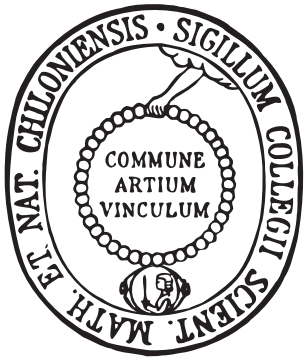
Digital soil mapping at different spatial scales using machine learning algorithms and multivariate geostatistics in a Mediterranean basin



Dissertation

zur Erlangung des Doktorgrades
der Mathematisch-Naturwissenschaftlichen Fakultät
der Christian-Albrechts-Universität zu Kiel

vorgelegt von

Michael Blaschek

Kiel, 2015

Erster Gutachter: Prof. Dr. Rainer Duttmann

Zweiter Gutachter: Prof. Dr. Ralf Ludwig

Tag der mündlichen Prüfung: 23. Oktober 2015

Zum Druck genehmigt: 23. Oktober 2015

gez. Prof. Dr. Wolfgang J. Duschl, Dekan

Abstract

Spatially distributed soil information is essential for environmental management and research. In order to obtain continuous soil data from limited point measurements, digital soil mapping (DSM) techniques are recognised as effective interpolation tools for a wide range of spatial scales and different types of landscape. This thesis aims at creating DSM approaches to model soil textural fractions at the field and catchment scale in Mediterranean soil types. It intends to improve the prediction accuracy of soil spatial interpolation models by focusing on (i) the combination of traditional kriging methods with more complex, data-driven machine learning algorithms, and (ii) the integration of environmental covariates from geophysical sensor data. In addition, model outputs are delivered by a web-based dissemination platform (geoportal) as well as in form of a Web Map Service (WMS) to advance the long-term visibility and access to DSM data.

These enhanced DSM approaches are used to predict clay, silt and sand content of the topsoil at test sites in southern Sardinia (Italy). With regard to the catchment scale, results indicate that the proposed neural network residual cokriging model outperforms common DSM techniques such as ordinary kriging or kriging with external drift. Thus, the results of this thesis suggest that machine learning algorithms such as artificial neural networks are an efficient tool for multivariate, non-linear trend analysis in soil spatial interpolation models. At the field scale, regression-based soil texture mapping using electro-magnetical and gamma radiometric data as covariates shows good model performance, except for case studies characterised by great soil heterogeneities, and particularly in the presence of highly calcareous parent material. By presenting a flexible digital soil mapping framework, and by producing accurate soil property maps, this dissertation sets the stage for the determination of more complex soil parameters as required, for instance, in hydrological modelling and precision agriculture.

Zusammenfassung

Räumlich differenzierte Bodeninformationen sind wesentliche Einflussfaktoren für ein nachhaltiges Umweltmanagement und integraler Bestandteil geographischer Forschung. Um diese kontinuierlichen Bodendaten aus punktuellen Messwerten abzuleiten, sind in der Vergangenheit diverse Techniken des Digital Soil Mapping (DSM) erfolgreich auf verschiedenen Raumskalen und in unterschiedlichen Landschaften angewendet worden. Das Ziel dieser Dissertation ist es, geeignete DSM-Ansätze zu entwickeln, um Korngrößenfraktionen mediterraner Böden auf der Feld- und Landschaftsskala zu modellieren. Dabei geht es primär um eine verbesserte Vorhersagegenauigkeit räumlicher Interpolationsverfahren (i) durch Kombinieren traditioneller Kriging-Methoden mit komplexeren, datengetriebenen Modellen des Maschinellen Lernens sowie (ii) durch die Berücksichtigung geophysikalischer Prädiktoren. Darüber hinaus werden die erstellten Karten über ein Geoportal und in Form eines WMS-Dienstes verbreitet, um eine langfristige und breite Erreichbarkeit der DSM-Daten zu gewährleisten.

Die entwickelten DSM-Ansätze werden zur räumlichen Vorhersage der Hauptfraktionen Ton, Schluff und Sand im Oberboden zweier Testgebiete im Süden Sardiniens (Italien) verwendet. Auf der Landschaftsskala zeigen die Ergebnisse, dass das erweiterte Modell ("neural network residual cokriging") anderen DSM-Techniken überlegen ist. Verfahren des Maschinellen Lernens wie die hier eingesetzten neuronalen Netzwerke eignen sich daher besonders zur Trendanalyse in räumlichen Interpolationsmethoden. Auf der Feldskala führt die Anwendung regressions-basierter Modelle unter Einbezug geophysikalischer Kovariablen zu guten Vorhersageergebnissen. Dies gilt jedoch nur eingeschränkt für untersuchte Gebiete mit besonders heterogenen Bodenverhältnissen und für Standorte mit stark kalkhaltigem Ausgangssubstrat. Durch das Bereitstellen eines flexiblen DSM-Frameworks und angesichts der Produktion präziser Karten, liefert diese Dissertation die Grundlage für die Berechnung komplexerer Bodenparameter, die beispielsweise in der hydrologischen Modellierung oder in der Präzisionslandwirtschaft benötigt werden.

Contents

List of Figures

List of Tables

Nomenclature

Institutions and projects

AGRIS	Agricultural Research Agency of Sardinia
CARG	Geological CARtography
CAU	Christian-Albrechts-Universität zu Kiel
CLIMB	Climate Induced Changes on the Hydrology of Mediterranean Basins
EC	European Commission
EU-FP7	European Union's Seventh Framework Programme for Research
EuDASM	European Digital Archive on Soil Maps of the World
FAO	Food and Agriculture Organization of the United Nations
GDI-DE	Spatial Data Infrastructure Germany
IAEA	International Atomic Energy Agency
INSPIRE	Infrastructure for Spatial Information in the European Community
iSOIL	Interactions between soil related sciences – Linking geophysics, soil science and digital soil mapping

ISPRA	Italy's Institute for Environmental Protection and Research
ISRIC	International Soil Reference and Information Centre
IUSS	The International Union of Soil Sciences
IYS	International Year of Soils
JRC	Joint Research Centre of the European Commission
LMU	Ludwig-Maximilians-Universität München
OGC	Open Geospatial Consortium
UFZ	Helmholtz Centre for Environmental Research
USDA	United States Department of Agriculture
WRB	World Reference Base for Soils

Software and licenses

ArcGIS	Esri's GIS Mapping Platform
CC BY-SA	Creative Commons Attribution Share-Alike License
FOSS	Free and Open-Source Software
GeoNode	Open-source geospatial content management system for deploying spatial data infrastructures
GME	Geospatial Modelling Environment
IODL	Italian Open Data License
R	Software environment for statistical computing and graphics

SAGA	System for Automated Geoscientific Analyses

GIS and digital soil mapping

agl	additive generalised logistic (transform)
alr	additive log-ratio (transform)
BLUP	Best Linear Unbiased Prediction
clorpt	State equation for soil formation after Jenny (1941): climate (cl), organisms (o), relief (r), parent material (p), time (t)
CSW	Catalog Service for the Web
DBSS	Database Soil Sardinia
DEM	Digital Elevation Model
DSM	Digital Soil Mapping
DTA	Digital Terrain Analysis
EDA	Exploratory Data Analysis
EF	Model Efficiency after Nash and Sutcliffe (1970)
EMI	Electro-Magnetic Induction
EPSG	European Petroleum Survey Group
FAMD	Factor Analysis for Mixed Data
GIS	Geographical Information System
GLS	Generalised Least Squares

GPS	Global Positioning System
IDW	Inverse Distance Weighting
KED	Kriging with External Drift
KS	Kennard-Stone (algorithm)
LiDAR	Light Detection And Ranging
LMC	Linear Model of Coregionalisation
LOOCV	Leave-One-Out Cross-Validation
MLP	Multi-Layer Perceptron
MLR	Multiple Linear Regression
MPS	Multiple-Point Simulations
NN	Artificial Neural Network
NNRCK	Neural Network Residual CoKriging
OK	Ordinary Kriging
OLS	Ordinary Least Squares
PCA	Principal Component Analysis
RCOK	Regression COKriging
RK	Regression Kriging
RMSE	Root Mean Squared Error
Rprop	Resilient back-propagation (learning rule)

SAR	Synthetic Aperture Radar
scorpan	Spatial prediction function introduced by McBratney et al. (2003): other soil-related factors (s), climate (c), organisms including human activities (o), relief (r), parent material (p), the time factor (a), spatial position (n)
SDI	Spatial Data Infrastructure
SLD	Styled Layer Descriptor
STRESS	STandardized REsidual Sum-of-Squares
TIFF	Tagged Image File Format
UAV	Unmanned Aerial Vehicles
UK	Universal Kriging
UTM	Universal Transverse Mercator
WCS	Web Coverage Service
WGS84	World Geodetic System 1984
WMS	Web Map Service

Chapter 1

Introduction

1.1 Motivation

Spatially distributed soil data of adequate quality is essential for environmental management and research. This is particularly true in the light of changing climate conditions and rising demands for agricultural space due to an increasing global population. In their recent Science paper, Amundson et al. (2015) relate soil to food security emphasising the great relevance that soils will have for human prosperity and survival in the future. With similar motivation the United Nations declared 2015 as the International Year of Soils (IYS). Under the patronage of the IYS, numerous projects at national and international level are being launched to raise awareness about soils as major contributors to ecosystem services like food production or carbon sequestration. Besides educational objectives, IYS-claims also stress the need for massive soil data collection and monitoring programmes ranging from local to global scale. These data requirements apply in particular to semi-arid areas such as the Mediterranean region, where soil degradation and water resource shortages are already major concerns for sustainable land management (see Vacca 2012; La Jeunesse et al. 2015). This is aggravated by the fact that Mediterranean ecosystems will be largely affected through changing climate conditions and impact of global warming (Giorgi and Lionello 2008; Navarra and

Tubiana 2013; IPCC 2014). However, to establish appropriate regional soil and water protection strategies, governing factors that, for example, control both vertical and lateral movement of water need to be known. This includes appropriate quantification of the current state of the soil and requires high-resolution mapping of the spatial distribution of key soil attributes such as texture. The great importance of soil texture results from its strong influence on many physical and chemical characteristics of the soil including conductivity, nutrient and water retention capacity, organic carbon dynamics and mechanical properties (see Greve et al. 2012; Akpa et al. 2014). In addition, soil textural fractions are essential input factors for a great variety of environmental models and risk assessment tools (Ließ et al. 2012).

The increasing demand for high-quality soil data coincides with advances in computing and rapid technological developments providing geographically continuous auxiliary information on soil forming factors. Prominent sources of such readily available secondary data are proximal and remote sensing images as well as digital elevation models. In combination with observations from conventional soil surveys, these ancillary variables provide a promising basis for modelling the spatial distribution of functional soil properties. Modern techniques to produce soil maps considering both, soil measurements and secondary data, are generally associated with the term "digital soil mapping" (DSM). The spectra of DSM methods is dominated by standard geostatistical tools like the widespread kriging techniques. Moreover, they include novel approaches based on machine learning concepts which are usually data-driven, non-spatial and highly flexible.

Regardless of the methods used, plenty of DSM projects have been carried out successfully on different scales and in different types of landscape. On a global scale, important ongoing quantitative soil inventories are the GlobalSoilMap initiative (Arrouays et al. 2014) as well as the SoilGrids1km product of the ISRIC – World Soil Information foundation (Hengl et al. 2014). At continental level, spatial prediction of soil properties has been strongly focused on Africa for the past few years, leading to continuous 250 m resolution maps compiled by the AfSIS project (Hengl et al. 2015). With respect to the Mediterranean region, for instance, Vaysse and Lagacherie (2015) applied and compared several DSM

techniques to create soil attribute maps from legacy data. Vacca et al. (2014) described an intended geomatic approach for soil mapping across the isle of Sardinia (Italy) highlighting the need for modern soil inventories to support sustainable land use planning in a highly vulnerable region.

