has recognized their importance.

For the IntP and IntHP models, lasso sets a more sparse model structure by setting the penalty parameter to 0.8 and 362.1, respectively. Generally, all the main effects are additionally shrunk towards zero, while some of them (the least important) were equaled exactly to zero. This is undoubtedly the consequence of the inclusion of interactions between spatial covariates and depth, which is particularly true for the "depth" terms. It is apparent that the two approaches for fitting models with interactions (hierarchical and non-hierarchical) produce two significantly different models. The two main differences can be observed: first, the main effects are more shrunk towards zero in the case of IntHP model; second, the selection of interaction terms substantially differs. This is particularly expressed in interaction terms associated with the dummy variables, where the IntP model

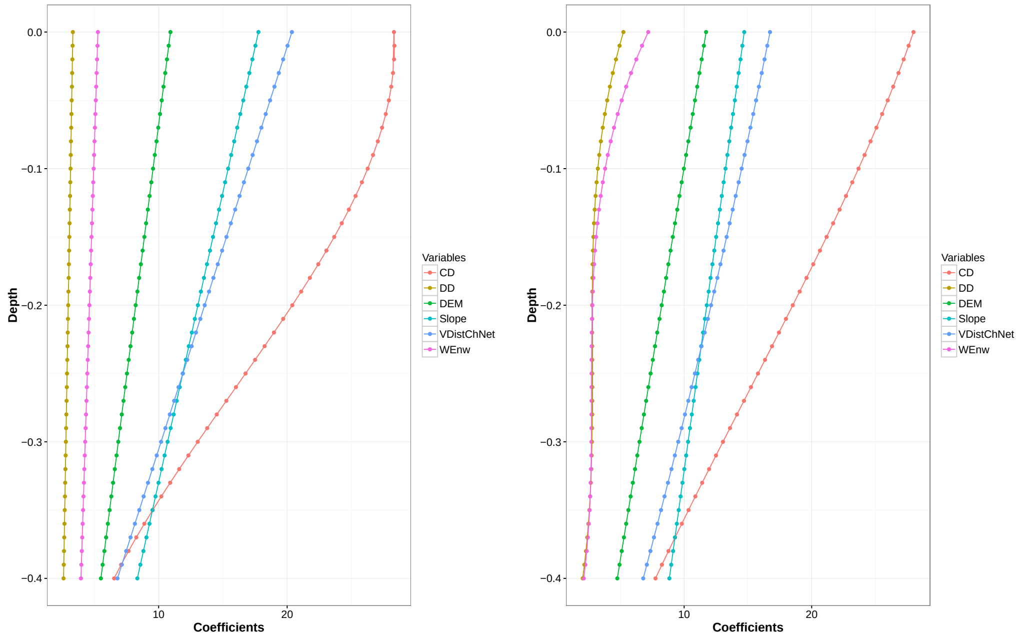
[3] The *"me"* columns refer to the coefficients of the mean effects, whilst the *"ie"* columns refer to the coefficients of interaction effects.

TABLE 5.4: Final models coefficients for *BaseP*, *IntP* and *IntHP* models for As concentration.

variable	BaseP, λ_{CV} = 0	IntP, λ_{CV} = 0.8				IntHP, λ_{CV} = 362.1			
	me	me	$ie(d)$	$ie(d^2)$	$ie(d^3)$	me	$ie(d)$	$ie(d^2)$	$ie(d^3)$
Int	36.020	37.248	–	–	–	–	–	–	–
DEM	9.100	7.350	2.996	0	0	7.129	3.856	0	0
Aspect	2.712	1.437	3.095	0	0	1.769	0.212	−1.557	0
Slope	17.017	11.552	5.218	0	0	10.840	3.252	0	0
TWI	14.227	6.710	0.955	0	0	6.889	0	0.728	0
ConvInd	-0.552	0	-2.762	0	3.625	−0.928	0	0	0.928
CrSectCurv	−8.445	−3.987	−1.052	0	0	−3.442	0	0	0
LongCurv	−1.632	0	0	0	0	0	0	0	0
ChNetBLevel	5.405	0	0	0	0	0	0	0	0
VDistChNet	11.288	11.419	7.516	0	0	10.490	5.751	−0.426	0
NegOp	−9.914	−0.328	−1.045	0	0	0	0	0	0
PosOp	11.090	0	0	0	1.785	0	0	0	0
WEeast	−12.376	−1.967	0	0	0	0	0	0	0
WEnw	7.894	4.404	0.730	0	0	2.740	0	−0.00001	2.633
DD	7.271	2.848	0.401	0	0	2.799	0	−0.737	2.063
CD	15.284	15.735	16.558	0	−4.231	15.035	12.191	0	−0.927
clc.231	−35.241	−22.685	0	1.793	−4.586	−2.064	0	0	0
clc.242	0.006	0	7.121	0	0	0	0	0	0
clc.243	−7.137	−6.779	0	0	0	−2.860	0	1.989	0
clc.311	5.026	2.289	9.301	0	0	1.982	0	0	0
clc.324	24.626	23.371	3.701	0	−31.590	4.247	0	0	0
CMca	53.284	43.280	0	0	−30.626	7.258	0	0	0
CMdy	−4.156	0	0	0	−1.057	0	0	0	0
LPdy	−1.438	0	7.883	0	−12.964	0	0	0	0
RGdy	−46.269	−35.590	0	0	0	−4.686	0	0	0
CMeu	3.341	−1.836	0	0	−1.653	−0.290	0	0	0
LPeu	5.007	−0.197	0	0	0	0	0	0	0
LPmo	17.375	0	3.151	0	0	0	0	0	0
VR	18.782	8.543	3.554	0	0	1.750	0	0	0.100
d	67.703	15.166	0	0	0	25.261	0	0	0
d2	104.247	0	0	0	0	5.437	0	0	0
d3	52.037	−4.739	0	0	0	−6.651	0	0	0

includes several non-negligible interactions, while the same interaction terms in IntHP model have zero coefficients. Therefore, the two models yield different interpretations. Thus, according to the IntP model, the abrupt changes in the vertical distribution of As concentration occur between the categorical classes, while the IntHP model does not yield the same results.

Figure 5.2(a) and Figure 5.2(b) depict the *paths* of effects for the most important continual predictors of the IntP and IntHP models, as functions of depth. Paths were presented at the depth interval of 0-40 cm. These graphs show how some variables influence the prediction at particular depths. The variation depends on the type of interactions included in the model. For example, the effects of Downwind Dilution and Wind Effect vary non-linearly in the IntHP model, whereas they have linear form in the IntP model. These figures also provide an examination of the relationship between the effects at particular depths. In other words, based on these graphs, the extent of which certain variables influence the prediction at particular depths can be examined. Therefore, different effects are observed and more pronounced near the soil surface whilst, the effects tend to equalize at deeper depths.



(a)

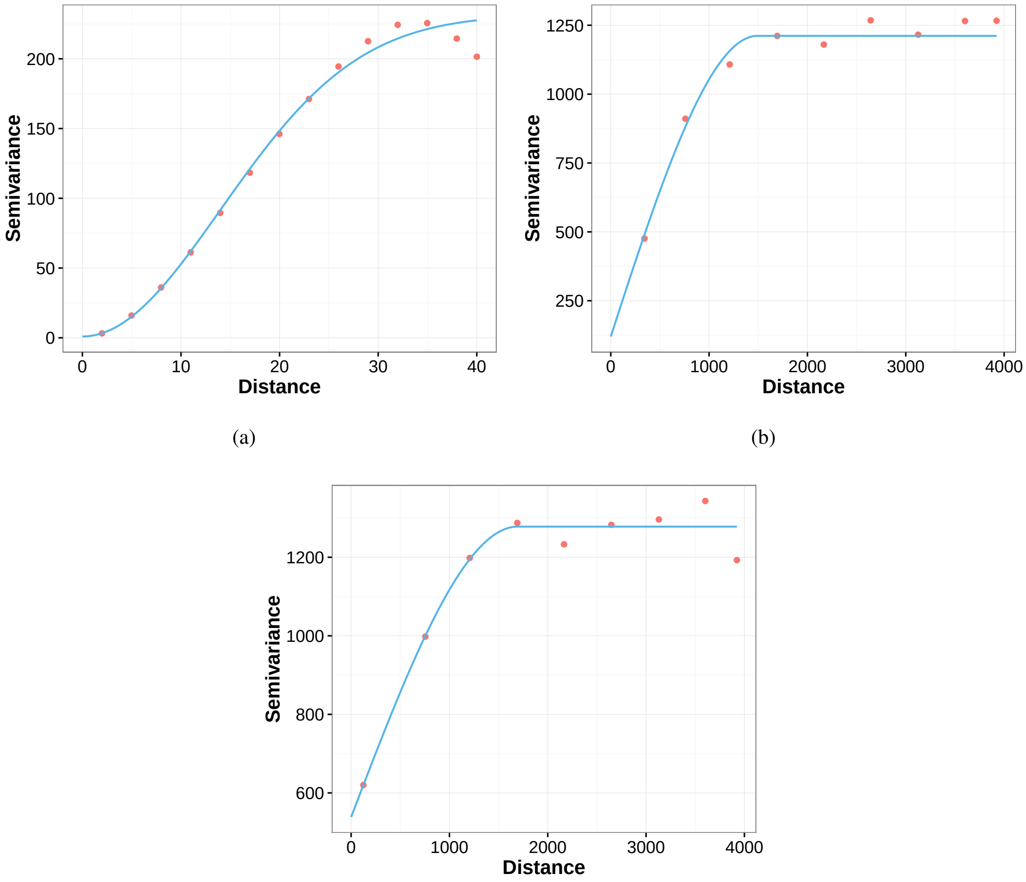
(b)

FIGURE 5.2: Coefficients path for As models: a) IntP model, b) IntHP model

The IntP model yielded the best predictive performance, and therefore it was used for further analysis. Residual variograms in the vertical direction, horizontal space and in 3D are depicted in Figure 5.3. Residuals were obtained via the IntP model. The first two variograms show a clear spatial dependence in both vertical and horizontal directions. Table 5.5 lists the fitted 3D variogram parameters. The geometric anisotropy is expressed as the value of horizontal distance that corresponds to the 5 cm depth. The Nugget/Sill ratio shows that residuals are moderately spatially structured according to Cambardella et al. (1994), indicating that the 3D kriging of residuals is certainly applicable.

TABLE 5.5: Parameters for the fitted 3D residual variogram model for As concentration

	Nugget	Sill	Range	Anisotropy (5cm depth=)	Nugget/Sill
As	581.47	1,256.00	1,632.34 m	343.66 m	0.46



(a)

(b)

(c)

FIGURE 5.3: Fitted residual variograms for As concentration data: a) Vertical (depth) variogram, b) Horizontal variogram, c) 3D variogram

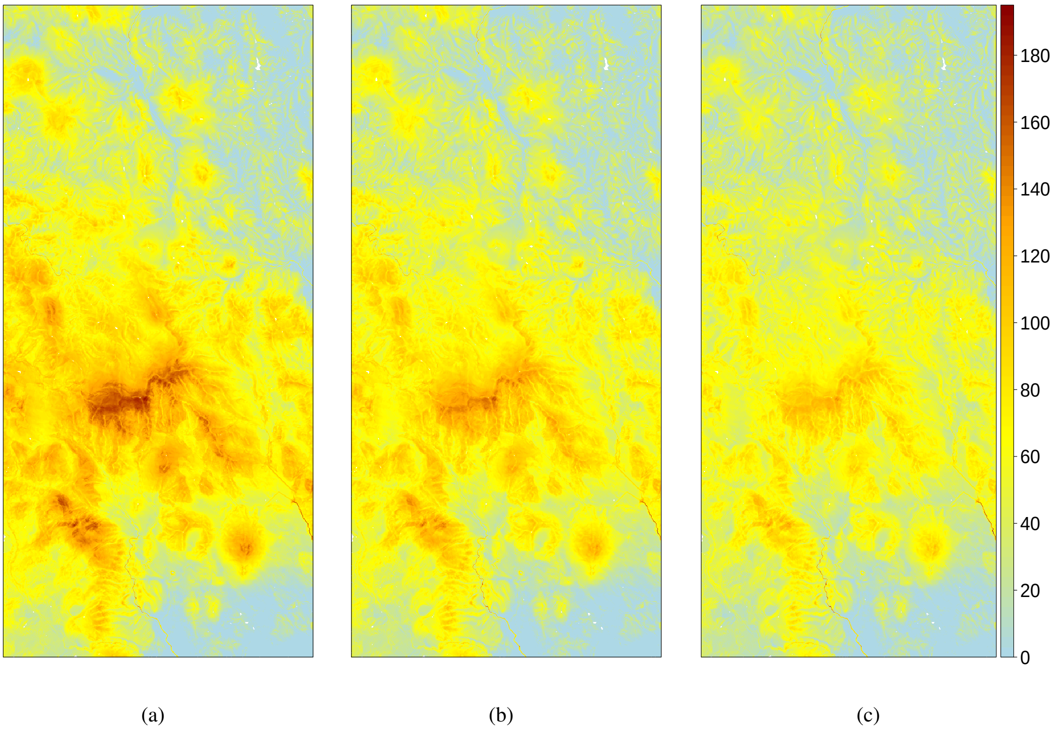
The overall accuracy parameters (R^2 and RMSE) achieved by the trend model (final IntP model) and 3D regression kriging are shown in Table 5.6. A considerable improvement in prediction accuracy was achieved by additional kriging of the residuals. The

RMSE measure has decreased from 39.95 mg/kg to 37.13 mg/kg while the R^2 measure has increased up to 0.51.

TABLE 5.6: Accuracy parameters for the IntP regression model and 3D regression kriging with IntP trend model for As concentration

Method	IntP		3D RK	
	RMSE	R^2	RMSE	R^2
As	39.95	0.43	37.13	0.51

The maps of predicted values As concentration over the entire area are given in Figure 5.4. The final predictions was performed for the depths of 10 cm, 20 cm and 30 cm.



(a)

(b)

(c)

FIGURE 5.4: Final prediction maps of As concentration produced by 3D regression kriging with IntP trend model: a) 0.1 m depth; b) 0.2 m depth; c) 0.3 m depth.

5.3.2 3D modeling and spatial prediction of SOM content

Table 5.7 shows the summary statistics for stratified folds of SOM content data. Values in brackets refer to the soil depth in cm. When observing each column separately, it can be noticed that the folds are well-stratified and provide similar statistics for each fold.

TABLE 5.7: Basic statistical parameters for stratified 5-fold data splitting of SOM content data

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
fold1	0.11(−1.08)	1.97(−0.35)	3.60(−0.18)	4.48(−0.26)	5.87(−0.10)	21.94(−0.02)
fold2	0.61(−0.96)	1.99(−0.37)	3.64(−0.15)	5.44(−0.23)	6.16(−0.09)	30.73(−0.02)
fold3	0.11(−0.87)	1.83(−0.40)	3.49(−0.18)	5.04(−0.27)	6.04(−0.11)	40.47(−0.02)
fold4	0.27(−1.25)	2.27(−0.38)	3.42(−0.16)	4.45(−0.28)	5.75(−0.10)	19.67(−0.02)
fold5	0.80(−0.85)	2.30(−0.38)	4.15(−0.20)	5.20(−0.26)	6.48(−0.12)	23.65(−0.01)

Comparison of RMSE and R^2 measures calculated via stratified nested 5-fold cross-validation for BaseL(P), IntL(P) and IntHL(P) regression models is shown in Table 5.8. Considering the pronounced higher variation of SOM content in the upper soil layers (Figure 3.5), the interaction models yielded considerably better predictive performance, in comparison to the benchmark Base models. In addition, an inclusion of quadratic and cubic depth terms and their interactions improves the prediction accuracy to some extent.

TABLE 5.8: Results of nested 5-fold cross-validation for SOM models

Model	Base		Int		IntH	
	L	P	L	P	L	P
RMSE	3.62	3.53	3.49	3.40	3.53	3.48
R^2	0.39	0.41	0.43	0.46	0.42	0.43

Table 5.9 lists the estimated coefficients for BaseP, IntP and IntHP regression models, respectively. It is apparent that the lasso produces quite different models depending on whether the interactions are considered or not. By including the interactions into the consideration, many of the main effects are excluded from the model. In the case of BaseP model, lasso has retained all the input variables in the model although some of them have coefficients that are very close to zero, which is obviously a consequence of selecting the penalized parameter equal to 0.3. The importance of dummy variables is noticeable.

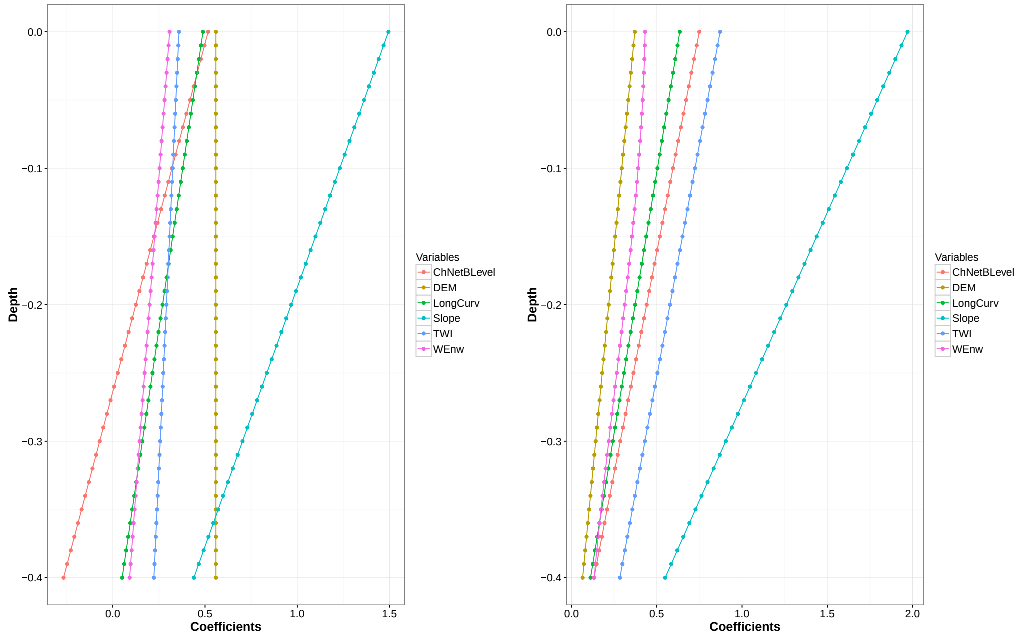
Although the depth terms were expected to be significant, their coefficients turn out to be surprisingly large in comparison to the other terms. The inclusion of interactions resulted in a sparse structure in terms of main effects (λ=0.1 for IntP model, while for IntHP model λ=28.6). This is particularly expressed in the non-hierarchical (IntP) model. The output for the IntP model shows that, among 30 "main" variables, 15 variables are excluded to be non-informative for prediction. Some of them, like positive openness, CMca, RGdy, LPmo, VR and d^2 have very strong effects on prediction in the non-interaction model. As with As concetration models, the coefficients in the interaction models are additionally shrunk towards the zero. In comparison to the magnitude of the main effects, the interaction effects in both the IntP and IntHP models are relatively strong. The IntP model will be used for further analysis since this model provided the best predictive performance.

TABLE 5.9: Final models coefficients for *BaseP*, *IntP* and *IntHP* models for SOM conctent.

variable	BaseP, λ_{CV} = 0.3	IntP, λ_{CV} = 0.1				IntHP, λ_{CV} = 28.6			
	me	me	$ie(d)$	$ie(d^2)$	$ie(d^3)$	me	$ie(d)$	$ie(d^2)$	$ie(d^3)$
Int	3.254	4.549	–	–	–	–	–	–	–
DEM	−0.019	0.559	0	0	0	0.169	0.169	0	0
Aspect	0.111	0	0.179	0	0	0.039	0.081	0	0
Slope	1.901	0.799	0.584	0	0	1.033	0.787	0	0
TWI	0.945	0.269	0.075	0	0	0.484	0.325	0	0
ConvInd	0.105	0	0	0	0	0	0	0	0
CrSectCurv	−0.193	0	0	0	0	0	0	0	0
LongCurv	0.357	0.200	0.243	0	0	0.290	0.290	0	0
ChNetBLevel	0.656	0	0.434	0	0	0.342	0.342	0	0
VDistChNet	−0.094	0	0	0	0	0	0	0	0
NegOp	0.075	0	0	0	0	0	0	0	0
PosOp	1.039	0	0	0	0	0.178	0	0	0
WEeast	−0.112	0.092	0	0	0	0.043	0	0	0
WEnw	0.342	0.165	0.120	0	0	0.253	0.210	0	−0.042
clc.231	−0.952	0	0	0	0	0	0	0	0
clc.242	0.682	0	0	0	0	0	0	0	0
clc.243	0.082	−0.450	0	0	−0.115	−0.265	−0.105	0.152	0
clc.311	1.764	0.615	1.219	0	−0.344	0.335	0.335	0	0
clc.324	2.527	0.637	1.390	0	0	0.353	0	0	0
CMca	1.755	0	0	0	0	0.045	0	0	0
CMdy	0.936	0.015	0	0	0	0.079	0	0	0
LPdy	0.269	0	0.766	0	−1.109	−0.012	0.012	0	0
RGdy	1.982	0	1.423	0	0	0.131	0	0	0
CMeu	0.818	−0.026	0	0	0	0	0	0	0
LPeu	0.471	−0.251	0	0	−0.083	−0.060	0	0	0
LPmo	2.228	0	1.410	0	0	0.180	0	0	0
VR	1.420	0	0	0	0	0.070	0	0	0
d	8.288	2.101	0	0	0	3.121	0	0	0
d2	10.907	0	0	0	0	0.416	0	0	0
d3	4.794	−0.425	0	0	0	−0.754	0	0	0

Figure 5.5 depicts the coefficient paths for the six most important continuous variables of the IntP and IntHP models. These figures graphically illustrate changes in the

magnitudes of their effects on prediction along 40 centimeters of depth. The pronounced effect of Slope is dominant along the whole depth interval. These figures also reveal a pronounced decreasing trend in effects of other variables in the IntHP model when compared to the IntP model.

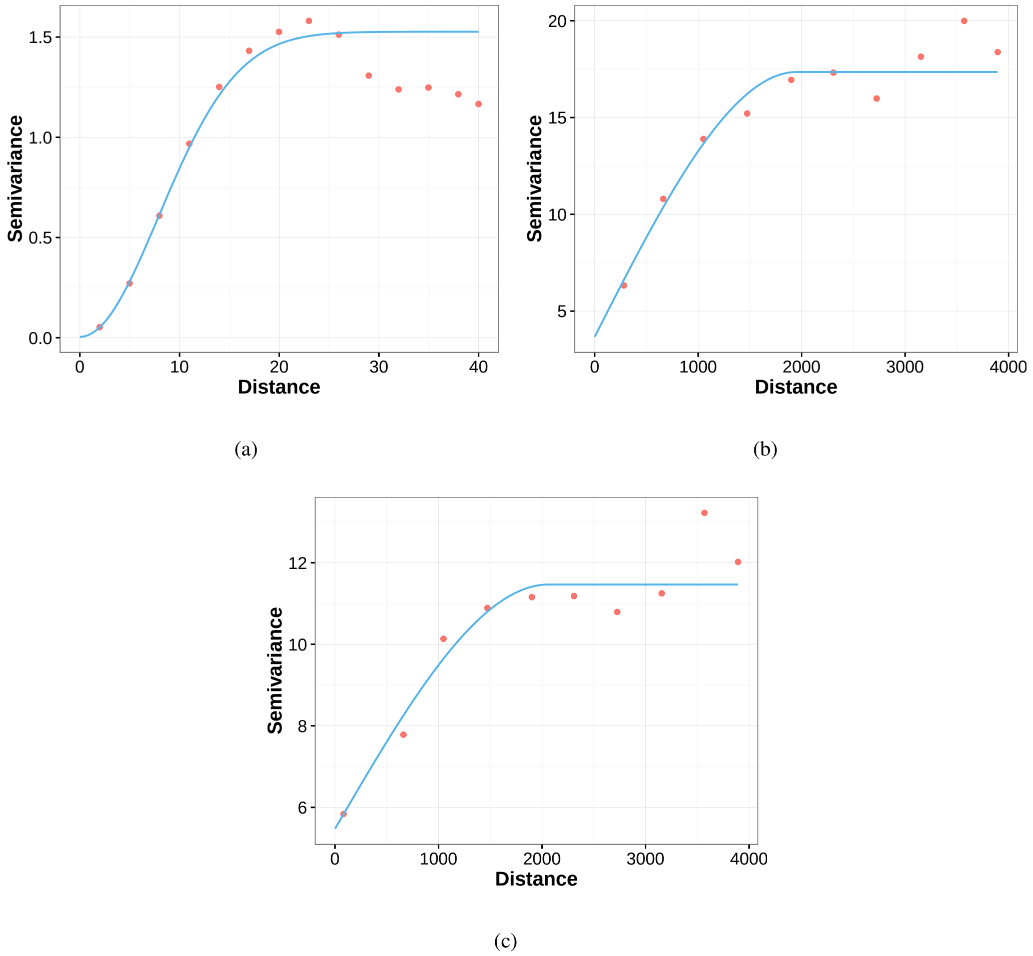


(a)

(b)

FIGURE 5.5: Coefficients path for SOM models: a) IntHP model, b) IntP model

Figure 5.6 and Figure 5.6(b) show fitted vertical and horizontal residual variograms. These figures reveal that the residuals are clearly correlated in both vertical and horizontal sense. The variograms suggest that the residuals are correlated horizontally up to 2000 m in distance, whereas the correlation reaches the sill at 20 cm in the vertical direction. The horizontal variogram, and the 3D variogram, are fitted by the spherical function (Figure 5.6(c)). Table 5.10 lists the fitted parameters for the 3D residual variogram model.



(a)

(b)

(c)

FIGURE 5.6: Fitted residual variograms for SOM concentration data: a) Vertical (depth) variogram, b) Horizontal variogram, c) 3D variogram

TABLE 5.10: Parameters for the fitted 3D residual variogram model for SOM content

	Nugget	Sill	Range	Anisotropy	Nugget/Sill
				(5cm depth=)	
SOM	5.50	11.34	2,017.58 m	932.30 m	0.48

Table 5.11 shows the accuracy parameters obtained by 3D regression kriging along with the accuracy parameters of the trend model. The trend is estimated by the IntP model.

The subsequent interpolation of the residuals by 3D kriging resulted in an improvement of the prediction accuracy. Figure 5.7 shows the predicted SOM content via the 3D regression kriging model over the entire area at three different depths: 10 cm, 20 cm and 30 cm.