This thesis focuses on digital mapping of soil textural fractions at the field and landscape (or catchment) scales. It is closely linked to the international research project CLIMB (Climate Induced Changes on the Hydrology of Mediterranean Basins) financed by the Seventh Framework Programme of the European Union and coordinated at the University of Munich (Germany). The CLIMB collaboration ran for slightly more than four years until February 2014 investigating the impact of climate change on catchment hydrology at seven Mediterranean test sites (Ludwig et al. 2010). The Rio di Costara catchment, as part of the Rio di San Sperate basin in southern Sardinia, is one of the investigated areas of CLIMB and represents the regional focus of this research. It represents a typical Mediterranean basin experiencing strong climatic variability and intensive agricultural activity. Moreover, the selected catchment is characterised by scarcity of soil data. In response, this work provides mapped soil attributes to serve as input data in hydrological modelling projects within that particular CLIMB test site. In addition to the provision of input data for hydrological modelling at the catchment scale, a second set of geographically continuous soil maps is produced for two agricultural fields located at the Azienda San Michele, an agronomic research farm managed by the Agricultural Research Agency of Sardinia (AGRIS). Field-related texture mapping is part of a cross-cooperation of the CLIMB consortium and the iSOIL (Interactions between soil related sciences – Linking geophysics, soil science and digital soil mapping, Werban et al. 2010) project.

1.2 Objectives and outline of the thesis

The overall objective of this work is to present a comprehensive approach for digital soil mapping at multiple spatial scales in an assigned Mediterranean study area. Producing accurate soil property maps is related to a variety of relevant

methodological research topics including soil sampling and laboratory analysis, co-variable selection, (geo)statistical modelling as well as web-based data delivery. These requirements result in a number of specific objectives, which are:

- to provide geographically continuous spatial information on soil textural fractions at two different spatial scales considering limited point measurements and ancillary co-variables,
- to assess the efficiency of using log-ratio transformations prior to spatial distribution modelling of clay, silt and sand content, and thus taking into account their compositional (sum-to-100) character,
- to test and evaluate the ability of primarily non-spatial artificial neural networks for improving the predictive power of digital soil mapping frameworks in comparison with frequently used traditional (kriging) techniques,
- to determine whether novel sets of covariates from geophysical data fusion support regression-based soil spatial interpolation approaches, and
- to advance the long-term visibility and access to soil mapping products by deploying a web-based dissemination platform using free and open-source software and following common international standards.

Following this introduction, chapter 2 describes the concepts of digital soil mapping and reviews the state-of-the-art in this particular field of research.

Chapter 3 introduces the study area of this work focusing on details of climatic, pedological and geological conditions within the Rio di Costara catchment in southern Sardinia. In addition to literature reviews, this study site characterisation is partly based on own soil profile analysis and recorded meteorological data such as temperature, precipitation and wind speed.

Chapter 4 presents the datasets and illustrates the methodology used in this thesis to achieve the objectives listed above. Following a general overview of the substantial steps in (geo)statistical data processing, chapter 4 explains the sampling strategies and laboratory work related to soil data collection. It subsequently derives the covariates serving as possible predictors in the digital soil

mapping approaches applied to this work. Special attention is also given to geophysical measurements. Afterwards, exploratory (spatial) data analysis steps are described including correlation and factor analysis as well as global trend detection and exploratory variography. The data preparation section then focuses on the theory of log-ratio transforms to account for the compositional character of the target variables and data splitting techniques used for proper model selection during neural network training. The latter is of particular interest in the frame of this thesis and thoroughly described in the section on spatial interpolation. Additionally, the concepts behind multiple linear regression and inverse distance weighting are introduced, followed by a detailed derivation of the classical approach to multi-variate geostatistics in matrix notation. Next, the promoted hybrid spatial prediction methods at the field (regression cokriging) and the landscape (neural network residual cokriging) scale are laid out. In the last section of chapter 4 the validation procedures used to test and compare model performances are presented, as well as the technical implementation details regarding the web-based delivery and documentation of soil mapping products.

Chapter 5 illustrates the results of spatial prediction. It distinguishes between approaches applied at the field scale (fields 21 and 33 located at the San Michele farm) and the landscape scale (the entire Rio di Costara catchment).

Chapter 6 discusses the results in coherence with the formulated objectives and places them into scientific context. Moreover, it focuses on the applicability of geophysical measurements as covariates in field-level digital soil mapping and evaluates the potential of combined neural network and cokriging approaches for spatial interpolation.

The thesis concludes with an overview of the main achievements and identifies possible ways forward in chapter 7.

Appendix A lists important software and applications as well as commented screenshots concerning the final geoportal solution disseminating the produced soil property maps. Appendix B shows R-packages and -scripts used in the frame of this thesis. Appendix C provides all recorded data regarding clay, silt and sand content of the top 30 cm soil layer at both the field and the landscape scale.

Chapter 2

Current state of research

Since the mid-1990s, spatial prediction or interpolation of soil properties is closely linked to the research fields of pedometrics and digital soil mapping (DSM). The scientific discipline of pedometrics officially aims at "the application of mathematical and statistical methods for the study of the distribution and genesis of soils" (see Heuvelink 2003, p. 11). It operates as an interdisciplinary subject combining elements from soil science, applied statistics and mathematics as well as geoinformation science (Hengl 2003). By contrast, digital soil mapping can be seen as a particular application within the branch of pedometrics, focusing on the development of methods for an automated production of soil (property) maps (Lagacherie and McBratney 2006).

Both, pedometrics and digital soil mapping, evolved rapidly during the past decades and occupied important positions in the organisational network of the International Union of Soil Sciences (IUSS). In 2002 pedometrics became a commission of the IUSS (1.5) which subsequently established, along with the commission on Soil Geography (1.2), an "International Working Group on Digital Soil Mapping". This DSM task force started its work by organising the first of a series of biennial meetings, which was held in Montpellier (France) in 2004. The proceedings of that meeting and the ones associated with its follow-up workshops provide an excellent introduction into the topic of DSM (Lagacherie et al. 2006; Hartemink et al. 2008; Boettinger et al. 2010; Minasny et al. 2012). Apart from

those collections, one of the most important and comprehensive formalisation of digital soil mapping is given by McBratney et al. (2003). Gessler et al. (1995) and Scull et al. (2003) describe the concepts of soil-landscape modelling and predictive soil mapping, respectively, which are very similar to DSM but specified in a slightly different manner. For details about the concepts and early developments in pedometrics refer to Webster (1994) and Burrough et al. (1994). Lark (2005) provides a more recent review of pedometrics. A detailed presentation of specific DSM (or pedometric mapping) techniques can be found in Grunwald (2006), Minasny et al. (2008) and Grunwald (2009). In an influential paper McBratney et al. (2000) classified the techniques of digital soil mapping into (i) geostatistics, (ii) the so-called clorpt or scorpan models and (iii) hybrid methods that combine central elements of the first two groups.

The field of geostatistics as used in present-day practice has its origin in the mining industry and was most importantly defined by G. Matheron in the early 1960s (see Journel and Huijbregts 1978). Its key concept is based on stochastic considerations which Matheron (1971) developed and summarised under the theory of regionalised variables. In honour of earlier work by D. G. Krige, the procedure of geostatistical prediction is commonly referred to as "kriging" (see Cressie 1990). Standard (ordinary) kriging basically operates on auto-correlation evaluated from spatially distributed measurements of some target variable. It is therefore an interpolation technique that strongly depends on the number and arrangement of available observations (Hengl 2009). Starting with ore reserve estimation (Journel and Huijbregts 1978), numerous uni- and multivariate variants of the krige interpolator have been applied to date within different scientific fields including hydrogeology (Kitanidis 1997), epidemiology (Lawson 2013) and ecology (Dale and Fortin 2014).

Focusing on soil science, Webster (2000) provided a noteworthy framework for geostatistical modelling that considers the specific characteristics of spatial soil variation. More recent developments in soil-related geostatistics are described, for instance, by Lark (2012b), who discusses the advantages of linear mixed model theory as an alternative framework for geostatistical prediction according to the concepts developed by Stein (1999) and Diggle and Ribeiro Jr. (2007).

In the presence of compositional data such as soil separates, (geo)statistical modelling requires some data transformation prior to interpolation (Aitchison 1982; Pawlowsky et al. 1995). Odeh et al. (2003), Lark and Bishop (2007) and Buchanan et al. (2012) provide successful examples of performing such additive log-ratio transformations prior to digital mapping of soil particle-size fractions. Alternative ways of dealing with constant-sum constraints in kriging applications were suggested by De Gruijter et al. (1997) and Walvoort and de Gruijter (2001), who modified the kriging equations rather than transforming the target quantities.

A major assumption of the standard kriging model is mean and (co)variance stationarity, which is hardly realistic in soil-scientific applications due to usually large spatial heterogeneities (Webster and Oliver 2007). In response to a varying spatial mean, also referred to as trend or drift, ancillary data is often used to separately approximate any possible trend component. By contrast to the stochastic kriging system, these clorpt (Jenny 1941) or scorpan (McBratney et al. 2003) models rely on known deterministic relationships between environmental variables and the soil property of interest:

$$S_a = f(s, c, o, r, p, a, n) + \epsilon \tag{2.1}$$

where any soil attribute S_a is a function of other soil-related factors (s), climate (c), organisms including human activities (o), relief (r), parent material (p), the time factor (a), spatial position (n), and ϵ being the spatially correlated errors (Minasny et al. 2013; following McBratney et al. 2003). Note that the scorpan function for spatial prediction of soil properties is strongly linked to the concept of environmental correlation introduced by McKenzie and Ryan (1999). Starting from simple regression (Moore et al. 1993; Gessler et al. 1995), numerous scorpan models have been applied differing in the method used to formalise the relationship between measured soil quantities and available predictors. These methods are nowadays often based on machine learning algorithms (Brungard et al. 2015). The term machine learning refers to a broad variety of models meant for pattern analysis in data, also known as data mining, and making data-driven predictions (Witten and Frank 2005). These learning algorithms are extremely powerful with

respect to soil properties, because often little is known about the complex relationship between the studied soil attributes and available scorpan factors. Recent applications of machine learning algorithms in soil science included the use of generalized additive models (Poggio et al. 2013), support vector machines (Kovačević et al. 2010), Bayes networks (Taalab et al. 2015), Cubist (Lacoste et al. 2014), classification trees (Scull et al. 2005; Kim et al. 2012), random forests (Grimm et al. 2008; Wiesmeier et al. 2011; Barthold et al. 2013) and neural networks (Behrens et al. 2005; Malone et al. 2009; Ozturk et al. 2011; Guo et al. 2013). The statistical theory related to many of these techniques is comprehensively outlined in the textbooks by Hastie et al. (2009) and Abu-Mostafa et al. (2012).

Focusing on soil textural fractions, Greve et al. (2012) applied regression trees on a national scale (Denmark) to map clay, silt and sand contents of the top 30 cm soil layer using different environmental covariates including continuous terrain attributes and categorical information on parent material. Akpa et al. (2014) used legacy soil profile data in combination with various scorpan factors to predict the spatial distribution of soil particle-size fractions across Nigeria using a random forest model. Both, regression trees and random forest models were applied and compared by Ließ et al. (2012). They focused on terrain attributes as covariates for soil texture mapping in a meso-scale catchment of the southern Ecuadorian Andes. Digital mapping of soil textural fractions based on neural networks has been done, for instance, by Zhao et al. (2009) and Priori et al. (2014). Zhao et al. (2009) predicted spatial distribution of soil texture from hydrographic parameters obtained from a digital elevation model in a Canadian catchment. By contrast, Priori et al. (2014) focused on agricultural fields comparing the performance of support vector machines and neural networks on the basis of gamma radiometric data to model the variation of clay and sand content in seven vineyards of a Tuscany farm (Italy). A major advantage of all outlined machine learning algorithms is their flexibility to adequately address the usually non-linear and complex dependencies between soil properties and environmental covariates. However, these kinds of prediction methods are non-spatial and thus ignore any possible similarities due to geographical proximity. In addition, they show weak performances in case studies where only little correlations occur between the target quantities

and available environmental predictors. Moreover, common machine learning algorithms do not consider any random variation. As a consequence, and in order to combine the benefits from geostatistics with the advantages of scorpan-based modelling, McBratney et al. (2000) promoted another group of spatial interpolation techniques often referred to as hybrid methods.

Hybrid methods are spatial interpolation techniques that consider both deterministic and stochastic elements. Important examples of combined models are universal kriging (Matheron 1969), regression kriging (Odeh et al. 1995; Hengl et al. 2004; Hengl et al. 2007) and kriging with external drift (Goovaerts 1997). Although being very similar, these spatial prediction tools exhibit some subtle differences which are broadly discussed by Hengl et al. (2004). Regression kriging (RK) is a univariate digital soil mapping technique coined by Odeh et al. (1995) that additively combines the regression of some target quantity on a set of scorpan factors with (ordinary) kriging of the regression residuals. Examples of applications of this approach in soil sciences are provided by Sumfleth and Duttmann (2008), Zhang et al. (2012) and Piccini et al. (2014). Gobin et al. (2001) and Hengl et al. (2004) substituted original predictors by principal components (PCs) resulting in better prediction performances of the RK model. Instead of using linear models, McBratney et al. (2000) and Minasny et al. (2008) expanded the list of hybrid interpolation techniques by more flexible and particularly non-linear methods for scorpan-based trend estimation such as regression trees or neural networks. The latter was coupled with a multivariate cokriging approach by Kanevski et al. (1997) in order to map soil contamination due to Chernobyl fallouts. They later substituted the kriging component of their spatial prediction model by sequential Gaussian simulations providing a measure of uncertainty related to radioactive soil contamination mapping (Kanevski et al. 2004). A recent example of combining neural networks with residual estimation using ordinary kriging is given by Dai et al. (2014) who mapped soil organic matter contents in the Tibetan highlands.

Spatial interpolation techniques based on explanatory factors such as the above mentioned scorpan or hybrid models require geographically continuous maps of environmental covariates. These maps have been extensively derived from digital

elevation models or scanned copies of geological representations. However, during the past decades other sources have become relevant, focusing primarily on non- or minimal-invasive observation methods. Mulder et al. (2011) provide a very recent review of applications using remote sensing images for soil mapping projects at regional and coarser scales. Case studies that focus on the field scale more often rely on proximal sensing techniques such as electromagnetic induction (Triantafilis and Lesch 2005; Cockx et al. 2009) and gamma-ray spectrometry (Dierke and Werban 2013; Priori et al. 2014). Very few soil mapping projects combine the different kinds of sensor data. For instance, Castrignanò et al. (2012) successfully used PCs of raw covariates from GPS heights, electromagnetic induction and gamma-ray spectrometry for the delineation of homogeneous soil zones. Piikki et al. (2013) presented an application for mapping topsoil clay content on the basis of ancillary data from electromagnetic induction measurements and gamma-ray spectrometry at an agricultural field in Sweden. Contrasting soils were studied by Rodrigues et al. (2015) who made spatial predictions of clay, silt and sand contents based on stepwise regression procedures and sensor data fusion using principal component analysis.

Chapter 3

Study site characterisation

The Rio di Costara test site is located in the province of Cagliari in southern Sardinia (Italy). It is a river catchment of 16.44 km^2 size, divided between the municipalities of Ussana (W), Serdiana (E), Donori (N) and Dolianova (S) according to the GADM database of Global Administrative Areas (Hijmans et al. 2012). The Rio di Costara catchment ranges in elevation from 85 to 271 metres with an average of 148 metres above sea level (m. a. s. l.). It has a gently undulating topography and is part of the border area between the eastern edge of the Campidano plain and the western foothills of the mountainous Sarrabus-Gerrei sub-region.

The study area includes main parts of the Azienda San Michele, an agronomic research farm managed by the Agricultural Research Agency of Sardinia (AGRIS). The San Michele farm enabled and greatly supported the extended field campaigns related to this thesis. Figure 3.1c on page 14 shows the exact position of the farm in respect to the Rio di Costara catchment boundaries. It covers an area of 4.36 km^2 with central coordinates of 9°6' E, 39°25' N. Two adjacent fields in the area of the Azienda San Michele, marked in figure 3.6 on page 22, are of particular interest to this work. Both, field 21 (4.86 ha) and field 33 (10.74 ha) were investigated within various past research projects such as iSOIL (Werban et al. 2010) and CLIMB (Ludwig et al. 2010). To ensure comparability with the previous surveys, former field labels were maintained.

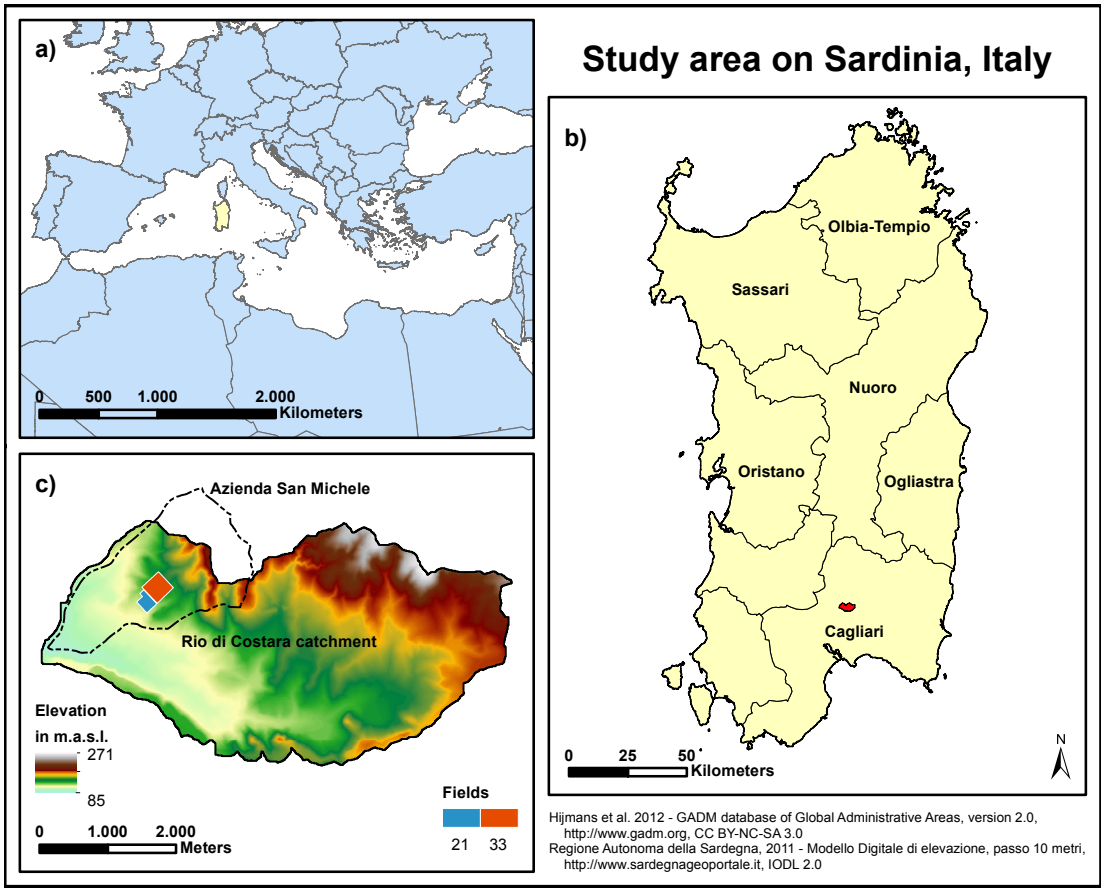
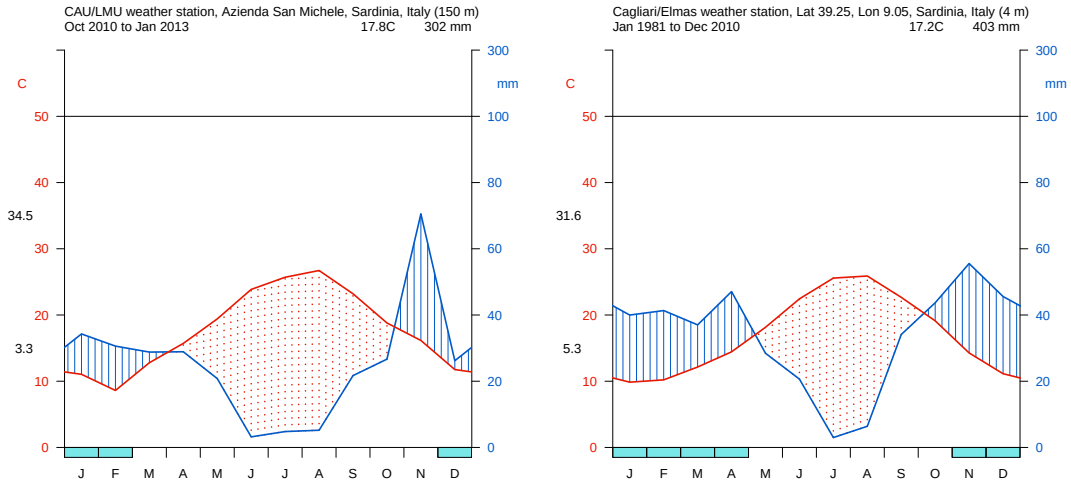


Figure 3.1: Study area in southern Sardinia (Italy) – a) Sardinia in its Mediterranean context, b) Provinces of Sardinia, c) Digital elevation model of the Rio di Costara catchment

3.1 Climate conditions

The climate of Sardinia is typically Mediterranean with dry summer months and cool but rainy winter days. It can be classified as Csa according to the modified Köppen system (Chessa and Delitala 1997; following Critchfield 1983). Benzi et al. (1997) recognised April and September as transition months between the two dominant seasons with regard to temperature (Chessa and Delitala 1997). Mean monthly temperatures within the region of interest in southern Sardinia range from 9 °C in January to 25 °C in July and August. Average annual precipitation is 680 mm (Mascaro et al. 2013, p. 4145). By contrast, Vacca et al. (2002) defines

a slightly lower precipitation value of below 500 mm for the southern Sardinian region and an average annual rainfall of 700–900 mm for the inner, hilly areas of the island. Regardless of differing average values, several authors consistently emphasise that overall occurrence of rainfall is almost limited to the cold season (e.g. Delitala et al. 2000; Mascaro et al. 2013). Furthermore, annual precipitation varies significantly between years (Delitala et al. 2000; Vacca et al. 2002). The prevailing wind on Sardinia is the Maestrale reaching the island from north-west direction. It is characterised by dry, cold breezes, stable air and clear sky (Dalu and Cima 1983). The second important, but less frequent wind on Sardinia is the Scirocco. Especially in winter, this south-easterly wind gives complete cloud coverage and is responsible for continuous rainfall (Dalu and Cima 1983).