TABLE 5.11: Accuracy parameters for the IntP regression model and 3D regression kriging with IntP trend model for SOM content

Method	IntP		3D RK	
	RMSE	R^2	RMSE	R^2
SOM	3.40	0.46	3.24	0.51

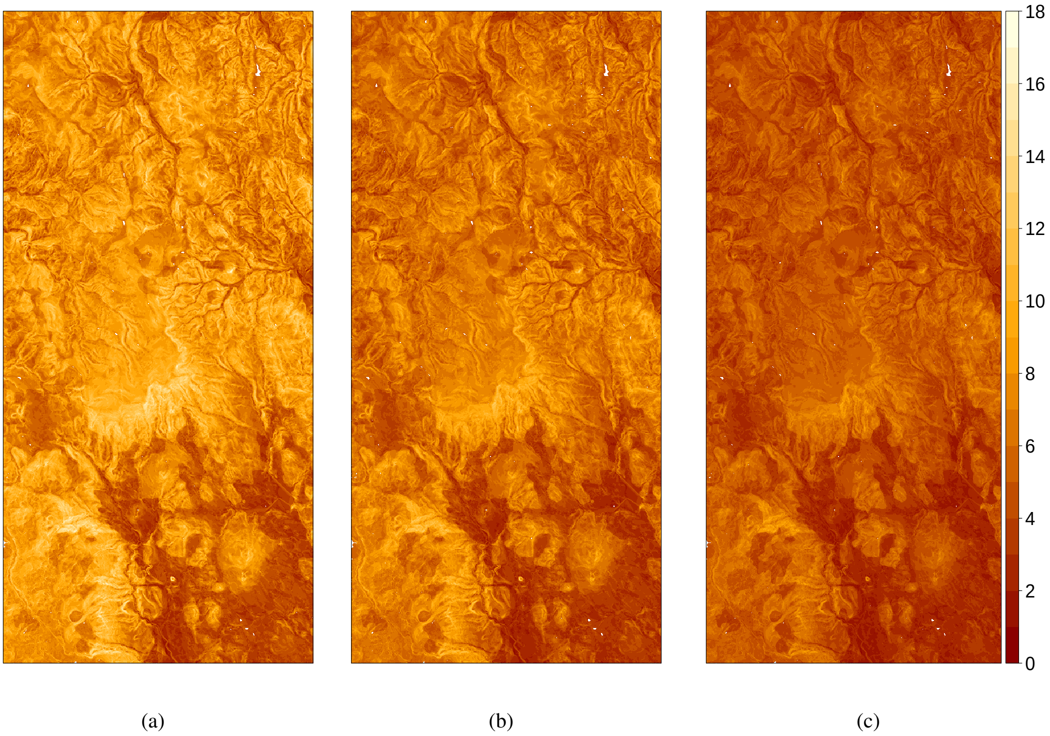


FIGURE 5.7: Final prediction maps of SOM content produced by 3D regression kriging with IntP trend model: a) 0.1 m depth; b) 0.2 m depth; c) 0.3 m depth.

5.3.3 3D model and spatial prediction of pH

As with As and SOM data, the pH profile data was split into 5 stratified folds according to the sampling strategy described in Section 5.2.4. Table 5.12 summarizes folds according to the pH values and depth.

TABLE 5.12: Basic statistical parameters for stratified 5-fold data splitting of pH data

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
fold1	3.80(−1.18)	4.74(−0.43)	5.50(−0.21)	5.38(−0.31)	5.91(−0.14)	7.30(−0.03)
fold2	3.45(−0.85)	4.66(−0.39)	5.20(−0.18)	5.29(−0.27)	5.92(−0.14)	7.40(−0.02)
fold3	3.60(−0.87)	4.60(−0.41)	5.04(−0.24)	5.22(−0.30)	5.84(−0.12)	7.70(−0.02)
fold4	3.90(−0.85)	4.60(−0.38)	5.25(−0.19)	5.21(−0.26)	5.78(−0.12)	6.90(−0.02)
fold5	3.70(−1.25)	4.65(−0.43)	5.10(−0.24)	5.22(−0.30)	5.83(−0.12)	6.95(−0.05)

Table 5.13 summarizes the results of 5-fold nested cross-validatoin for BaseP(L), IntP(L) and IntHP(L) models for pH data. All models perform almost the same, indicating that the inclusion of interactions do not contribute much to predictive performance. This can be attributed to the quite constant variation of pH along the depth (see Figure 3.5). However, inspection of Table 5.14 reveals that the three models (BaseP, IntP and IntHP) take quite different forms. The BaseP model included all 30 variables that constitute the initial input set of predictors. On the other hand, the IntP and IntHP regressions started the model selection with an almost three times larger initial set; however, they resulted in much simpler models (although the interaction terms were not included). In the case of non-hierarchical regression (IntP), only 6 variables were selected, while in the case of hierarchical regression (IntHP) a model with 17 parameters was selected. Considering that the IntP model performs similarly as other models, but with much simpler parameter settings, it will be used for further analysis and mapping.

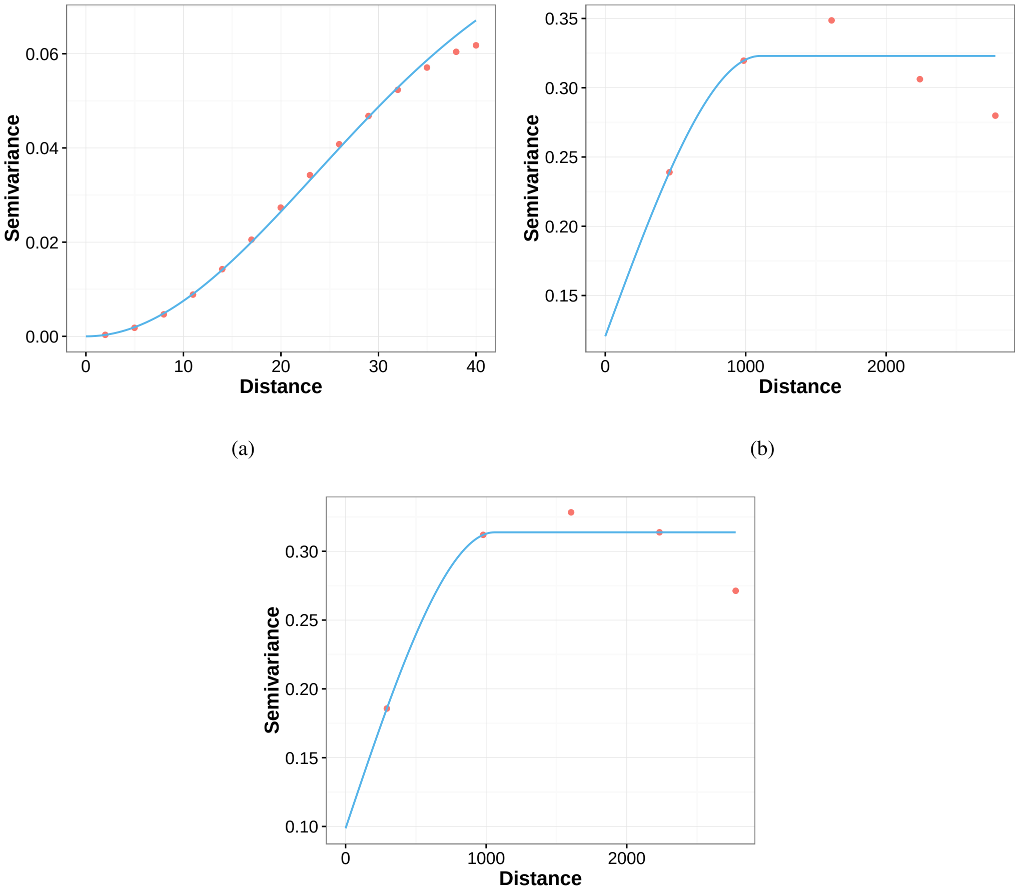
TABLE 5.13: Results of nested 5-fold cross-validation for pH models

Model	Base		Int		IntH	
	L	P	L	P	L	P
RMSE	0.59	0.57	0.60	0.60	0.59	0.59
R^2	0.52	0.54	0.52	0.52	0.52	0.52

TABLE 5.14: Final models coefficients for *BaseP*, *IntP* and *IntHP* models for pH.

variable	BaseP, λ_{CV} = 0	IntP, λ_{CV} = 0.1				IntHP, λ_{CV} = 12.1			
	me	me	$ie(d)$	$ie(d^2)$	$ie(d^3)$	me	$ie(d)$	$ie(d^2)$	$ie(d^3)$
Int	5.356	5.301	–	–	–	–	–	–	–
DEM	−0.187	−0.098	0	0	0	−0.131	0	0	0
Aspect	0.072	0	0	0	0	0.004	0	0	0
Slope	−0.052	0	0	0	0	0	0	0	0
TWI	−0.089	0	0	0	0	−0.014	0.014	0	0
ConvInd	−0.004	0	0	0	0	0	0	0	0
CrSectCurv	0.011	0	0	0	0	0	0	0	0
LongCurv	-0.119	0	0	0	0	-0.053	0	0	0
ChNetBLevel	−0.237	−0.276	0	0	0	−0.211	0	0	0
VDistChNet	−0.005	0	0	0	0	0	0	0	0
NegOp	0.088	0	0	0	0	0	0	0	0
PosOp	−0.103	0	0	0	0	0	0	0	0
WEeast	0.312	0	0	0	0	0.045	0	0	0
WEnw	−0.052	0	0	0	0	0	0	0	0
clc.231	0.138	0	0	0	0	0.002	0	0	0
clc.242	0.335	0.146	0	0	0	0.097	0	0	0
clc.243	0.045	0	0	0	0	0	0	0	0
clc.311	−0.180	−0.086	0	0	0	−0.079	0	0	0
clc.324	−0.134	0	0	0	0	−0.014	0	0	0
CMca	0.276	0	0	0	0	0.044	0	0	0
CMdy	−0.569	0	0	0	0	−0.069	0	0	0
LPdy	−0.240	−0.155	0	0	0	−0.103	0	0	0
RGdy	−0.430	0	0	0	0	−0.032	0	0	0
CMsu	−0.090	0	0	0	0	0	0	0	0
LPeu	−0.125	0	0	0	0	0	0	0	0
LPmo	0.706	0	0	0	0	0.075	0	0	0
VR	0.073	0	0	0	0	0.054	0	0	0
d	−0.352	−0.149	0	0	0	−0.216	0	0	0
d2	−0.043	0	0	0	0	0	0	0	0
d3	0.091	0	0	0	0	0	0	0	0

Figure 5.8 depicts the residual variogram models in the vertical and horizontal directions as well as in the 3D, fitted to the residuals resulting from the IntP model. The spatial correlation in both the vertical and horizontal senses is observed. The shape of the vertical variogram model shows a higher continuity of pH residuals in the vertical direction than in the case of As and SOM residuals. On the other hand, the horizontal variogram obviously has a much shorter range and reaches a sill at a distance shorter than 1000 m. It is reflected by a much lower anisotropy ratio than in with As and SOM residuals. A correlation at a distance of 5 cm in the vertical direction corresponds to the 165 m in horizontal direction (see Table 5.15), while in the case of As and SOM the corresponding horizontal distances are ˜900 and ˜350 m. The fitted parameters for the resulting 3D variogram model are given in Table 5.15.



(a)

(b)

(c)

FIGURE 5.8: Residual variograms for IntP model: a) Vertical (depth) variogram, b) Horizontal variogram, c) 3D variogram

TABLE 5.15: Parameters for the fitted 3D residual variogram model for pH

	Nugget	Sill	Range	Anisotropy (5cm depth=)	Nugget/Sill
pH	0.10	0.21	1,060.44 m	065.08 m	0.47

The final accuracy parameters for 3D regression kriging and the corresponding trend model are given in Table 5.16. As is evident in the table, the residual interpolation by 3D kriging shows remarkable improvements over the regression (IntP) model. Predicton of pH over the entire area at three different depths, 10 cm, 20 cm and 30 cm, are depicted in Figure 5.9.

TABLE 5.16: Accuracy parameters for the IntP regression model and 3D regression kriging with IntP trend model for pH

Method	IntP		3D RK	
	RMSE	R^2	RMSE	R^2
pH	0.60	0.49	0.57	0.54

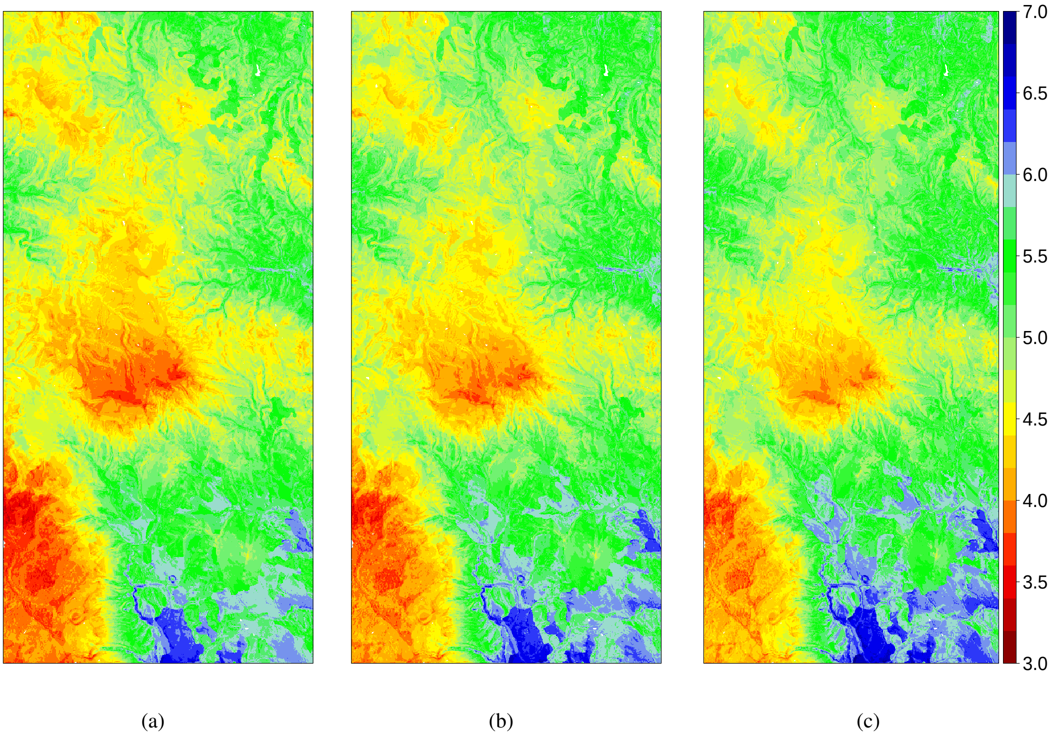


FIGURE 5.9: Final prediction maps of pH produced by 3D regression kriging with IntP trend model: a) 0.1 m depth; b) 0.2 m depth; c) 0.3 m depth.