(a) San Michele farm (b) Cagliari/Elmas

Figure 3.2: Walter and Lieth diagrams for southern Sardinia (Italy). Data in (a) calculated from own measurements, while data in (b) is taken from Tutiempo Network, S.L. (2014)

Figure 3.2a illustrates the temperature and precipitation conditions during campaign periods between 2010 and 2013 in form of Walter and Lieth diagrams. The presented values were collected at a weather station located close to the main building at the San Michele farm (see figure 3.6 on page 22). It was installed in cooperation with LMU Munich and AGRIS Sardegna during an extended field

campaign in early October 2010. The weather station ran until late March 2013 to support the hydrological modelling within the CLIMB project by estimating evapotranspiration from meteorological measurements. Accordingly, the following parameters were constantly recorded from different heights in a time interval of ten minutes: relative humidity in %, air temperature in °C, global radiation in W/m^2, radiation balance in W/m^2, wind direction at a height of 5 m, wind speed in m/s, soil temperature in °C and soil moisture in %. Figure 3.2b displays the climate conditions for a thirty-year time frame from 1981 to 2010 at Cagliari/Elmas located approximately 30 km south of the San Michele farm. It again emphasises the semi-arid character of the southern Sardinian climate and strongly resembles the conditions recorded for the field campaign period by own sensors. Figure 3.3 represents the wind system at the CAU/LMU weather station during the recording period. It verifies the dominant role of the north-western Maestrale and the secondary wind direction of the Scirocco blowing from south-easterly direction (see Dalu and Cima 1983). The presented wind rose also shows wind speed depending on the given directions. It is interesting to note that light air and light breezes in 5 m height are more likely to occur from the North. Very calm winds with wind speeds below 0.1 m/s were excluded from image processing.

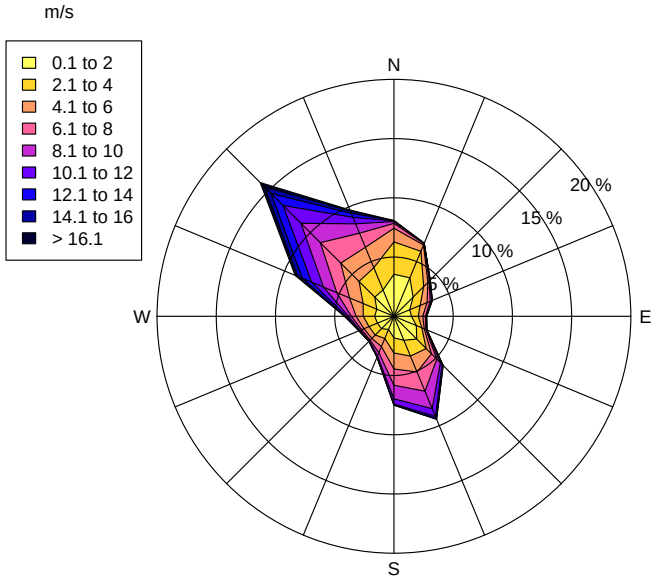


Figure 3.3: Wind direction and wind speed at the San Michele farm. Annual wind rose in a height of 5 m above ground. Circulating winds: 22.9 %.

3.2 Geological setting

The geological setting on Sardinia is rather complex and even today quite a few rock formations of different age outcrop throughout the island. The most ancient sediments are Cambro-Ordovician and were deformed under metamorphic conditions during the Variscan (or Hercynian) orogeny of the Early Carboniferous period (Boni et al. 2009; following Conti et al. 2001). In addition to the Paleozoic metamorphic basement, the Variscan collision between the northern Armorican and the southern Gondwana continents also produced the Late Hercynian magmatic complex. Predominant formations are granites in the northern Gallura and south-eastern Sarrabus-Gerrei regions of Sardinia (Carmignani et al. 1994). During the Mesozoic, long periods of marine sedimentation led to carbonate formation. Today these can be found, for instance, in the vicinity of the east-central gulf of Orosei as well as in the north-western Nurra region (Carmignani et al. 1994). After this rather calm era Sardinia (and Corsica) featured higher tectonic activities during the Cenozoic with phases of stretching, shortening, rifting and volcanism (Casula et al. 2001). In addition, Sardinia became an island at that time and drifted counterclockwise into its current position within the Burdigalian age approximately 16 to 20 million years ago (Casula et al. 2001). Deposits from two Oligo-Miocene sedimentary cycles are mainly present in south-central depressions of the Sardinian island (Vacca 2012). The most recent sediments from quaternary deposits are largely concentrated on the Pliocene Campidano graben located in the south-western region of Sardinia between the cities of Oristano and Cagliari (Casula et al. 2001).

In exception of the Mesozoic carbonates all described lithological complexes are relevant for the study area of this thesis. Figure 3.4 displays the main geological units of the Rio di Costara catchment according to a 1:25,000 map published by the geoportal of the Autonomous Region of Sardinia. There is also another image of an official 1:50,000 geological map available from the CARG project realised by ISPRA (http://www.isprambiente.gov.it). This larger scale map basically reveals the same entities as shown in figure 3.4. However, it provides an extensive description used to explain the given geological units (Funedda et al. 2013).

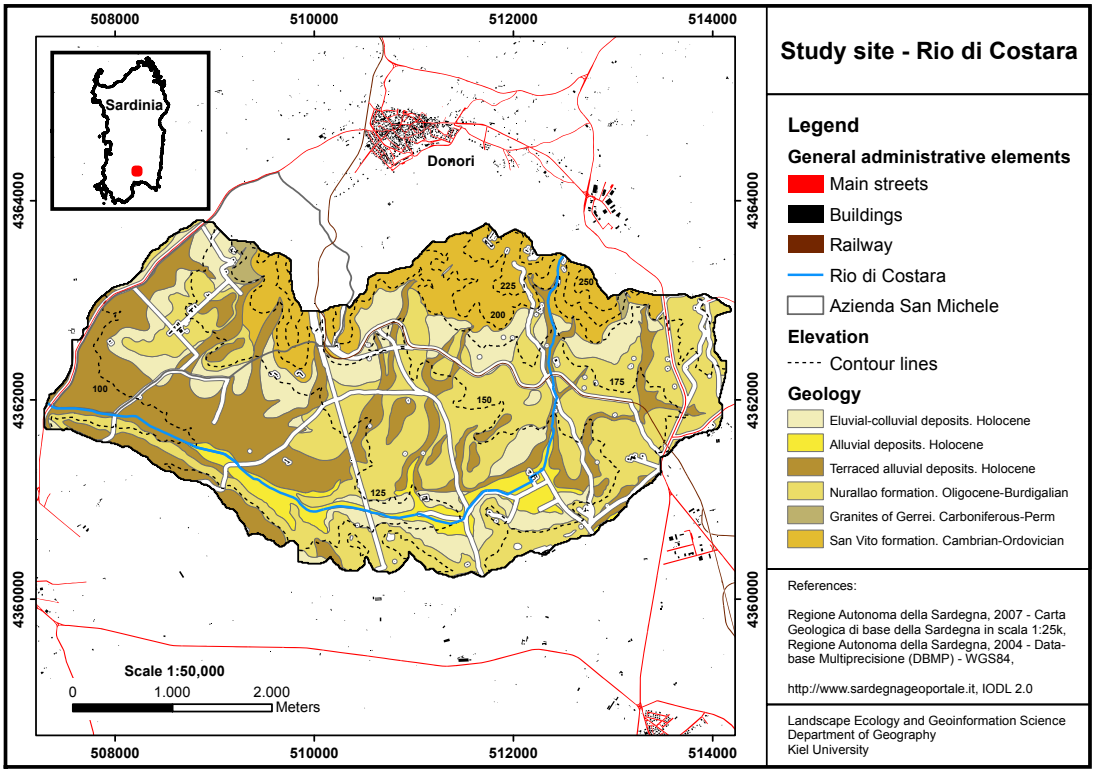


Figure 3.4: Geological overview of the Rio di Costara catchment

The more mountainous northern part of the Rio di Costara catchment largely consists of the Variscan metamorphic basement formation. To be exact, it belongs to the Sarrabus tectonic unit. Its prevalent San Vito formation is characterised by micaceous and quartz-rich metasandstones (Accornero and Marini 2007). West of this very ancient metamorphic material, a relatively small area of the test site is associated with intrusive granitoids of the Barrali unit. These Late Paleozoic rocks are mainly made of equigranular biotite monzogranites that might occur in grey to pinkish colour after alteration (Funedda et al. 2013). The Nurallao formation, dominating the central and eastern areas of the catchment, is part of the Cenozoic volcanic sedimentary cover. Deposited in shallow water during the Late Oligocene to Burdigalian, it belongs to the first sedimentary cycle of that era and is chiefly constituted by sandy-conglomerate alternations (Mancosu 2012). The second Miocene marine sedimentary cycle, represented by Gesturi Marls, is of minor interest related to the Rio di Costara catchment following

the classification of available geological maps. Nevertheless, it plays an important role in the adjacent Ussana area right at the southern border of the study site (Mancosu 2012). In addition, several soil surveys concerning the San Michele farm are claiming these Miocene marls as an important and widespread parent material. Recent quaternary deposits inside the lowlands of the study area are subdivided into three different classes. Most important in terms of surface area, especially in the western region, are pebbly terraced alluvial deposits with interbedded sand (Vacca 2012). A second group of alluvial sediments deposited in active rivers includes gravels of coarse to medium size (Funedda et al. 2013). The final present geological unit summarises deposits originating from gravitational transport and concentrating on footslope areas throughout the catchment.

Considering the geological complexity, as described up to this point, recent (southern) Sardinian landscapes are morphologically distinct. Inside the Rio di Costara catchment, a combination of weathering, erosion and intensive farming activity formed an undulating topography with small, rounded hills especially on Miocene and Holocene sedimentary successions. A slightly more rough relief developed on the metamorphic and granitic Paleozoic basement, located in the northern parts of the study area. For a visual impression on the given landscape, see figure 3.5 on page 20. The left photo was taken from the mountainous part of the San Michele farm in March 2010 in south-east direction. The right picture was taken from a more north-central position inside the Rio di Costara catchment in October 2010 and is oriented south.

3.3 Soil types

Current soil formation in the Mediterranean region is strongly influenced by the specific climate conditions described in chapter 3.1 with two well-distinguished seasons per year. Having most of the annual rainfall during wintertime followed by a dry and hot summer period of several months defines a xeric soil moisture regime according to Soil Taxonomy (Soil Survey Staff 1999) and its latest update (Soil Survey Staff 2010). In addition, Sardinia features a thermic temperature regime.

(a) San Michele farm (b) Rio di Costara catchment

Figure 3.5: Test site impressions from March/October 2010

This is confirmed by an annual soil temperature of approximately 19 °C measured in 50 cm depth at the Azienda San Michele from March 2011 to November 2012. Both regimes define the background for the most important formation processes in Mediterranean soil landscapes: the migration of clay and the redistribution of calcium carbonate during the rainy winter months, as well as the (red) oxidisation of iron compounds in the summer (Yaalon 1997). All these processes are highly dependent on the second important soil forming factor which is parent material. Based on the geological complexity of Sardinia discussed in chapter 3.2 existing parent material is rather diverse. Nevertheless, calcareous rocks are said to be predominant throughout the Mediterranean – which also holds true for the study area of this thesis – and do usually weather into a clayey texture (Verheye and de la Rosa 2006). Another important source of fine soil material affecting all terrestrial soils in the area of interest is Sahara dust (Yaalon 1997). Further relevant soil forming factors are topography and human influences. On the one hand, a calm relief allows for undisturbed soil development. This, for instance, might end in a well-differentiated Red Mediterranean Soil (Terra rossa) which is supposed to be the regional soil climax on hard limestone (Verheye and de la Rosa 2006). On the other hand, more hilly landscapes are prone to erosional processes leading to rejuvenated soils on slopes and accumulation of soil material at lower positions. In general, erosion is a severe problem in undulating landscapes that is often initiated or accompanied by human activities. Thus, a long history of crop

production, grazing and deforestation heavily influenced the recent distribution of soils on Sardinia (Aru 1985).

From the brief summary on soils in Sardinia presented by Aru (1985), regosols (Xerorthents) and cambisols (Xerochrepts) are the most widespread soils related to the study area of this thesis. (Categorical) information on the spatial variation of soils in the Rio di Costara catchment can be taken from a 1:200,000 soil map published by the Joint Research Centre of the European Commission (JRC) through the European Digital Archive on Soil Maps of the World (EuDASM). This soil map of southern Sardinia was originally created by Angelo Aru and Paolo Baldaccini in 1962 to 1964 and can be freely downloaded as PDF-document from http://eusoils.jrc.ec.europa.eu. Following the map and its inherent nomenclature, predominant soils in the western and southern parts of the catchment are reddish brown soils with carbonate accumulation. Their spatial distribution strongly corresponds to the location of quaternary deposits as indicated by the previously discussed geological map (see figure 3.4 on page 18). In contrast, metamorphic and granitic Paleozoic basement in the northern area developed into brown earths and lithosols. Small portions of the catchment especially at the eastern margin are covered by hydromorphic vertisols and the group of brown earths, regosols and vertisols on marls, sandstones and conglomerates.

For the Azienda San Michele a more detailed soil description on a 1:10,000 scale is available which was digitised in a former project by Oppo (2010) after a survey from Aru (1966). It is displayed in figure 3.6 and at a higher resolution focusing on the surveyed fields 21 and 33 in figure 4.3 on page 35. Taking everything into conclusion, both fields are dominated by soils derived from subrecent alluvial deposits which differ from each other due to erosion influences of varying intensity. Aru (1966) also observed the presence of brown soils and regosols from Miocenic marls and sandstones in significant parts of the San Michele farm.

On the following pages selected soil profiles from own field trips in March 2011 and 2013 are shown (see figure 3.6 for exact profile locations). Related field soil descriptions and laboratory analysis were done in accordance with the German soil classification system and its manual of soil mapping (AG Boden 2005). The

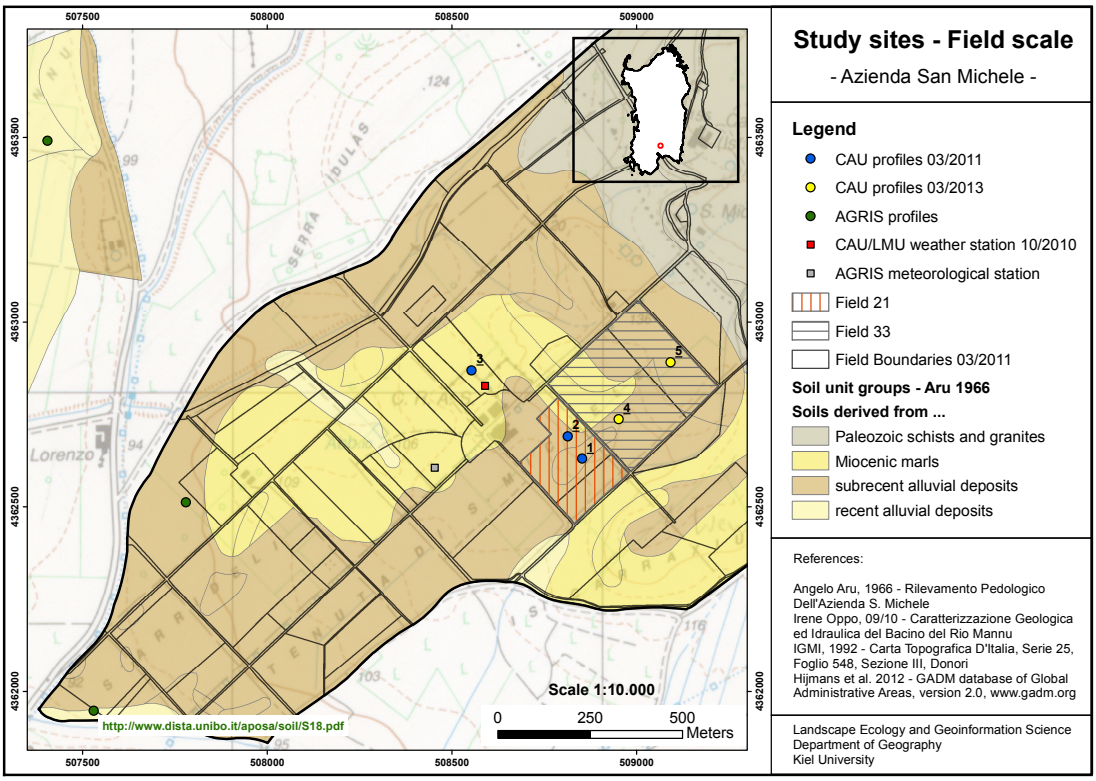


Figure 3.6: Soil map of the San Michele farm (after Aru 1966) and geographic location of the weather station, analysed soil profiles, and studied agricultural fields 21 and 33

final soil and horizon naming follows the World Reference Base for Soils (WRB) in its latest version (IUSS Working Group WRB 2006). Thus, using relatively recent guidelines, deviations from former surveys might occur since different classification and naming schemes were on hand. For instance, previously declared brown soils, surveyed in the 1960s at locations that are strongly influenced by erosional processes, might be classified as anthrosols today. Anthrosols, however, entered the FAO and later the WRB system only in the 1990s, long after the initial surveys from Aru and others. The decision upon location of soil profiles based on access constraints due to present crop arrangements. Additionally, different positions of the given morphological setting should be covered. Soil profile 3 was intentionally placed close to the weather station and represents soils developed from Miocene marls as indicated by the survey of Aru (1966).

Soil profile 1 – Calcaric Regosol

Location: Municipality of Ussana, Province of Cagliari, Sardinia, Italy
Easting: 508853, Northing: 4362630 (UTM 32 N, WGS 84)
Height: 122 m. a. s. l., Slope: 1°, Aspect: 140°, Land use: rape

Located on top of a small hill, profile 1 is strongly affected by loss of material due to erosion processes. This leads to a relatively small ploughed A horizon of about 40 cm thickness. It is followed by four Ck horizons where "k" describes the accumulation of secondary carbonates. Below the Ck4 horizon, the bedrock material consists of calcareous sandstone. The texture class of the first three horizons is clay and changes over clay loam in Ck3 to loam in the Ck4 horizon. The skeleton fraction of the soil is lower than 4 % throughout the whole profile. Bulk density is continuously low and ranges from 1.4 to 1.5 g/cm^3. Measured pH-values are consistently higher than 7 characterising the soil as slightly alkaline. The content of calcium carbonate is high varying between 21.9 and 51.8 %, while soil organic carbon never exceeds 0.6 %. The C/N-ratio of the Ap horizon is 11 rising to 139 in the Ck4 horizon.

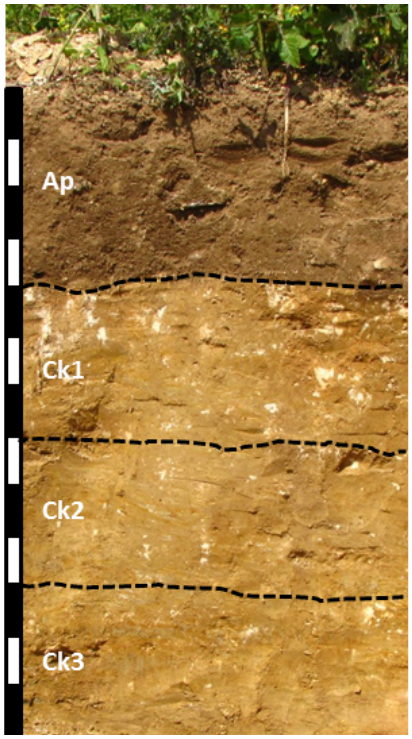


Figure 3.7: Soil profile no. 1

Table 3.1: Selected characteristics of soil profile no. 1

Horiz.	Depth	Sand	Silt	Clay	CaCO3	Carbon	BD[1]	pH
	cm			%			g/cm^3	$(CaCl_2)$
Ap	40	34.11	18.16	47.73	37.3	5.06	1.42	7.45
Ck1	70	30.89	19.62	49.49	51.8	6.76	1.47	7.56
Ck2	100	31.06	24.61	44.32	42.0	5.01	1.52	7.57
Ck3	130	34.69	30.71	34.60	21.9	3.18	1.41	7.65
Ck4	>130	49.65	32.11	18.25	33.2	3.96	-	7.68

[1] BD = Soil bulk density

Soil profile 2 – Calcaric Regosol

Location: Municipality of Ussana, Province of Cagliari, Sardinia, Italy
Easting: 508814, Northing: 4362690 (UTM 32 N, WGS 84)
Height: 121 m. a. s. l., Slope: 1°, Aspect: 250°, Land use: rape

Located only slightly lower than profile 1, soil profile 2 is also characterised by material loss due to erosional processes. The ploughed A horizon is approximately 45 cm thick and followed by three Ck horizons over calcareous sandstone.

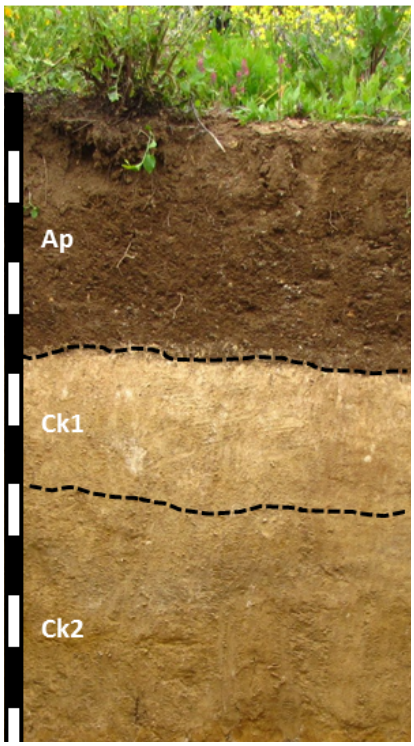


Figure 3.8: Soil profile no. 2

The texture class in the surface layer is clay and changes to clay loam in the first two Ck horizons and sandy loam in the Ck3 horizon. The soil skeleton content is relatively high in the Ap horizon with 8.3 %, but lower than 0.5 % in the subsequent horizons. Clay content significantly decreases from 46.89 % at the top to 18.44 % at the bottom of the profile. Bulk density is continuously low and ranges from 1.23 to 1.35 g/cm^3. Measured pH-values are consistently higher than 7 characterizing the soil as slightly alkaline. The content of calcium carbonate is highly variable between 15.1 and 72.6 %, while soil organic carbon never exceeds 0.7 %. The C/N-ratio of the Ap horizon is 11 and rises up to 115 in the Ck1 horizon.