5.4 Conclusion

This work presents the use of the shrinkage regression lasso method for building the 3D interaction linear trend models of three soil variables: arsenic concentration, soil organic matter (SOM) content and soil pH measured in water. The obtained models were further used as a part of 3D regression kriging for the interpolation over the entire 3D prediction domain. The main motivation behind the use of interaction models was to create a 3D trend model with the ability to change the effects of spatial covariates on a target variable with depth. Therefore, instead of considering all two-way interactions between all covariates, only the interactions between the spatial covariates and depth were considered. Two approaches for selecting the important interactions were examined: first, treating all the effects (main and interactions) individually; second, obeying the principle of hierarchy between the main and the interaction effects. Accordingly, two lasso implementations were compared: the widely popular R package glmnet with extremely efficient fitting procedure and hierNet R package, which were exclusively created for fitting interaction models subject to strong or weak hierarchy restriction.

The predictive power of each model was tested through the process of nested 5-fold cross-validation based on stratified samples. Samples were stratified based on a strategy specifically devised for this purpose. In general, the stratified sampling methodology applied in this study ensures that each sample is representative according to three criteria: 1) the lateral distribution of samples; 2) the depth of profiles and 3) the range of the observed target values. In general, the obtained results show that taking the interaction into account can improve the predictive capabilities of the trend model up to 20%. As expected, the greatest improvement was achieved with variables that have a strong decreasing trend along the depth, as well as a higher variation in the surface soil layers. Furthermore, it is evident that the inclusion of interactions contribute to the exclusion from the model of some less or moderately important variables that are eventually selected in the non-interaction model. In other words, the inclusion of interactions contribute to the sparsity of the model, in terms of main effects.

The spatial structure of 3D residuals was analyzed by computing the variograms in the vertical direction, in horizontal space as well as in 3D. The problem of the interval scale of residuals in the vertical sense was overcome by interpolating the residuals using the equal-area spline function. The geometric anisotropy was determined as a ratio between vertical and horizontal variogram ranges, and then incorporated in the 3D variogram model. The presence of even, moderate and spatially structured residuals was utilized for further residual interpolation. In all cases, geostatistical interpolation of residuals by 3D kriging has resulted in remarkable improvements in the accuracy of prediction.

Chapter 6

PenInt3D - package for 3D soil mapping based on penalized interaction models

6.1 Introduction

This chapter presents a set of functions, developed under the R environment, that constitute the core of the PenInt3D package, which is under development in our laboratory. The PenInt3D package is envisaged to provide a semi-automatic technique for exploring and mapping soil variables using linear penalized interaction methods within the 3D regression kriging framework. In short, the PenInt3D package provides the functions for exploring spatial continuity of raw profile data in 3D by:(1) making stratified data partitioning based on several criteria, including spatial location, soil depth and range of observed values of soil property; (2) building 3D interaction trend models based on soil profile data; (3) performing model selection based on stratified n-fold cross-validation; (4) performing model assessment through nested cross-validation; (5) exploring spatial continuity of residuals in horizontal and vertical directions; (6) fitting 3D variogram models; and (7) finally, making spatial prediction at specified depths by 3D regression kriging. PenInt3D combines the functionality of several R packages including: aqp, gstat, GSIF, glmnet, hierNet, plyr and ggplot. Therefore, the essentials of R environment, together with all the previously mentioned R packages, will be presented in the following sections.

6.2 R environment and related packages

R (Team, 2013) is a language and environment for statistical computation and graphics that provides programming facilities, high-level graphics, interfaces to other languages, and debugging facilities. R implements a language similar to the S language that was originally developed by John Chambers (Chambers, 2008). The main difference is in the license statement, because R is a free and open source software under the terms of the GNU General Public License in contrast to the S language. R is a fully functional interpreter which permits the creation of functions and calculations within an environment defined by a command line window or a graphical user interface (Grunsky, 2002). R is organized as a collection of packages designated for specific tasks.

The R package system has been one of the key factors in the overall success of the R project (Team, 2013). The R contains the base system that enables statistical computation, linear algebra computation, graphics creation, and other similar features. A package is a related set of functions, help, and data files that have been bundled up together. It is not necessary to install the specific packages if they are not necessary to the user.

6.2.1 aqp package

The aqp (Algorithms for Quantitative Pedology) package (Beaudette et al., 2013a) is an unavoidable tool for dealing with profile based soil data. As stated in the manual of the aqp, this package was developed to address some of the difficulties associated with processing soils information; specifically related to visualization, aggregation, and classification of soil profile data. The aqp package defines the S4 class, 'SoilProfileCollection' which is used for the storage of soil profile data. It also defines the methods of summarizing, printing, and plotting the soil data.

6.2.2 sp package

The sp package (Pebesma and Bivand, 2005) provides classes and methods for dealing with spatial data in R. The spatial data classes implemented are: points, grids, lines, rings and polygons (each with or without the attribute data).

6.2.3 gstat package

gstat (Pebesma, 2004) is the most accessible geostatistical package. It can be used to calculate sample variograms, fit valid models, plot variograms, calculate (pseudo) cross variograms, and calculate and fit directional variograms and variogram models. Ordinary and simple kriging, ordinary or simple co-kriging, universal kriging, external drift kriging, Gaussian conditional or unconditional simulation or co-simulation, can also be done.

6.2.4 GSIF package

GSIF (Global Soil Information Facilities) is a generic framework developed in the ISRIC institute to support the production of global soil information. The GSIF package for R is just one of several components of the GSIF framework. The GSIF R package contains tools to handle the soil data and produce gridded soil property maps using a fully automated approach. It implies that model fitting, prediction and visualization are run using fully automated and reproducible workflows, thereby providing easy access to new data integration, map updating and output validation. Methodologically, the GSIF R package implements the 3D regression kriging to provide the point based prediction of soil properties. In addition, the R package defines several S4 classes, including: 'geosample', 'GlobalSoilMaps', 'SoilGrids', 'FAO.SoilProfileCollection', and 'GlobalSoilMap'.

6.2.5 glmnet and hierNet packages

glmnet (Friedman et al., 2010) is a package that fits a generalized linear model via a penalized maximum likelihood. The regularization path is computed for the Lasso or elastic-net penalty, by using a grid of values for the regularization parameter. The algorithm is extremely fast. It uses cyclical coordinate descent, which successively optimizes the objective function over each parameter, while the others remain fixed, and cycles repeatedly until convergence.

HierNet (Bien et al., 2013) R package fits sparse interaction models for continuous and binary responses that are subject to the strong (or weak) hierarchy restriction.

6.3 PenInt3D

PenInt3D package[1] is designed to provide a semi-automatic method for the 3D mapping of soil variables, by using a combination of penalized interaction models and 3D ordinary kriging. The PenInt3D package contains several functions that are designed to perform two main tasks: model selection/prediction and model assessment. Both model selection and model assessment follow the two-step approach of regression kriging. The main difference between them lies in cross-validation. The model assessment procedure uses the stratified n-fold nested cross-validation to assess the accuracy, while the model selection procedure is based on the standard stratified n-fold cross-validation. Only the concepts and core functions of the PenInt3D package will be presented within the following sections.

6.3.1 Creating penint3D object

penint3D object is designed to be the common input for both the model selection/prediction and the prediction accuracy assessment tasks. The main motivation for the introduction of penint3D lies in the need for consistency within these two procedures. A penint3D object is designed to hold all the necessary input information for these two main tasks. By creating the penint3D object, double defining the same input parameters, as well as the repetition of common steps, such as spatial overlay, data-pre-processing and data-partitioning, are avoided. In order to create the penint3D object, a penint3D function must be run. There are a number of input arguments (parameters) that must be defined,

[1] https://github.com/pejovic/int3D

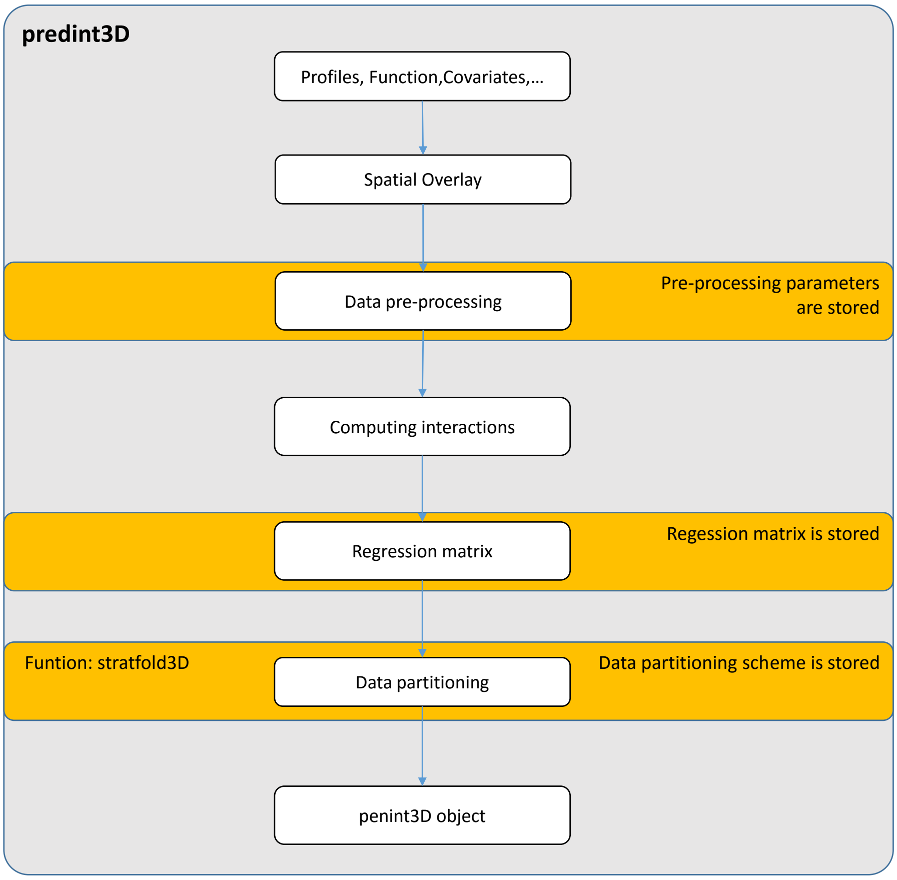


FIGURE 6.1: Creating of penInt3D object

including: profile data (as SoilProfileCollection object), gridded covariates (as SpatialPixelsDataFrame), formula object that relates the target variable and covariates (depth must be included in the formula), a number of folds, etc.

Figure 6.1 depicts the main steps in creating a penint3D object. At the beginning, spatial overlay is performed by extracting the covariate values and merging them with observations according to profile locations. The data-preprocessing step involves the standardization of continual variables to have zero mean and standard deviation equals one, and the dummy coding of categorical variables. The standardization parameters and coding schemes are stored in the penint3D object (item "pre-processing") for further usage. In the next step, the interactions between spatial covariates and depth are calculated and merged with other data to create an overall data matrix (item "data"). Considering that cross-validation is an integral part of both the model assessment and prediction procedures, the penint3D function performs the data-partitioning. The data partitioning step invokes a new function named stratfold3D that partitions the data into a number of stratified folds. The stratification can be performed in accordance with several parameters: 1) spatial distribution of observed target values; 2) profile depths and 3) range of measurements of target variables. The vector with indexes identifying what fold each observation contains, is also stored in the penint3D object (item 'folds'). This step is particularly important because properly conducted data partitioning ensures greater representativeness on characteristics of interest within each fold. For this reason, data partitioning can be checked by running the function plotfold3D with the item folds as input arguments. Running this functions generates the statistical 'summary' of each fold together with the 2D plots of each sample. For example, Table 5.7 depicts the summary statistics of created folds for SOM data, while the Figure 5.1 shows the spatial allocation of samples (folds). Finally, penint3D is an object of the 'list' class, with the following structure:

1. cogrids - *gridded covaraites*
2. data - *data matrix*
3. pre-processing - *pre-processing parameters. Mean and standard deviation for each continual variable and dummy coding scheme for each categorical variable*,
4. folds - *output from stratfold3D function*
5. lambda - *values of regularization parameters*

6.3.2 Prediction

The prediction involves the sequential running of two functions: predint3D and krige3D. Figure 6.2 shows the work-flow algorithm for the prediction procedure.

6.3.2.1 Trend modeling

predint3D function provides an automatic method for building a 3D trend model, which includes the interactions between spatial covariates and depth, using lasso to perform

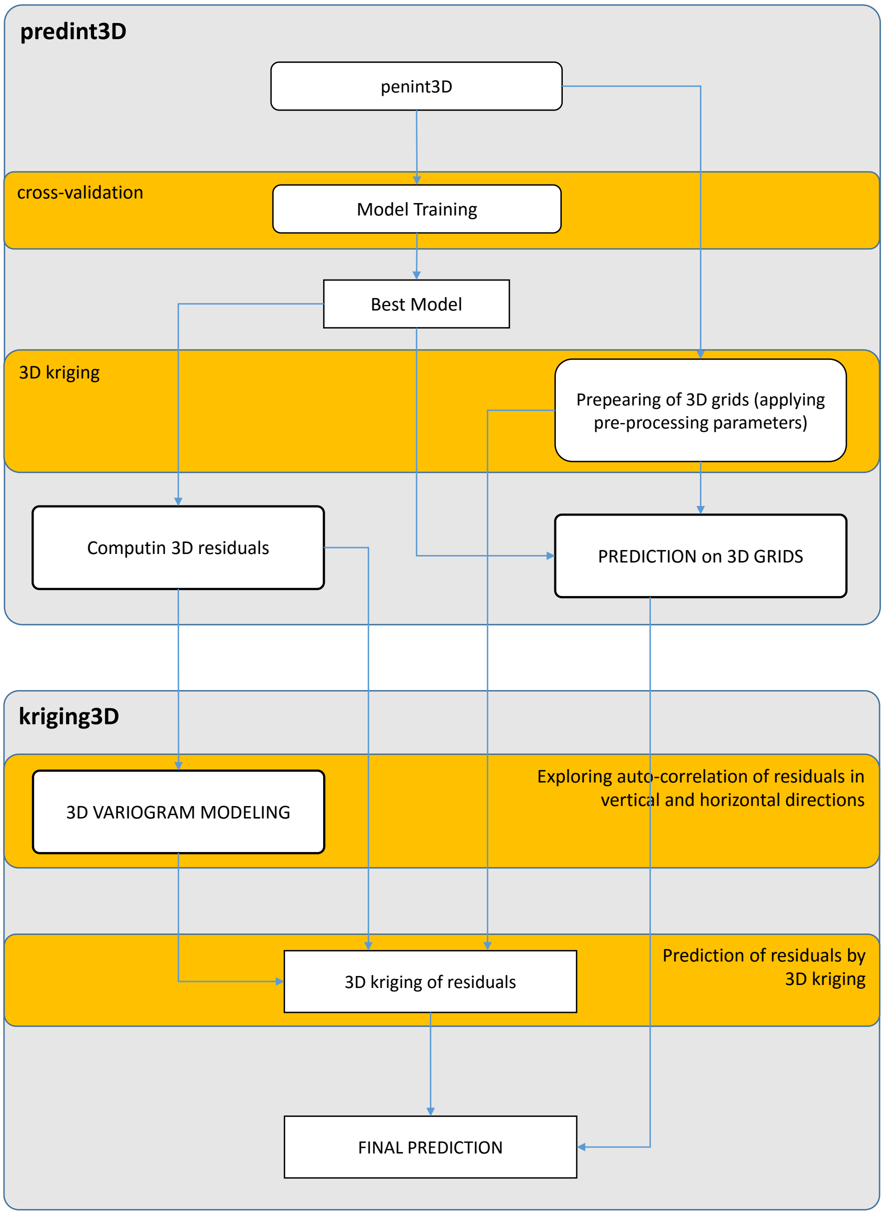


FIGURE 6.2: Prediction algorithm in PenInt3D package

model selection and parameter optimization. The optimal model is selected through the process of repeated n-fold cross-validation that is based on stratified samples stored in the penint3D object (item $folds), as well as on the pre-specified values of regularization parameter (item penint3D$lambda). Considering that the level of regularization (λ) directly affects the model settings, cross-validation error e_{cv} is computed. The optimal value of λ yields the lowest e_{cv}. The function decides whether the hierarchy constraints should be included or not, depending on whether the argument hier is TRUE or FALSE.