Table 3.2: Selected characteristics of soil profile no. 2

Horiz.	Depth	Sand	Silt	Clay	CaCO3	Carbon	BD[1]	pH
	cm			%			g/cm^3	$(CaCl_2)$
Ap	45	30.67	22.44	46.89	22.1	3.28	1.23	7.51
Ck1	70	43.52	22.36	36.09	72.6	8.79	1.23	7.56
Ck2	120	42.15	22.55	32.46	50.4	6.53	1.35	7.62
Ck3	>120	67.80	13.76	18.44	15.1	1.91	-	7.69

[1] BD = Soil bulk density

Soil profile 3 – Regic Anthrosol

Location: Municipality of Ussana, Province of Cagliari, Sardinia, Italy
Easting: 508554, Northing: 4362869 (UTM 32 N, WGS 84)
Height: 114 m. a. s. l., Slope: 2°, Aspect: 285°, Land use: grassland

Soil profile 3 is located slightly separated from the other profiles on the same field as the weather station and is strongly influenced by colluvic material (Ap1, Ap2).

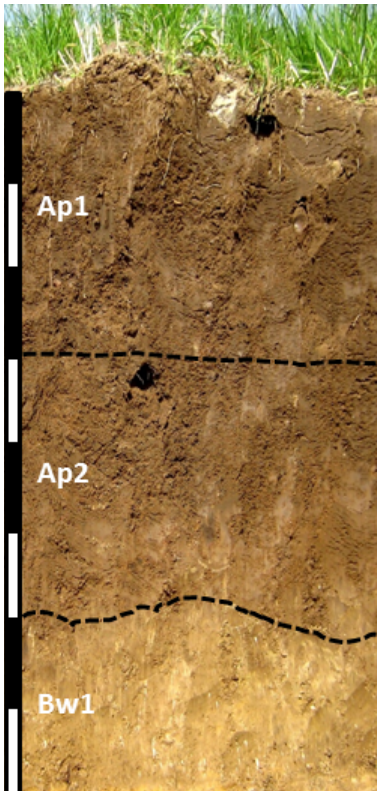


Figure 3.9: Soil profile no. 3

The two upper horizons are followed by a well-developed Bw1 horizon and a thin Bw2 horizon with high content of calcium carbonate that originates from the underlying high porous calcareous sandstone right below 85 cm. The texture class of all horizons is clay loam with a clay content between 32 and 43 %. Skeletal components are continuously below 3 %, while bulk density exhibits values around 1.5 g/cm^3. Measured pH-values are consistently higher than 7 indicating the influence of carbonate. Soil organic carbon content never exceeds 0.8 %. C/N-ratios from 8 to 9 in the upper horizons indicate good conditions for microbial activity until 80 cm depth. An overall Ap-thickness of 60 cm qualifies the whole profile as Anthrosol. The prefix regic- results from the fact that no buried horizons were identified.

Table 3.3: Selected characteristics of soil profile no. 3

Horiz.	Depth	Sand	Silt	Clay	CaCO3	Carbon	BD[1]	pH
	cm			%			g/cm^3	($CaCl_2$)
Ap1	30	43.13	24.56	32.31	1.8	0.99	1.51	7.38
Ap2	60	42.07	24.74	33.19	1.6	0.87	1.46	7.38
Bw1	80	34.80	23.06	42.14	0.5	0.54	1.56	7.35
Bw2	>80	33.99	26.08	39.93	31.5	4.23	1.53	7.50

[1] BD = Soil bulk density

Soil profile 4 – Chromic Cambisol

Location: Municipality of Ussana, Province of Cagliari, Sardinia, Italy
Easting: 508952, Northing: 4362737 (UTM 32 N, WGS 84)
Height: 127 m. a. s. l., Slope: 2°, Aspect: 210°, Land use: clover

Located inside a very small depression, profile 4 is more or less protected from erosion-induced loss of soil material and initial soil development took place. Two Ap horizons are followed by Bw1 and Bw2 horizons that show no clear migration of clay, but significant changes in colour and bulk density. The latter increases downwards along the profile from 1.4 to 1.7 g/cm^3. Clay content varies around 35 % with sand content between 40 and 46 % classifying most of the profile's texture as (sandy) clay loam. Measured pH-values are consistently higher than 7 indicating the influence of carbonate. The content of calcium carbonate starts slightly below 4 % in the surface horizon and continuously decreases with depth. Strong reddening (rubefication) of the Bw horizons from iron oxidisation is documented by changes from 10 YR to 7.5 YR 4/4 following the Munsell colour system and helps classifying the given soil profile as chromic cambisol.

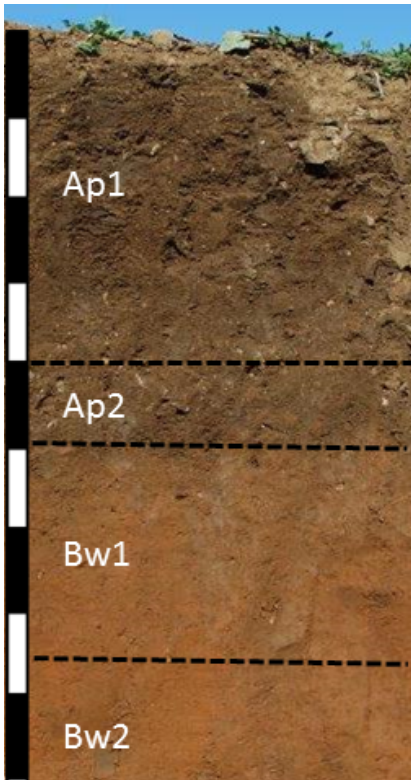


Figure 3.10: Soil profile no. 4

Table 3.4: Selected characteristics of soil profile no. 4

Horiz.	Depth	Sand	Silt	Clay	CaCO3	Carbon	BD[1]	pH
	cm			%			g/cm^3	$(CaCl_2)$
Ap1	40	40.90	24.66	34.44	3.8	1.28	1.43	7.21
Ap2	50	40.64	24.43	34.93	1.7	0.82	1.43	7.29
Bw1	75	46.23	21.21	32.56	0.29	0.25	1.71	7.22
Bw2	>75	43.39	17.86	38.75	0.21	0.20	1.67	7.15

[1] BD = Soil bulk density

Soil profile 5 – Haplic Luvisol

Location: Municipality of Ussana, Province of Cagliari, Sardinia, Italy
Easting: 509093, Northing: 4362891 (UTM 32 N, WGS 84)
Height: 135 m. a. s. l., Slope: 2°, Aspect: 180°, Land use: wheat

Located at the northern edge of field 33, profile 5 reveals a mature soil which is characterised by strong clay illuviation into three Bt horizons. The third Bt horizon also exhibits high content of calcium carbonate (56 %), whereas the rest of the profile is almost carbonate-free. The texture class in the upper horizon is sandy clay loam and changes into clay in the Bt2 and Btk horizons. Bulk density is highest inside the Bt1 horizon (1.83 g/cm^3). Measured pH-values are consistently higher than 7 indicating slightly alkaline conditions. Soil organic carbon never exceeds 0.7 %, the C/N-ratio of the Ap horizon is 9. Soil profile 5 has been finally classified as haplic luvisol developed from alluvial deposits which show at a depth of approximately 160 cm. The sandy material contains pebbly components from the surrounding metamorphic and granitic Paleozoic hills that are only slightly rounded.

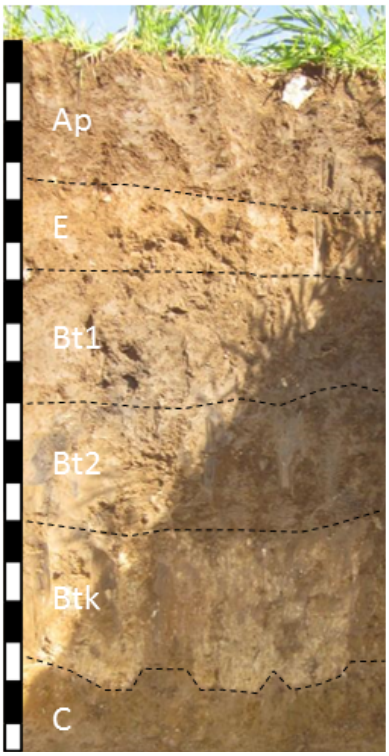


Figure 3.11: Soil profile no. 5

Table 3.5: Selected characteristics of soil profile no. 5

Horiz.	Depth	Sand	Silt	Clay	CaCO3	Carbon	BD[1]	pH
	cm			%			g/cm^3	$(CaCl_2)$
Ap	35	51.25	23.59	25.15	0.45	0.71	1.66	7.08
E	55	46.44	27.68	25.88	0.15	0.25	1.77	7.02
Bt1	90	37.71	23.63	38.67	0.08	0.26	1.83	7.00
Bt2	120	21.18	25.10	53.72	0.31	0.26	1.53	7.35
Btk	160	9.54	26.35	64.12	55.9	7.13	1.45	7.45
C	>160	65.68	10.68	23.64	1.3	0.23	1.57	7.37

[1] BD = Soil bulk density

3.4 Land use

Open forests and interrupted woodlands are the natural vegetation in the Mediterranean region including holm and cork oak, wild olive, lentisk and pine trees (Yaalon 1997). However, in recent days these kind of forests are scarce especially with respect to the Rio di Costara catchment which is mainly covered by agricultural fields and grassland. The more mountainous northern part of the basin is dominated by maquis shrubland (macchia) after intense deforestation and degradation. Macchia – widespread throughout the Sardinian island – is characterised by prickly and thorny vegetation in combination with stunted evergreen oaks and aromatic species such as lavender, myrtle and oleander (Verheye and de la Rosa 2006). The left picture in figure 3.12 from March 2010 shows some shrubs and plants of the macchia located in the northern region of the San Michele farm.

(a) Maquis shrubland (b) Rape – Field 21

Figure 3.12: Macchia vegetation photograph and agricultural land cover picture of field 21 at the San Michele farm

Regarding the fields of interest at the San Michele farm land use has been monitored during sporadic field campaigns from 2005. Being classified as fallow land until 2010, field 21 was covered in March 2011 by maize (in its southern parts) and rape as shown in figure 3.12b. In March 2013 it was partly cultivated with wheat. Field 33 remained completely bare during March 2011. However, it was subdivided into seven horizontal pieces during the last survey in March 2013 either raising field beans and wheat or being slightly covered with clover.

Chapter 4

Materials and methods

This chapter introduces all datasets that are relevant in the frame of this thesis and explains the methodology used to achieve the objectives described in chapter 1.2. In particular, a comprehensive summary is given with respect to the interpolation techniques applied for mapping of soil properties at two different scales at the Sardinian test sites described in chapter 3.