Once the best trend model is selected, the prediction can be performed in any 3D point within the sampling area. Predictions are usually required to be made over the entire area and at specific depths, so as to form a list of maps relating to different soil depths. Accordingly, all spatial covariates, which are included in the model, have to be prepared (standardized), by using the same pre-processing parameters, as when the model was built. The whole procedure of making prediction grids (3D grids) consists of the following steps that are executed automatically:

1. Creating a list of gridded covariates that are included in the final model as SpatialPixelsDataFrame (SPDF) (one SPDF per each prediction depth). The depth must be added as third coordinate.
2. Adding 'depth' (or polynomial depth terms) also as new variable(s) in each SPDF.
3. Using preprocessing parameters from penint3D$preprocessing to transform SPDFs.
4. Computing the interactions between spatial variables and depth terms for each SPDF.
5. Adding interactions to corresponding SPDF as new variables.

Once the 3D prediction grids are defined, the final model is run to produce trend predictions. Parallel to the prediction on grids, the final model is run to make a prediction on the observed data points, enabling the computation of the trend model residuals. Finally, the running of the predint3D function ends by generating the list object with the following structure:

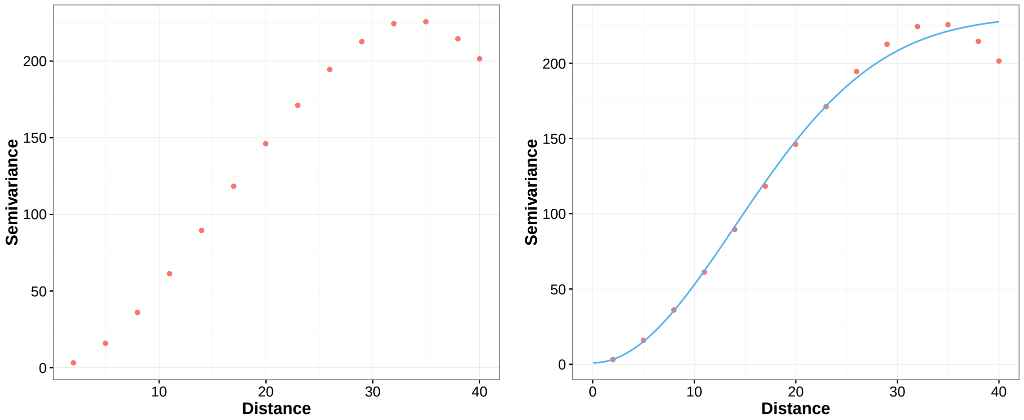
1. $prediction-SpatialPixelsDataFrame with predictions at different depths,

2. $summary- a list object consisting of: 1) Accuracy measures; 2) Model definition; 3) Preprocessing parameters; 4) Coefficients 5) Prediction (data frame with following variables: ID, longitude, latitude, depth, observed, predicted, and residuals.

6.3.2.2 Residual modeling and spatial prediction

krige3D function provides residual dependency analysis, 3D variogram modeling and 3D kriging prediction. Each task requires the running of the krige3D function separately with different input arguments.

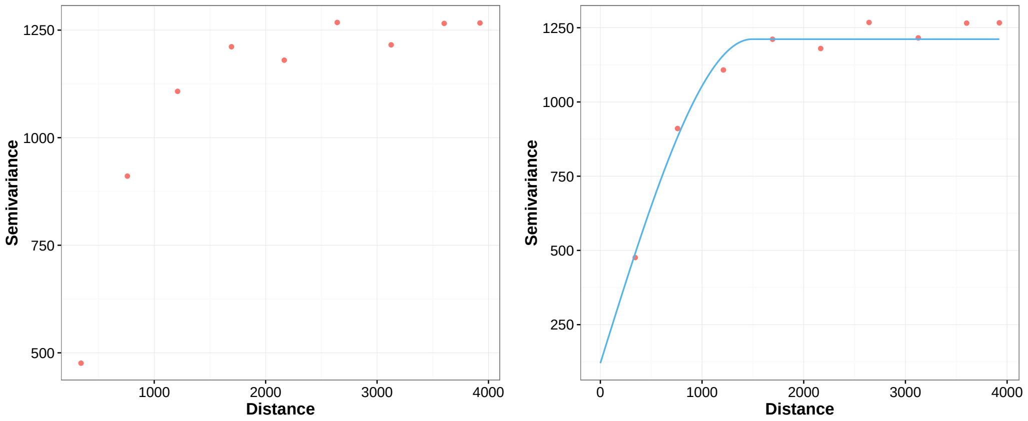
Residual dependency analysis refers to the process of exploring the sample residual variograms in horizontal and vertical directions. The main objective of this phase is to detect the range of spatial dependency in both directions and to establish the level of anisotropy as the ratio between two ranges. For this purpose, krige3D function requires as input: 1) output from the predict3D function; 2) horizontal and vertical 'cutoff'; 3) horizontal and vertical 'width' parameter. The variogram calculation and modeling routines are provided using the variogram and fit.variogram methods from the gstat R package. However, the computation of variograms in vertical directions is met with two main difficulties. The first issue rests in the fact that soil profile measurements refer to the specific soil depth intervals, which are mostly soil horizons. In point scale geostatistics, such measurements refer to the horizon mid-point, resulting in the loss of sensitivity in detecting local variation in vertical direction. The second issue is that the number of measurements in a soil profile is usually small (often between 2 and 5), and is mostly taken from upper soil layers that may introduce an additional uncertainty in variogram estimation. In order to overcome these issues, the krige3D function uses the *mass-preserving* spline function to interpolate the residuals between the horizon mid-points. In this way, the step-wise form of the residuals in each profile is modeled by a continuous function providing the possibility to compute semi-variances for any vertical distance with a sufficient amount of data. However, such a transformation introduces an additional error in the characterization of vertical variation. But, as long as the the main objective is only to establish the range of spatial dependency in the vertical direction, it makes sense. Figure 6.3(a) and Figure 6.4(a) show the plots of sample variograms computed in vertical and horizontal directions, respectively.



(a) (b)

FIGURE 6.3: Vertical residual variogram: a) Sample variogram b) Fitted variogram model

Based on visual inspection of variogram graphs, the theoretical variogram models with similar shape can be selected. The selected variogram models can be fitted by re-running the krige3D function, whereby, the selected variogram model is provided as an input argument. The variogram model must be defined as variogramModel class (output of vgm method from gstat package). Figure 6.3(b) and Figure 6.4(b) show the fitted variogram models plotted on graphs of sample variograms in vertical and horizontal directions, respectively.

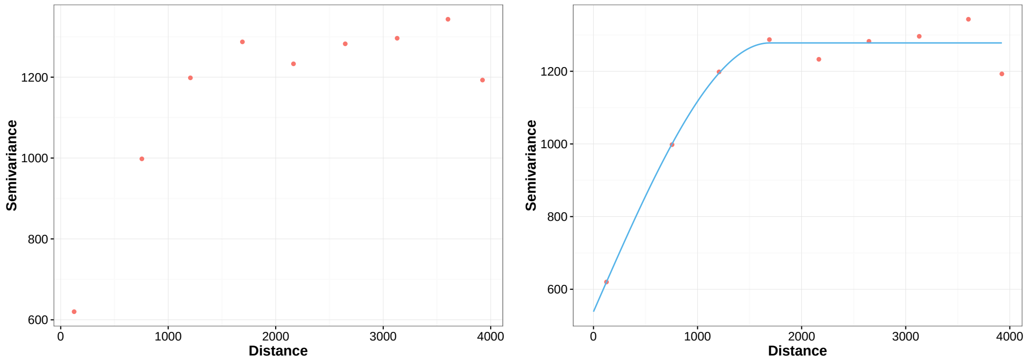


(a) (b)

FIGURE 6.4: Horizontal residual variogram: a) Sample variogram b) Fitted variogram model

If the fitted variogram models look reasonable, they can be adopted. If not, the other model should be selected and fitted as long as the chosen criterion is not satisfied, which is met through either visual inspection or statistical measures of goodness of fit.

The second phase involves the 3D variogram modeling. In the current phase of development of the PenInt3D package, krige3D enables the user to create only the single-structure 3D variogram model. Once the horizontal and vertical variograms are fitted, the anisotropy ratio is automatically determined as a ratio between the two variogram ranges. The 3D residual variogram graph can also be explored in order to define the appropriate theoretical model. Estimated anisotropy is automatically incorporated into initial parameters for the theoretical model of a 3D variogram. Therefore, it is only required to provide the initial variogram parameters without the additional parameters that pertain to the geometrical anisotropy. Figure 6.5 shows the 3D sample residual variogram (a) and the fitted 3D variogram model (b):



(a)

(b)

FIGURE 6.5: 3D residual variogram: a) Sample variogram b) Fitted variogram model

The last phase refers to the 3D prediction. Once the 3D variogram model is adopted, residual 3D kriging can be run. Depending on the logical argument krige=TRUE/FALSE, the krige3D function decides whether the prediction should be run. By running the 3D kriging on 3D prediction grids, the final maps are produced and stored in the output as 'list' of SpatialPixelsDataFrame objects.

6.3.3 Accuracy assessment

The entire accuracy assessment in the PenInt3D package is based on nested n-fold cross-validation. The implementation of nested cross-validation for the two-step procedure, such as regression kriging, implies the ensurance of consistency throughout the whole modeling procedure. This means that the same test data (test folds) must be used for assessment of both the trend model and for 3D kriging of the residuals.

Figure 6.6 depicts the flow-chart of the nested cross-validation procedure for 3D regression kriging implemented in the PenInt3D package. Similarly, as for the 'prediction' procedure, model assessment requires the sequential running of two functions: penint3Dncv and krige3Dncv with the penint3D object as a starting input parameter.

Accuracy assessment starts by running the penint3Dncv function with the aim to assess the accuracy of the trend model. In the first step, the first fold (test data) is held out, while the other $k-1$ folds (training data) enter the inner loop. Within the inner loop of the nested cross-validation (see Algorithm 2), stratfold3D function splits the training data into new k stratified folds that are then used for model selection (model training) through standard cross-validation. The resultant optimal model is then run to predict on both the 'outer' test data, as well as on the training data set. Therefore, each step of the outer loop involves the storing of one optimal trend model along with the associated *test* and the *training* predictions, within the *prediction storage*. Consequently, at the end of the process, the prediction storage will contain the k sets of test and training predictions. Each set corresponds to one data partitioning from the outer loop.

Calling the penint3Dncv function generates the object with the following structure:

1. $measure - data frame with accuracy measures (R^2 and RMSE) calculated for each outer-loop step as well as for overall test prediction
2. $coef - sparse matrix of class 'dgCMatrix' with models coefficients.
3. $folds - penint3D$folds
4. $train.results - list of k data frames that hold the predictions of each model on training data. Each data frame contains six columns: ID, observed, predicted, longitude, latitude, and depth.

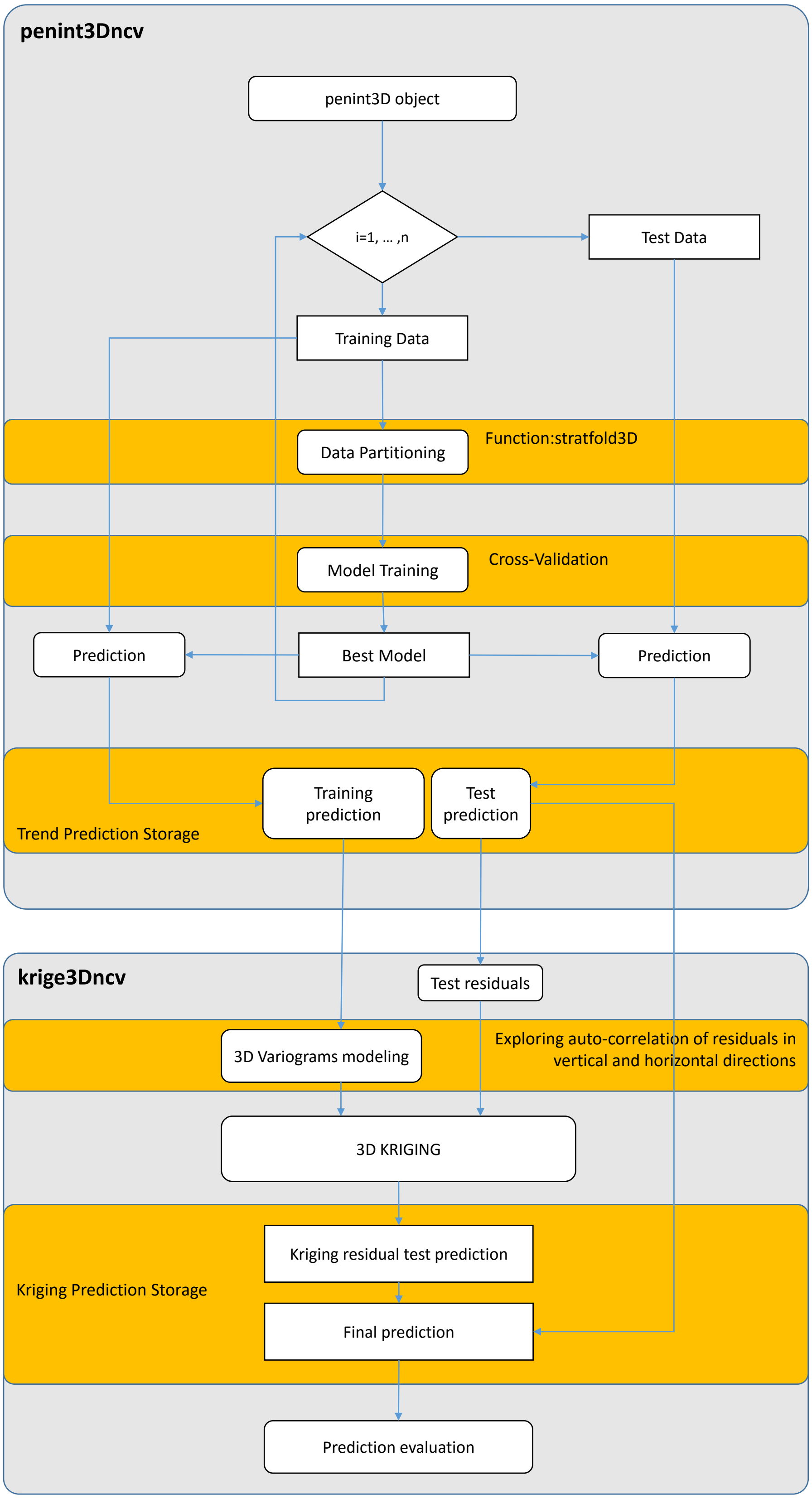


FIGURE 6.6: Nested cross-validation procedure for 3D regression kriging implemented in PenInt3D package

5. $test.results - list of *k* data frames that hold the predictions of each model on test data. Each data frame contains six columns: ID, observed, predicted, longitude, latitude, and depth.

The next step of the accuracy assessment procedure refers to the assessment of 3D regressin kriging prediction. At the begining, 3D spatial dependence structure of each set of training residuals must be checked. The krige3Dncv function provides the possibility for jointly processing 3D variogram modeling in the same way as it was done for the prediction modeling. Each training prediction may have a different structure. However, the spatial dependence structure would not differ too much considering that the majority of data are common (for 5-fold cross-validation 80% of data are common). Nevertheless, the first call of the krige3Dncv function in this phase is intended for exploring the horizontal and the vertical sample residual variograms. Therefore, the initial parameters for theoretical variogram models are not required. After visual inspection, theoretical models can be fitted to each residual variograms separately, or one to all, depending on how data sets differ between each other. Figure 6.7, Figure 6.8 and Figure 6.9 show the 1D, 2D and the 3D variogram models fitted to each sample variogram. Each variogram corresponds to one training set.

Each 3D variogram model is then used for 3D regression kriging of the test residuals. The final 3D regression kriging prediction is obtained by adding the test residual kriging prediction to the test trend model prediction. The final accuracy can then be assessed by combining the observed data with final test predictions and computing the accuracy measures.

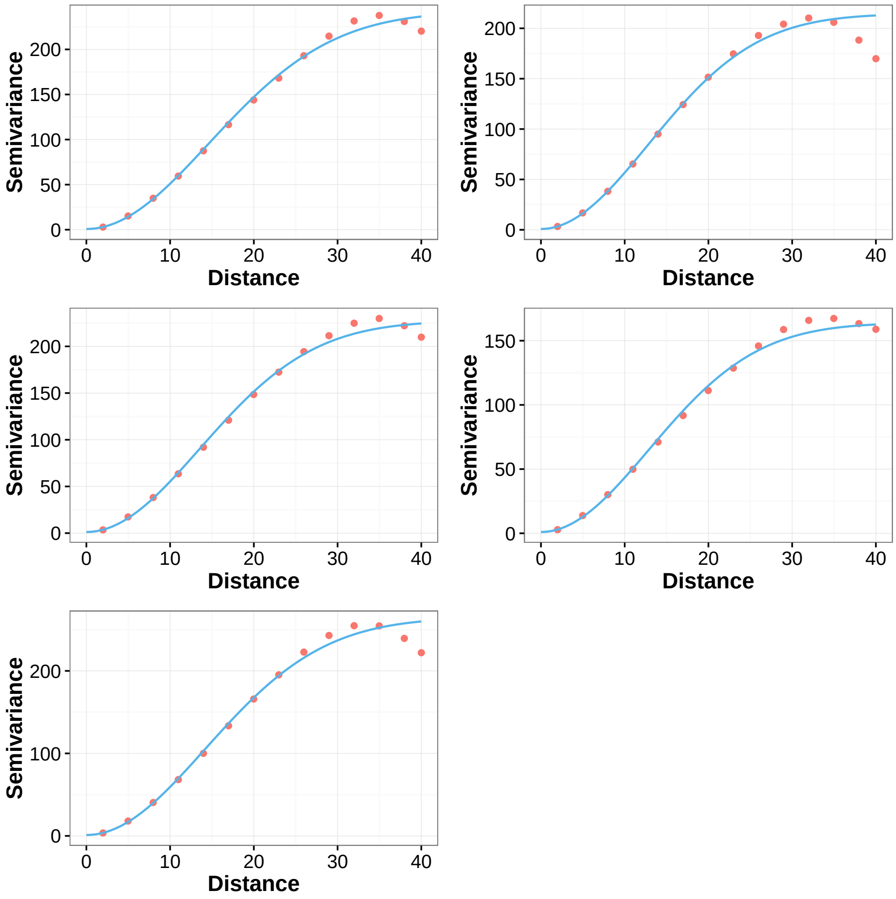


FIGURE 6.7: Five vertical variograms fitted to each training data set within the 5-fold nested cross-validation

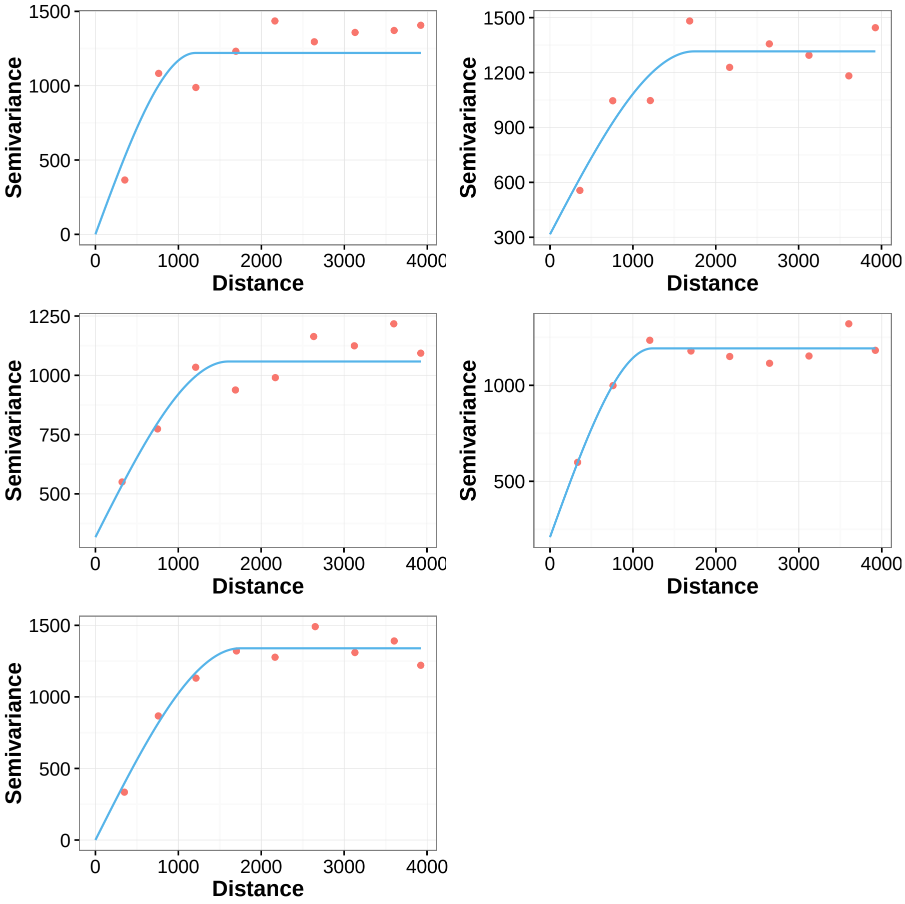


FIGURE 6.8: Five horizontal variograms fitted to each training data set within the 5-fold nested cross-validation

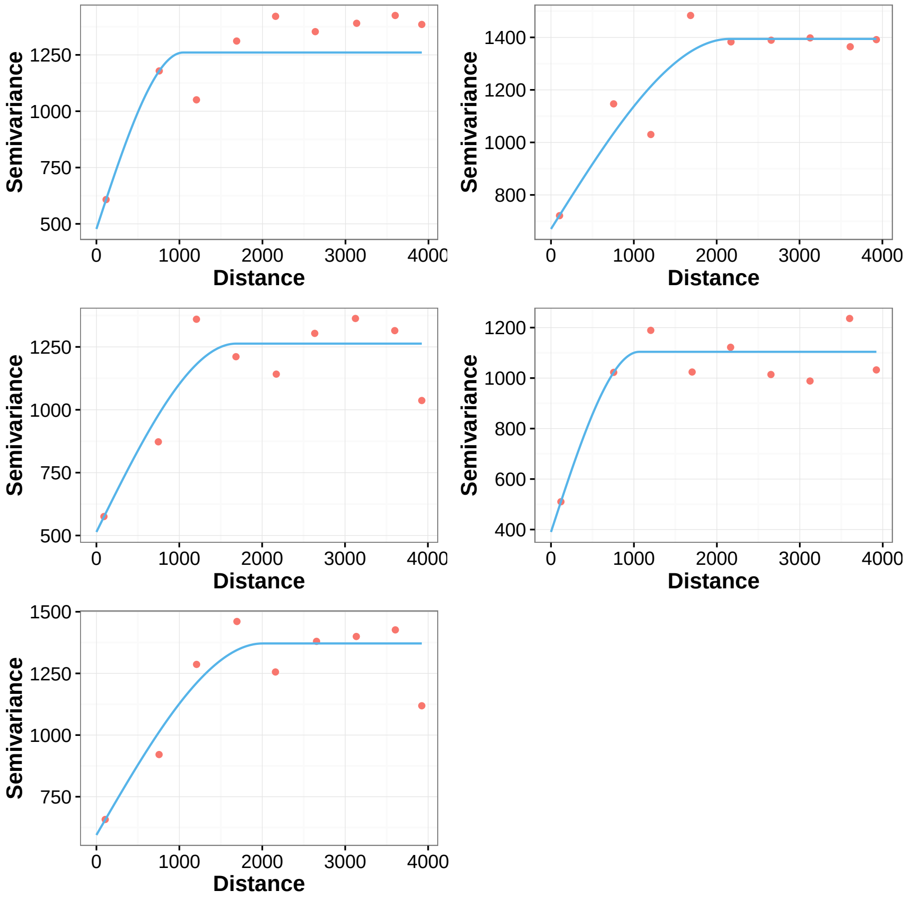


FIGURE 6.9: Five 3D variograms fitted to each training data set within the 5-fold nested cross-validation

Chapter 7

Conclusion

The aim of this thesis was to contribute to the development of a geostatistical approach for the 3D mapping of soil properties. The contributions are primarily reflected in investigating the existing methods, as well as in proposing inovative modeling and accuracy evaluation approaches. The most important results of this thesis are presented in Chapters 4, 5 and 6.

In Chapter 4, the two aspects of terrain exposure - geometrical and topographical - were considered and incorporated into the Spline-Than-Krige approach to produce maps of atmospherically deposited arsenic concentration at different soil depths. The aim of this research was to explore the extent to which the commonly available information, such as the prevailing wind direction or location of the source of pollution, in combination with digital elevation model (DEM), can be used to improve the spatial prediction of the deposited pollutants at several soil depths. The geometrical aspect was approximated by two parameters that quantify the exposure of any cell, regarding the distance to the source of pollution, or angular distance from the major wind direction which is measured by the location of source. On the other hand, topographic parameters take into consideration the neighboring topography of each cell in order to quantify the topographical protection from the wind. Therefore, the topographical exposure was quantified by: DEM, modified Exposure towards the Wind Flux (EWF) index and the Morphometric Protection Index (MPI). A modification of EWF index was performed to account for the location of the pollution source, with the aim of emphasizing the effects of topographical exposure to the known source. This modification was achieved by replacing the wind direction with the azimuth between each grid cell and the source of pollution. In this regard, only the topography along this direction was considered. This new parameter was denoted as Exposure toward the Source (ES). Regression analysis confirmed the presence of a significant statistical association between the As data and all exposure parameters. The trend model showed good overall accuracy, explaining 52% of the variance in As data for the surface soil layer (0-5 cm), 49% for the middle layer (5-15 cm), and 35% for the deepest layer (15-30 cm). Relative predictors importance analysis revealed the importance of considering a more general model that includes interactions between exposure parameters. The kriging interpolation of residuals improved to some extent, the regression accuracy for all three layers with R^2 values ranging from 55% for the surface layer to 36% for the deepest soil layer. Generally, a relatively high RMSE characterizes the prediction on each soil layer and indicates that such a model cannot still be used for decision making purposes. However, in a situation when the wind has an important role on the spatial distribution of soil pollutants, the integration of topographic exposure parameters could be useful for a prediction even on deeper soil layers.