4.1 Workflow

Substantial steps of the (geo)statistical data processing are summarised in figure 4.1. The interpolation approaches used at the field and landscape scale are based on sets of sampled topsoil data and different predictor maps. Covariates supporting the spatial prediction at the entire Rio di Costara test site are derived from a digital elevation model and a geological map, whereas ancillary information at field level comes primarily from geophysical sensor data. After an extensive exploratory data analysis, important preprocessing steps include data splitting and (co-)variable selection. The interpolation procedure itself is divided into trend estimation and multivariate geostatistical modelling of the trend residuals. The final model is validated either based on leave-one-out cross-validation (field scale) or using an independent dataset (landscape scale) that has not been used at any

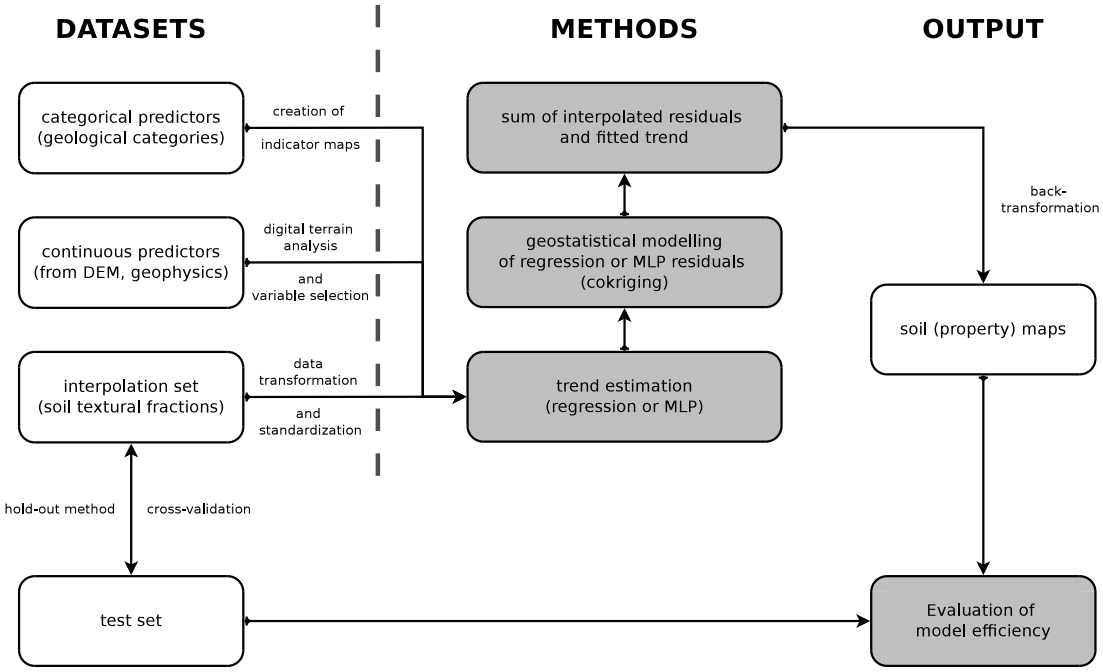


Figure 4.1: Flow chart showing the interpolation data processing scheme

stage of the model calibration. At the landscape scale, the developed neural network residual cokriging model is compared to common digital soil mapping techniques such as ordinary (co)kriging or regression (co)kriging. Finally, each soil property map is visualised and documented inside a web-based geoportal.

4.2 Soil sampling and laboratory work

Soil sampling is meant for collecting data in order to estimate some statistical parameter of an entire region (population) or to allow the prediction of soil properties at unvisited locations. A sampling strategy is defined by the primary goal of a soil survey, i.e. the estimation or prediction criterion, and its sampling design (Brus and de Gruijter 1997). The latter describes the procedure to select designated sample locations and basically follows two fundamental approaches: design-based or model-based sampling. The design-based approach relates to classical sampling theory and is regarded as ideal with respect to unbiased esti-

mation of global quantities such as the spatial mean (De Gruijter et al. 2006). It is also sometimes referred to as probability sampling with randomness being introduced by the design itself, whereas model-based or purposive sampling treats the values at any given location as random (De Gruijter and ter Braak 1990). Thus, model-based sampling is strongly connected to geostatistical theory and generally more efficient for spatial prediction purposes than probability sampling (Brus and Heuvelink 2007). However, model-based sampling requires prior knowledge of the underlying model (variogram), which is rarely the case in general-purpose surveys. For a more detailed discussion on design-based and model-based sampling philosophies refer to Brus and de Gruijter (1997), De Gruijter et al. (2006) or Allen et al. (2010). During the past two decades, several concepts of sampling strategy optimisation have been reported in soil science. Some methods aimed at well-dispersed samples in geographical space such as spatial coverage sampling (Royle and Nychka 1998). Other techniques focused on numerical optimisation, e.g. using simulated annealing, to obtain best sample locations with respect to minimised prediction variances in geostatistical applications (Van Groenigen et al. 1999). In the presence of secondary information, Gessler et al. (1995) and McKenzie and Ryan (1999) were among the first who stratified their sampling based on the values of correlated environmental variables. Hengl et al. (2003) suggested to use an equal range design on principal components of ancillary variables. Minasny and McBratney (2006) introduced a conditioned Latin hypercube method for optimal sampling using explanatory variables. However, optimisation in feature space only reflects the estimation of regression coefficients or neural network weights, while a good coverage of the investigation area is not accomplished. In order to balance these two conflicting requirements, Brus and Heuvelink (2007) proposed to incorporate secondary variables and to provide an optimal geographical spread by minimising the universal kriging variance. This procedure was later extended to the multivariate case by Vašát et al. (2010).

Once the decision on the sampling strategy is made, the number of desired sample points needs to be determined. In general, the final sample size for spatial prediction purposes depends on requirements regarding accuracy and spatial resolution of the target maps (Hengl 2009). The finer and more accurate resulting

soil property maps are supposed to be, the more samples are needed. However, soil surveys are expensive and time-consuming which in turn strongly limits the effective number of soil samples. In the context of this thesis, determination of sample size was guided by recommendations from literature regarding the minimum number of samples needed to adequately apply digital soil mapping techniques. Independent from scale, for instance, in kriging approaches proper (isotropic) variogram estimation requires between 100 and 150 samples (Webster and Oliver 1992; Kerry and Oliver 2008). Neural networks being data-driven techniques also strongly benefit from larger (training) set size depending in numbers, however, on the network complexity. As a rule of thumb, the amount of training cases should be 10 times the number of weights inside the network (Abu-Mostafa et al. 2012). Minasny et al. (2008) point out that data-mining tools such as neural networks usually relate to datasets ranging from 200 to over 1000 samples.

4.2.1 Soil sampling at the landscape scale

Corresponding to the preliminary considerations, a design-based, stratified, two-stage sampling design was used at the landscape scale aiming at the spatial prediction of target quantities at individual locations. This strategy follows the procedure of McKenzie and Ryan (1999) which has also been adopted recently, for instance, by Wiesmeier et al. (2011).

A first stratifying variable was chosen from the geological map described in chapter 3.2 by summarising the given units into four main categories: Quaternary deposits, Oligo-Miocene sedimentary deposits, Paleozoic intrusive rocks and Paleozoic metamorphic rocks. Two more stratifying variables – topographic wetness index (TWI) and potential incoming solar radiation (INSOLAT) – were derived from an SAR based digital elevation model (DEM) of ten metre resolution published through the geoportal of the Autonomous Region of Sardinia (http://www.sardegnageoportale.it). The parameter TWI was selected from a range of possible terrain attributes because of its proven value to the prediction of soil properties reported in numerous studies during the past decades (e.g. Moore et al. 1993; McKenzie and Ryan 1999; Hengl et al. 2003; Minasny and

McBratney 2006). While TWI is related to soil moisture, INSOLAT helps to reveal small-scale variation of climatic indicators such as air and soil temperature based on topographic position. Both selected relief parameter are almost uncorrelated to each other at the present study site and thus redundant dependencies to target variables are probably avoided. In order to classify the selected continuous DEM derivatives for stratification purposes, their density functions were used to calculate quantiles. Each pixel of the DEM raster was then grouped into one of three (INSOLAT) or one of four quantile classes (TWI). Combined with the four geological units this procedure led to an array of 30 different classes in the Rio di Costara catchment.

Up to six polygons of each available class were selected randomly, excluding areas smaller than 1 ha to avoid incorrect location of the points in the field. Further exclusion rules were applied before polygon selection masking out roads and buildings using a 20 m buffer. The selection procedure was repeated ten times and the set of polygons with the best geographical spread were chosen. In a final step, exact point locations were selected randomly from inside the chosen polygon features. At the same time, a second selection based on the same stratification and following the same procedure (selecting one polygon instead of six) was done in order to create an appropriate test set. Taking all this into consideration, the field trip in the autumn of 2010 produced an initial interpolation set of 121 samples and a test set of 24 points with locations as shown in figure 4.2 on page 34. Supplementary sets of purposive samples were obtained during a second survey in late spring 2011. This second sampling focused on polygons that have either not been considered during the first phase at all or were not adequately represented with respect to feature size. A total of 35 additional samples were collected for interpolation purposes, while the test set was enlarged by 17 more points.

The selection of sample point locations was done using Esri software (ArcGIS) extended by Hawth's Tools and later on its replacement the Geospatial Modelling Environment (GME) (Beyer 2011). At the landscape scale, 88 % of all desired points could actually be reached in the field and have been successfully sampled.

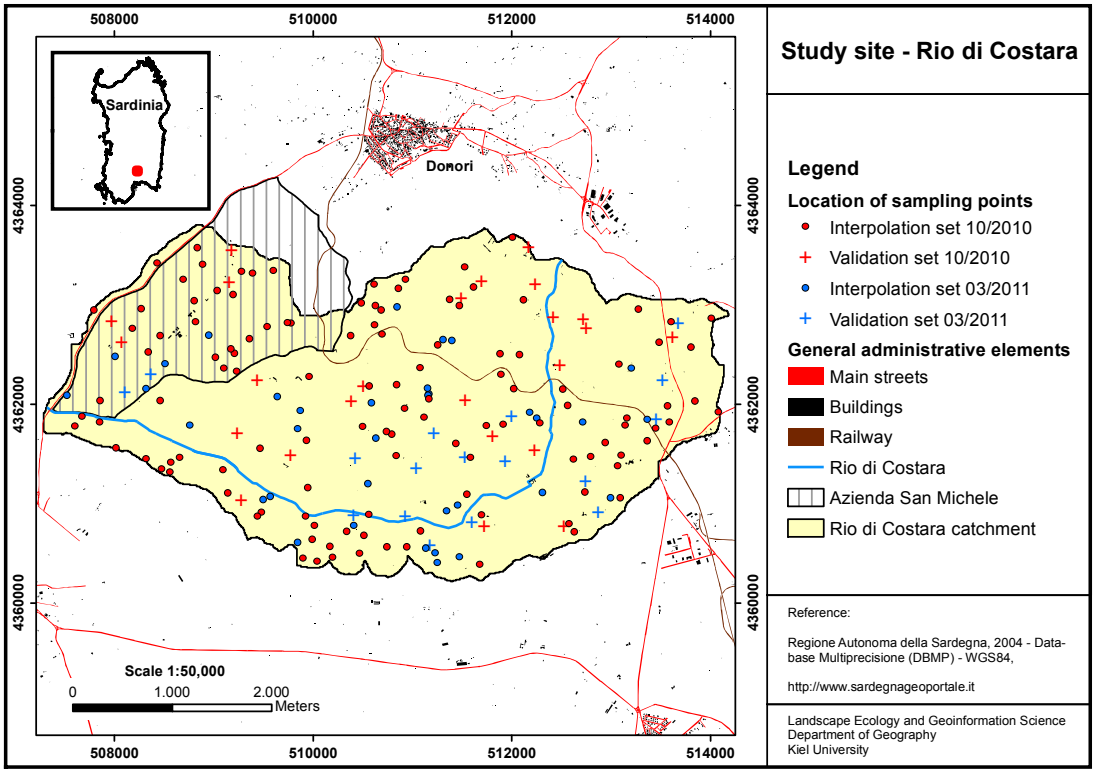


Figure 4.2: Location of sampling points in the Rio di Costara catchment

4.2.2 Soil sampling at the field scale

Both studied fields at the San Michele farm were sampled using a combination of systematic, regular sampling plans and design-based, stratified components. At field 21, a total of 43 soil samples were collected during two field campaigns. In early October 2010 the first 23 locations were sampled both from a triangular grid and along a transect running parallel to the maximum slope gradient. Additional 20 soil samples were taken in March 2011 following a stratified random sampling design (De Gruijter et al. 2006). The stratification based on three soil types according to Aru (1966) and two higher geological units which correspond to the field boundaries indicated in figure 4.3. Depending on the size of the stratifying polygons, three to seven points were chosen randomly from each strata with a minimum distance of 10 m between possible sample locations using GME (Beyer 2011). Field 33 was investigated mainly during a single field campaign in March

2013. In order to base any analysis on a data collection as similar as possible compared to field 21, a regular sampling scheme was supplemented by additional samples from five different soil units shown in figure 4.3. A total of 64 samples was available at field 33 after the final field campaign in spring 2013 including 7 points from the very first survey in 2010.