Chapter 5 focuses on the introduction of the shrinkage regression method lasso within the 3D regression kriging, as a tool for creating the 3D interaction model of soil properties. The main motivation of creating interaction models is to provide better utilization of spatial covariates, by allowing their interaction with soil depth. One of the principal advantages of using lasso to fit the 3D model lies in its ability to automatically select all the important variables, including the interaction terms, simultaneously with the optimization of the parameters, in order to provide the best possible prediction. As a result, an interpretable predictive 3D trend model is selected. It is important to emphasize here that the consideration of two-way interactions, even between the spatial covariates and depth, doubles the number of predictors that must be considered. For that reason, an automatic selection of important variables is preferred. Considering that the model selection in the case of lasso is based on selecting the shrinkage parameter that minimizes the mean prediction error, the complete process runs through stratified n-fold cross-validation. However, the resulting model is still linear and therefore retains all the limitations related to linear models, such as high bias or sensitivity to outliers. Residuals of the regression model are analyzed for dependency in 3D space to check the applicability of 3D kriging. The presence of residual dependency in a vertical and horizontal sense is then used for constructing the anisotropic lag distance and ultimately the anisotropic 3D variogram model. Once the 3D residual variogram is modeled, 3D kriging of residuals can be run. The final prediction is obtained by summing the estimates of trend models and the estimates of the 3D kriging of residuals. The inventiveness of the methodology proposed in this work largely lies in the implementation of nested n-fold cross-validation as a method for accessing the prediction accuracy of the two-step procedure of the 3D regression kriging. Nested n-fold cross-validation, in each step, keeps one fold (testing set) independent of the data used for modeling the trend and residual variogram. In this way, the computation of a more reliable accuracy measures is provided. This process is especially recommended for models for which the trend model is selected via the standard n-fold cross-validation. The results of applying the proposed methodology to the profile data of arsenic concentration, SOM content and soil pH (measured in H_2O) generally emphasizes the importance of the inclusion of the interactions between spatial covariates and depth in the 3D trend model. The amount of benefits that could be received from such a model largely depends on causal linkage between the modeled variable and the environmental factors. The interactions would be more valuable if the depth-wise distribution of the modeled variable is affected by the environmental factors more. For that reason, these factors should be approximated by spatial covariates as close as possible. Therefore, the 3D interaction models for As concentration and SOM content yielded a 20% improvement over the non-interaction models. A detailed spatial and depth-wise explanatory data analysis is also highly recommended. It can provide sufficient information as to whether, and to what extent, the soil is affected by external influences. Today, it can be easily conducted by using a variety of packages which are exclusively developed for this purpose, such as: `aqp`, `sp`, `spatstat`, and many other R packages.

Another important consequence of considering the interactions is related to the model selection. Generally, the inclusion of interactions between spatial covariates and depth has lead to models that have a smaller number of variables than non-interaction models, i.e. more sparse structure. Even if more predictors (main effects and interactions) are included in the interaction models, a smaller number of individual variables actually participates in the model, since interactions are not the newly introduced variables, but rather the derivates of the existing variables.

The complete computational framework was implemented in the set of R functions created by the author of this thesis, with the aim to constitute an R package (penint3D) for 3D soil mapping, by using interaction penalized models. The R package penint3D uses profile based soil data and gridded spatial covariates as main input arguments. The package is still under development and future work will be mostly concentrated on improving the modeling of the 3D covariance structure.

Bibliography

Adhikari, K., Hartemink, A. E., Minasny, B., Kheir, R. B., Greve, M. B., and Greve, M. H. (2014). Digital mapping of soil organic carbon contents and stocks in Denmark.

Adhikari, K., Kheir, R. B., Greve, M. B., Bøcher, P. K., Malone, B. P., Minasny, B., McBratney, A. B., and Greve, M. H. (2012). High-Resolution 3-D Mapping of Soil Texture in Denmark. *Soil Science Society of America Journal*, 77(3):1–17.

Akaike, H. (1974). A new look at the statistical model identification. *Automatic Control, IEEE Transactions on*, 19(6):716–723.

Antonić, O. and Legović, T. (1999). Estimating the direction of an unknown air pollution source using a digital elevation model and a sample of deposition. *Ecological Modelling*, 124(1):85–95.

Arrouays, D., McBratney, A. B., Minasny, B., Hempel, J. W., Heuvelink, G. B. M., MacMillan, R. A., Hartemink, A. E., Lagacherie, P., and McKenzie, N. J. (2014). The GlobalSoilMap project specifications. *GlobalSoilMap: Basis of the global spatial soil information system*, page 9.

Bajat, B., Pejović, M., Luković, J., Manojlović, P., Ducić, V., and Mustafić, S. (2013). Mapping average annual precipitation in serbia (1961–1990) by using regression kriging. *Theoretical and applied climatology*, 112(1-2):1–13.

Beaudette, D., Roudier, P., and O'Geen, A. (2013a). *Algorithms for Quantitative Pedology: A Toolkit for Soil Scientists.*

Beaudette, D. E., Roudier, P., and O'Geen, a. T. (2013b). Algorithms for quantitative pedology: A toolkit for soil scientists. *Computers and Geosciences*, 52:258–268.

Beckett, P. H. T. and Webster, R. (1971). Soil variability: a review. *Soils and fertilizers*, 34(1):1–15.

Bien, J., Taylor, J., and Tibshirani, R. (2013). A lasso for hierarchical interactions. *Annals of Statistics*, 41(3):1111–1141.

Bien, J. and Tibshirani, R. (2014). *hierNet: A Lasso for Hierarchical Interactions*. R package version 1.6.

Bishop, T. F. a., McBratney, a. B., and Laslett, G. M. (1999). Modeling soil attribute depth functions with equal-area quadratic smoothing splines. *Geoderma*, 91(1-2):27–45.

Brus, D. J., Yang, R. M., and Zhang, G. L. (2016). Three-dimensional geostatistical modeling of soil organic carbon: A case study in the Qilian Mountains, China. *Catena*, 141:46–55.

Burke, I. C., Yonker, C., Parton, W., Cole, C., Schimel, D., and Flach, K. (1989). Texture, climate, and cultivation effects on soil organic matter content in us grassland soils. *Soil science society of America journal*, 53(3):800–805.

Cambardella, C., Moorman, T., Parkin, T., Karlen, D., Novak, J., Turco, R., and Konopka, A. (1994). Field-scale variability of soil properties in central iowa soils. *Soil science society of America journal*, 58(5):1501–1511.

Chambers, J. (2008). *Software for data analysis: programming with R*. Springer Science & Business Media.

Cox, D. R. (1984). Interaction. *International Statistical Review / Revue Internationale de Statistique*, 52(1):1.

Cressie, N. (1993). Statistics for Spatial Data: Wiley Series in Probability and Statistics.

Dayani, M. and Mohammadi, J. (2010). Geostatistical assessment of Pb, Zn and Cd contamination in near-surface soils of the urban-mining transitional region of Isfahan, Iran. *Pedosphere*, 20(5):568–577.

De Visscher, A. (2014). *Air Dispersion Modeling. Foundations and Applications*, volume 53. John Wiley & Sons.

Diggle, P. J. (2011). Model-based Geostatistics Peter J Diggle Lancaster University and Johns Hopkins University School of Public Health July 2011 Approximate timetable. (July).

EPA, S. (1995). User's Guide for the Industrial Source Complex (ISC3) Dispersion Models, Volume I-User Instructions. Technical report, EPA-454/B-95-003a. Office of Air Quality Planning and Standards, US Environmental Protection Agency, Research Triangle Park, NC 27711.

Erickson, T. A., Williams, M. W., and Winstral, A. (2005). Persistence of topographic controls on the spatial distribution of snow in rugged mountain terrain, Colorado, United States. *Water Resources Research*, 41(4).

Eswaran, H., Lal, R., Reich, P., et al. (2001). Land degradation: an overview. *Responses to Land degradation*, pages 20–35.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, 33(1):1.

García-Sánchez, a., Alonso-Rojo, P., and Santos-Francés, F. (2010). Distribution and mobility of arsenic in soils of a mining area (Western Spain). *Science of the Total Environment*, 408(19):4194–4201.

Gessler, P. E., Moore, I. D., McKenzie, N. J., and Ryan, P. J. (1995). Soil-landscape modelling and spatial prediction of soil attributes. *International journal of geographical information systems*, 9(4):421–432.

Goovaerts, P. (1999a). Geostatistics for Natural Resources Evaluation. *Journal of Environment Quality*, 28(3):1044.

Goovaerts, P. (1999b). Geostatistics in soil science: State-of-the-art and perspectives. *Geoderma*, 89(1-2):1–45.

Goovaerts, P. (2001). Geostatistical modelling of uncertainty in soil science. *Geoderma*, 103(1-2):3–26.

Goovaerts, P., Trinh, H. T., Demond, A., Franzblau, A., Garabrant, D., Gillespie, B., Lepkowski, J., and Adriaens, P. (2008a). Geostatistical modeling of the spatial distribution of soil dioxins in the vicinity of an incinerator. 1. Theory and application to Midland, Michigan. *Environmental science & technology*, 42(10):3648–3654.

Goovaerts, P., Trinh, H. T., Demond, A. H., Towey, T., Chang, S. C., Gwinn, D., Hong, B., Franzblau, A., Garabrant, D., Gillespie, B. W., Lepkowski, J., and Adriaens, P. (2008b). Geostatistical modeling of the spatial distribution of soil dioxin in the vicinity of an incinerator. 2. Verification and calibration study. *Environmental Science and Technology*, 42(10):3655–3661.

Goovaerts, P., Webster, R., and Dubois, J.-P. (1997). Assessingthe risk of soil contamination in the Swiss Jura using indicator geostatistics. *Environmental and Ecological Statistics*, 4:31–48.

Grömping, U. and Others (2006). Relative importance for linear regression in R: the package relaimpo. *Journal of statistical software*, 17(1):1–27.

Grunsky, E. (2002). R: a data analysis and statistical programming environment–an emerging tool for the geosciences. *Computers & Geosciences*, 28(10):1219–1222.

Guastaldi, E. and Del Frate, A. A. (2012). Risk analysis for remediation of contaminated sites: The geostatistical approach. *Environmental Earth Sciences*, 65(3):897–916.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). The Elements of Statistical Learning. *Elements*, 1:337–387.

Hastie, T., Tibshirani, R., and Wainwright, M. S. (2015). *Statistical Learning with Sparsity: The Lasso and Generalizations*.

Hengl, T. (2015). GSIF: Global Soil Information Facilities.

Hengl, T., de Jesus, J. M., MacMillan, R. a., Batjes, N. H., Heuvelink, G. B. M., Ribeiro, E., Samuel-Rosa, A., Kempen, B., Leenaars, J. G. B., Walsh, M. G., and Gonzalez, M. R. (2014). SoilGrids1km–global soil information based on automated mapping. *PloS one*, 9(8):e105992.

Hengl, T. and Heuvelink, G. B. M. (2013). *Global Soil Information Facilities*. ISRIC. in preparation.

Hengl, T., Heuvelink, G. B. M., Kempen, B., Leenaars, J. G. B., Walsh, M. G., Shepherd, K. D., Sila, A., MacMillan, R. A., Mendes de Jesus, J., Tamene, L., and Tondoh, J. E. (2015). Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. *PloS one*, 10(6):e0125814.

Hengl, T., Heuvelink, G. B. M., and Rossiter, D. G. (2007). About regression-kriging: From equations to case studies. *Computers and Geosciences*, 33(10):1301–1315.

Hengl, T., Heuvelink, G. B. M., and Stein, A. (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. *Geoderma*, 120(1-2):75–93.

Heuvelink, G. B. M. and Griffith, D. A. (2010). Space-time geostatistics for geography: A case study of radiation monitoring across parts of Germany. *Geographical Analysis*, 42(2):161–179.

Isaaks, E. H. and Srivastava, R. M. (1990). *An Introduction to Applied Geostatistics*. Oxford University Press, 1 edition.

James, G., Witten, D., Hastie, T., and Tibshirani, R. (2013). *An Introduction to Statistical Learning with Applications in R*.

Jenny, H. (1941). Factors of Soil Formation. *Soil Science*, 52(5):415.

Jones, Z. and Linder, F. (2015). Exploratory data analysis using random forests. In *Prepared for the 73rd annual MPSA conference*.

Kanevski, M. (2013). A Methodology for Automatic Analysis and Modeling of Spatial Environmental Data. (c):105–107.

Khalil, a., Hanich, L., Bannari, a., Zouhri, L., Pourret, O., and Hakkou, R. (2013). Assessment of soil contamination around an abandoned mine in a semi-arid environment using geochemistry and geostatistics: Pre-work of geochemical process modeling with numerical models. *Journal of Geochemical Exploration*, 125:117–129.

Kilibarda, M. and Bajat, B. (2012). plotgooglemaps: The r-based web-mapping tool for thematic spatial data. *Geomatica*, 66(1):37–49.

Kilibarda, M., Hengl, T., Heuvelink, G. B. M., Gräler, B., Pebesma, E. J., Perčec Tadić, M., and Bajat, B. (2014). Journal of Geophysical Research: Atmospheres. *Journal of Geophysical Research: Atmospheres*, 119:2294–2313.

Kisić, I. (2012). *Sanacija onečišćenog tla*. Agronomski fakultet Sveučilišta.

Komnitsas, K. and Modis, K. (2006). Soil risk assessment of As and Zn contamination in a coal mining region using geostatisretics. *Science of the Total Environment*, 371(1-3):190–196.

Kovačević, R., Jovašević-stojanović, M., Tasić, V., Milošević, N., Petrović, N., Stanković, S., and Matić-besarabić, S. (2010). Preliminary analysis of Levels of Arsenic and other metalic elements In PM10 sampled Near Copper Smelter Bor (Serbia)*. *Chemical Industry & Chemical Engineering Quarterly*, 16(3):269–279.

Krige, D. (1951). *A statistical approach to some mine valuation and allied problems on the Witwatersrand: By DG Krige*.

Krstajic, D., Buturovic, L. J., Leahy, D. E., and Thomas, S. (2014). Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics*, 6(1):1–15.

Kuhn, M. (2008). Building Predictive Models in R Using the caret Package. *Journal Of Statistical Software*, 28(5):1–26.

Lacoste, M., Minasny, B., McBratney, A., Michot, D., Viaud, V., and Walter, C. (2014). High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. *Geoderma*, 213:296–311.

Lagacherie, P. and Mcbratney, A. B. (2007). Spatial Soil Information Systems and Spatial Soil Inference Systems : Perspectives for Digital Soil Mapping. *Development in Soil Science*, 31(2004):3–22.

Lark, R. M., Cullis, B. R., and Welham, S. J. (2006). On spatial prediction of soil properties in the presence of a spatial trend: The empirical best linear unbiased predictor (E-BLUP) with REML. *European Journal of Soil Science*, 57(6):787–799.

Lindeman, R. H., Merenda, P. F., and Gold, R. Z. (1980). *Introduction to bivariate and multivariate analysis*. Scott, Foresman Glenview, IL.

Lindsay, J. B. and Rothwell, J. J. (2008). Modelling Channelling and Deflection of Wind. *Advances in Digital Terrain Analysis*, page 383.

Malone, B. P., Jha, S. K., Minasny, B., and 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.

Malone, B. P., McBratney, A. B., Minasny, B., and Laslett, G. M. (2009). Mapping continuous depth functions of soil carbon storage and available water capacity. *Geoderma*, 154(1):138–152.

Matheron, G. (1963). Principles of geostatistics. *Economic Geology*, 58(8):1246–1266.

Matheron, G. (1969). Le krigeage universel.

McBratney, A. B., Odeh, I. O. a., Bishop, T. F. a., Dunbar, M. S., and Shatar, T. M. (2000). *An overview of pedometric techniques for use in soil survey*, volume 97.

McBratney, A. B., Santos, M. M., and Minasny, B. (2003). *On digital soil mapping*, volume 117.

McKenzie, N. J. and Ryan, P. J. (1999). Spatial prediction of soil properties using environmental correlation. *Geoderma*, 89(1-2):67–94.

Meirvenne, M. V., Maes, K., and Hofman, G. (2003). Three-dimensional variability of soil nitrate-nitrogen in an agricultural field. *Biology and Fertility of Soils*, 37(3):147–153.

Michéli, E., Schad, P., Spaargaren, O., Dent, D., and Nachtergaele, F. (2006). *World reference base for soil resources: 2006: a framework for international classification, correlation and communication*. FAO.

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

Minasny, B., McBratney, A. B., and Salvador-Blanes, S. (2008). Quantitative models for pedogenesis - A review. *Geoderma*, 144(1):140–157.

Moore, A. W., Russel, J. S., and Ward, W. T. (1972). NUMERICAL ANALYSIS OF SOILS: A COMPARISON OF THREE SOIL PROFILE MODELS WITH FIELD CLASSIFICATION. *Journal of Soil Science*, 23(2):193–209.

Moore, I. D., Gessler, P., Nielsen, G. A., and Peterson, G. A. (1993). Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal*, 57(2):443–452.

Mulder, V. L., Lacoste, M., Richer-de Forges, A. C., Martin, M. P., and Arrouays, D. (2016). National versus global modelling the 3D distribution of soil organic carbon in mainland France. *Geoderma*, 263:16–34.

Nestorov, I., Protic, D., and Nikolic, G. (2007). Land cover mapping in serbia. *Wetlands*, 21176:0–27.

Odeha, I., a.B. McBratney, and Chittleborough, D. (1994). Spatial prediction of soil properties from landform attributes derived from a digital elevation model. *Geoderma*, 63(3-4):197–214.

Odgers, N. P., Libohova, Z., and Thompson, J. a. (2012). Equal-area spline functions applied to a legacy soil database to create weighted-means maps of soil organic carbon at a continental scale. *Geoderma*, 189-190:153–163.

Oliver, M. A. and Webster, R. (2014). A tutorial guide to geostatistics: Computing and modelling variograms and kriging. *Catena*, 113:56–69.

Oliver, M. A. and Webster, R. (2015). The Variogram and Modelling. In *Basic Steps in Geostatistics: The Variogram and Kriging*, pages 15–42. Springer.

Orton, T. G., Pringle, M. J., and Bishop, T. F. A. (2016). A one-step approach for modelling and mapping soil properties based on profile data sampled over varying depth intervals. *Geoderma*, 262:174–186.

Parton, W. J., Schimel, D. S., Cole, C., and Ojima, D. (1987). Analysis of factors controlling soil organic matter levels in great plains grasslands. *Soil Science Society of America Journal*, 51(5):1173–1179.

Pebesma, E. J. (2004). Multivariable geostatistics in s: the gstat package. *Computers & Geosciences*, 30(7):683–691.

Pebesma, E. J. (2006). The role of external variables and GIS databases in geostatistical analysis. *Transactions in GIS*, 10(4):615–632.

Pebesma, E. J. and Bivand, R. S. (2005). Classes and methods for spatial data in r. *R news*, 5(2):9–13.

Perović, G. (2005). *Least Squares*. Faculty of Civil Engineering.

Plattner, C., Braun, L. N., and Brenning, a. (2004). Spatial variability of snow accumulation on Vernagtferner, Austrian Alps, in winter 2003/2004. *Zeitschrift für Gletscherkunde und Glazialgeologie*, 39(1):43–57.

Ponce-Hernandez, R., Marriott, F. H. C., and Beckett, P. H. T. (1986). An improved method for reconstructing a soil profile from analyses of a small number of samples. *Journal of Soil Science*, 37(3):455–467.

SAGA, G. I. S. (2014). System for automated geoscientific analyses. *Online) http://www. sagagis. org/en/index. html.*

Saito, H. and Goovaerts, P. (2001). Accounting for source location and transport direction into geostatistical prediction of contaminants. *Environmental science & technology*, 35(24):4823–4829.

Scull, P., Franklin, J., Chadwick, O. a., and McArthur, D. (2003). Predictive soil mapping: a review. *Progress in Physical Geography*, 27(2):171–197.

Serbula, S. M., Ilic, A. A., Kalinovic, J. V., Kalinovic, T. S., and Petrovic, N. B. (2014). Assessment of air pollution originating from copper smelter in bor (serbia). *Environmental earth sciences*, 71(4):1651–1661.

Serbula, S. M., Kalinovic, T. S., Kalinovic, J. V., and Ilic, A. A. (2013). Exceedance of air quality standards resulting from pyro-metallurgical production of copper: a case study, Bor (Eastern Serbia). *Environmental earth sciences*, 68(7):1989–1998.

Stockmann, U., Minasny, B., and McBratney, A. B. (2011). *Quantifying Processes of Pedogenesis*, volume 113.

Tavares, M. T., Sousa, a. J., and Abreu, M. M. (2008). Ordinary kriging and indicator kriging in the cartography of trace elements contamination in Sao Domingos mining site (Alentejo, Portugal). *Journal of Geochemical Exploration*, 98(1-2):43–56.

Team, R. C. (2013). R: A language and environment for statistical computing.

Tibshirani, R. (1996). Regression Selection and Shrinkage via the Lasso.

Unit, S. C. (2013). Science for Environment Policy In-depth Report: Soil Contamination: Impacts on Human Health. Technical report, University of the West of England.

Webster, R. and Oliver, M. A. (2007). *Geostatistics for environmental scientists*. John Wiley & Sons.

Welling, S. H. (2016). *forestFloor: Visualizes Random Forests with Feature Contributions*. R package version 1.9.1.

Winstral, A., Elder, K., and Davis, R. E. (2002). Spatial snow modeling of wind-redistributed snow using terrain-based parameters. *Journal of Hydrometeorology*, 3(5):524–538.

Winstral, A. and Marks, D. (2002). Simulating wind fields and snow redistribution using terrain-based parameters to model snow accumulation and melt over a semi-arid mountain catchment. *Hydrological Processes*, 16(18):3585–3603.

Yaalon, D. H. (1975). Conceptual models in pedogenesis: Can soil-forming functions be solved? *Geoderma*, 14(3):189–205.

Yokoyama, R., Shirasawa, M., and Pike, R. J. (2002). Visualizing topography by openness: a new application of image processing to digital elevation models. *Photogrammetric engineering and remote sensing*, 68(3):257–266.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc. B*, 68(1):49–67.

Žibret, G. and Šajn, R. (2008). Modelling of atmospheric dispersion of heavy metals in the Celje area, Slovenia. *Journal of Geochemical Exploration*, 97(1):29–41.

Biography

Milutin Pejović was born in Vrbas, on March 30th, 1983. He completed elementary school (in 1998) and gymnasium for Natural sciences and mathematics in 2002 in Vrbas. In 2002, he started study of Geodesy at University of Belgrade, Faculty of Civil Engineering, Department of Geodesy and Geoinformatics. In 2009, he finished study with average mark 8.57 (max 10.00).

In 2009, he enrolled in PhD study, within the same University. During the study he completed exams and started with research related to geostatistical modeling of soil variables in 3D. In August 2016, he submitted PhD thesis entitled "Geostatistical modeling of geochemical variables in 3D". During his PhD study Milutin Pejović published as author or co-author papers: 2 journal paper (from SCI list), 4 in Serbian journals and 15 international conference papers.

Since June 2010, he has been a teaching assistant in the field of Engineering Geodesy, Department of geodesy and geoinformatics, University of Belgrade. He was involved as researcher in 1 research project founded by Serbian Ministry of Science.

Изјава о ауторству

Име и презиме аутора: Milutin Pejović

Број индекса: 5/09

Изјављујем

да је докторска дисертација под насловом

ГЕОСТАТИСТИЧКО МОДЕЛИРАЊЕ ГЕОХЕМИЈСКИХ ПРОМЕНЉИВИХ

У 3Д ПРОСТОРУ

- резултат сопственог истраживачког рада;
- да дисертација у целини ни у деловима није била предложена за стицање друге дипломе према студијским програмима других високошколских установа;
- да су резултати коректно наведени и
- да нисам кршио/ла ауторска права и користио/ла интелектуалну својину других лица.

Потпис аутора

У Београду, ________________

____

Изјава о истоветности штампане и електронске верзије докторског рада

Име и презиме аутора: Милутин Пејовић

Број индекса: 5/09

Студијски програм: Геодезија и Геоинформатика

Наслов рада: ГЕОСТАТИСТИЧКО МОДЕЛИРАЊЕ ГЕОХЕМИЈСКИХ ПРОМЕНЉИВИХ У 3Д ПРОСТОРУ

Ментори: Проф. др Бранислав Бајат дипл. геод. инж. и В. Проф. др Загорка Госпавић дипл. геод. инж.

Изјављујем да је штампана верзија мог докторског рада истоветна електронској верзији коју сам предао/ла ради похрањена у **Дигиталном репозиторијуму Универзитета у Београду.**

Дозвољавам да се објаве моји лични подаци везани за добијање академског назива доктора наука, као што су име и презиме, година и место рођења и датум одбране рада.

Ови лични подаци могу се објавити на мрежним страницама дигиталне библиотеке, у електронском каталогу и у публикацијама Универзитета у Београду.

Потпис аутора

У Београду, ________________________

____

Изјава о коришћењу

Овлашћујем Универзитетску библиотеку „Светозар Марковић“ да у Дигитални репозиторијум Универзитета у Београду унесе моју докторску дисертацију под насловом:

ГЕОСТАТИСТИЧКО МОДЕЛИРАЊЕ ГЕОХЕМИЈСКИХ ПРОМЕНЉИВИХ

У 3Д ПРОСТОРУ

____

____

која је моје ауторско дело.

Дисертацију са свим прилозима предао/ла сам у електронском формату погодном за трајно архивирање.

Моју докторску дисертацију похрањену у Дигиталном репозиторијуму Универзитета у Београду и доступну у отвореном приступу могу да користе сви који поштују одредбе садржане у одабраном типу лиценце Креативне заједнице (Creative Commons) за коју сам се одлучио/ла.

1. Ауторство (CC BY)

2. Ауторство – некомерцијално (CC BY-NC)

3. Ауторство – некомерцијално – без прерада (CC BY-NC-ND)

4. Ауторство – некомерцијално – делити под истим условима (CC BY-NC-SA)

5. Ауторство – без прерада (CC BY-ND)

6. Ауторство – делити под истим условима (CC BY-SA)

(Молимо да заокружите само једну од шест понуђених лиценци.
Кратак опис лиценци је саставни део ове изјаве).

Потпис аутора

У Београду, ________________________

____

1. **Ауторство**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце, чак и у комерцијалне сврхе. Ово је најслободнија од свих лиценци.

2. **Ауторство – некомерцијално**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела.

3. **Ауторство – некомерцијално – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца не дозвољава комерцијалну употребу дела. У односу на све остале лиценце, овом лиценцом се ограничава највећи обим права коришћења дела.

4. **Ауторство – некомерцијално – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца не дозвољава комерцијалну употребу дела и прерада.

5. **Ауторство – без прерада**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, без промена, преобликовања или употребе дела у свом делу, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце. Ова лиценца дозвољава комерцијалну употребу дела.

6. **Ауторство – делити под истим условима**. Дозвољавате умножавање, дистрибуцију и јавно саопштавање дела, и прераде, ако се наведе име аутора на начин одређен од стране аутора или даваоца лиценце и ако се прерада дистрибуира под истом или сличном лиценцом. Ова лиценца дозвољава комерцијалну употребу дела и прерада. Слична је софтверским лиценцама, односно лиценцама отвореног кода.