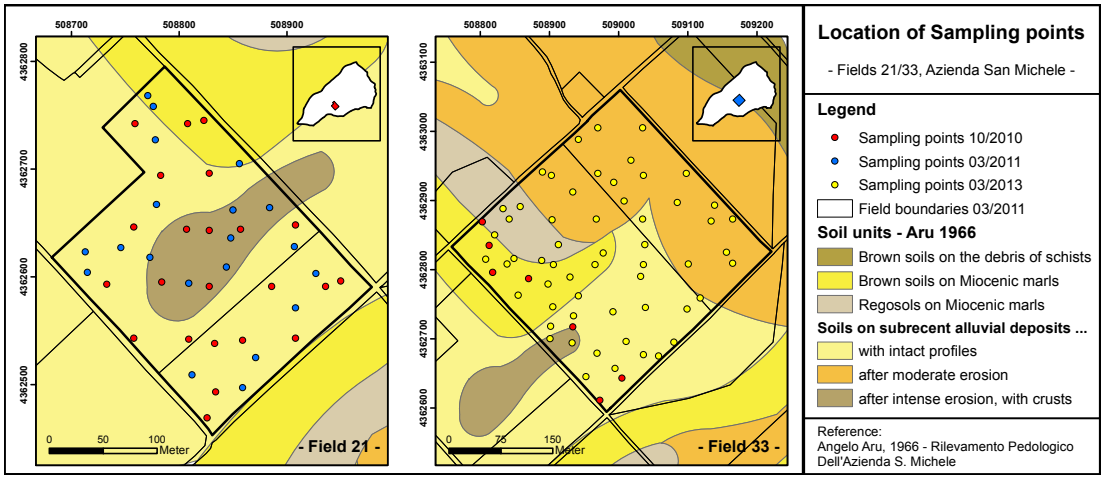


Figure 4.3: Location of sampling points at field sites

To collect enough soil material from up to three standard depths (0–30, 30–60, and 60–90 cm), each location was sampled at least three times using a gouge auger of the type Pürckhauer. Composite samples of the top 30 cm soil layer are of particular interest regarding the digital soil mapping projects of this thesis. Their horizontal support size, i.e. the area associated with measurements that form the composite sample at one single site, is approximately 1 m^2. All sample point locations were georeferenced using the Trimble Juno SB handheld device with a positional accuracy of about 2–5 m.

4.2.3 Laboratory analysis

Apart from soil bulk density, the complete laboratory analysis was performed at the Department of Geography in Kiel (Germany). The most relevant analysis with respect to the target variables of this thesis was the grain size distribution analysis. Following the sieve-pipette method in accordance with Köhn required

manifold pretreatment of soil samples (see Gee and Or 2002; Hartge and Horn 2009). In a first step, dried soil material (at 30 °C) was sieved to separate coarse particles from fine earth fractions (< 2 mm). The latter were subsequently treated with several chemical solutions to remove coatings that bind particles together. For instance, organic matter was removed by hydrogen peroxide (H_2O_2) and carbonates were eliminated by acidifying the sample using hydrogen chloride (HCl). The succeeding removal of iron oxides was done using a bicarbonate-buffered, sodium dithionite-citrate system after Mehra and Jackson (1960). Finally, the remaining material was dispersed in a sodium pyrophosphate ($Na_4P_2O_7$) solution. Following these preparation steps, wet sieving was applied for measuring sand fractions (2,000–630, 630–200 and 200–63 µm) and a sedimentation analysis after Köhn was used to determine silt (63–20, 20–6.3 and 6.3–2 µm) and clay (< 2 µm) contents (in weight percentages).

Besides soil texture, other parameters were measured, such as calcium carbonate content ($CaCO_3$) using Scheibler equiqment (DIN ISO 10693), and total carbon and nitrogen (C/N-Analyser EURO EA, Hekatech). Soil organic carbon (SOC) was calculated as the difference between measured total carbon and the inorganic C content of the quantified $CaCO_3$. Soil pH values were determined in 0.01 M calcium chloride ($CaCl_2$) at a soil/solution ratio of 1 to 2.5 (pH 330, WTW). In addition, soil bulk density (BD) was measured at a subset of sample locations (N = 49) based on soil sample rings with an inner volume of 100 cm^3. Three undisturbed soil samples per point were analysed using the lab facilities at the Azienda San Michele in October 2010. Each sample was dried at 105 °C for approximately 24 hours and the remaining mass of the oven-dry soil was divided by its core volume (Hartge and Horn 2009). Final BD was then obtained by averaging the three independent replicates.

Material from soil profiles were also analysed for exchangeable cations (Ca, K, Mg, Na) using flame atomic absorption spectrometry (AAS) (4100, PerkinElmer). Moreover, available phosphorus and potassium content was measured in calcium lactate solution at a spectral photometer (LAMBDA 2 S UV/Vis, PerkinElmer) or AAS, respectively. Free iron oxides (FeD) were determined through sodium dithionite as proposed by Mehra and Jackson (1960). Amorphous iron oxides

(FeO) were measured using the acid ammonium oxalate method of Tamm (1932) modified by Schwertmann (1964). Results regarding the soil profiles are partly shown in chapter 3.3. Although measurements of soil textural classes were based on seven particle size fractions, further analysis refers to the summarised main fractions of clay, silt and sand.

4.3 Derivation of covariates

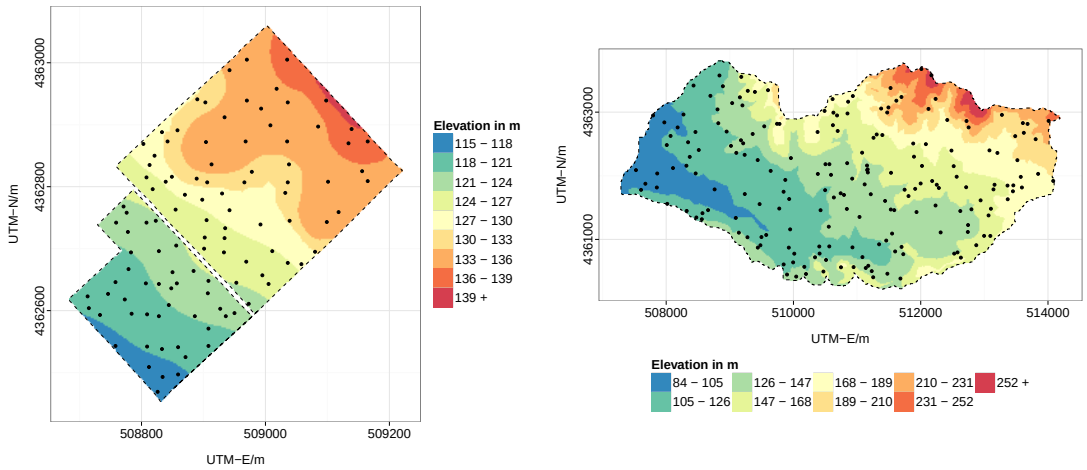
In the previous section, interpolation and test sets of soil data at two different scales were introduced. These datasets form the group of target variables in the modelling frame of this thesis. On the following pages, several environmental variables are described serving as predictors in the digital soil mapping approaches applied to this work. At the landscape scale, explanatory variables correspond to the soil forming factors topography and parent material. These variables were derived from a digital elevation model (DEM) and a geological map, respectively. However, regarding field scale, available geological maps were inaccurate in terms of spatial resolution. Hence, geophysical measurements from two different sources (gamma-ray spectrometry and electromagnetic induction) are used to support spatial predictions at fields 21 and 33.

4.3.1 DEM data and digital terrain analysis

Topography is an important soil forming factor according to early concepts of soil development (e.g. Jenny 1941), as well as more recent theories on environmental correlation (McBratney et al. 2003). It strongly influences small-scale meteorological conditions resulting in local soil moisture and temperature regimes. Moreover, topography determines the way water moves through the landscape and, therefore, affects the lateral transport of (soil) material (Florinsky 2011). In order to represent topography, a set of land-surface parameter is usually derived from DEMs using digital terrain analysis (Wilson and Gallant 2000). Digital terrain analysis (DTA) involves numerous calculation methods which are equivalent to

quantification techniques used in the fields of digital terrain modelling (Florinsky 2011) or geomorphometry (Pike et al. 2009).

Inside the Rio di Costara catchment, an SAR based DEM of originally 10 m resolution is used as source of information concerning topography. The given DEM originates from 2011 and is freely available at the geoportal of the Autonomous Region of Sardinia (http://www.sardegnageoportale.it). It is published under the Italian Open Data License in version 2.0 which is compatible with the Creative Commons Attribution Share-Alike License (CC BY-SA). The vertical and horizontal accuracy is reported with 2.5 m (Vacca et al. 2014). After resampling the DEM into grids with scale-specific target pixel size, digital terrain analysis provided 13 land-surface parameter listed in table 4.1. The calculation of terrain attributes was done in R using RSAGA (Brenning 2008) and SAGA GIS in version 2.0.8 (SAGA User Group Association 2011).



(a) Fields 21 and 33 (b) Rio di Costara catchment

Figure 4.4: Elevation grids used for digital terrain analysis: a) at the field scale, b) at the landscape scale. DEM of originally 10 m resolution was resampled to 2.5 m and 20 m, respectively. Black dots represent soil sample locations.

Morphological variables are commonly classified into primary and secondary attributes (Wilson and Gallant 2000). Primary attributes are usually calculated straight from a DEM, whereas secondary parameters require additional computation steps that are often built upon the primary ones. A slightly different

Table 4.1: Land-surface parameter from quantitative terrain analysis

Land-surface parameter	Abbreviation	Dimension
- Primary attributes -		
Elevation	ELEV	m. a. s. l.
Slope	SLOPE	m/m
Plan curvature	PLANC	1/100 m
Profile curvature	PROFC	1/100 m
Aspect	ASPECT	degrees clockwise from North
Divergence/convergence index	CONVG	-
- Secondary attributes -		
SAGA wetness index	SAGAWI	-
Topographic wetness index[1]	TWI	-
Stream power index	STREAMP	-
Length-slope factor	LS	-
Potential incoming solar radiation	INSOLAT	kWh/m^2, annual
Direct solar radiation	DIRECT	kWh/m^2, annual
Diffuse solar radiation	DIFFUS	kWh/m^2, annual

[1] TWI following Beven and Kirkby (1979)

classification scheme circulates among the geomorphometry community, however, distinguishing local morphology, hydrographic parameters and quantities related to climate (Pike et al. 2009).

Local morphology includes slope gradient and aspect as well as curvatures. In the frame of this thesis, local morphological variables are calculated following Zevenbergen and Thorne (1987). Accordingly, slope, aspect and curvature are functions of partial derivatives of a second-degree polynomial (Z) representing the land surface in a 3 x 3 moving window. Its first derivative, slope, is defined as the tangential plane to the central pixel of the 3 x 3 sub-matrix. The slope gradient reflects the angle of that plane with regard to a horizontal surface (in radian) and essentially influences gravity-driven mass transport. Aspect gives information on the maximum-slope direction (in degrees clockwise from north) revealing, for instance, differences related to incoming radiation. The second derivative of Z is curvature documenting the rate of which slope changes (Zevenbergen and Thorne 1987). There are basically two main directions of curvature considered in practice: First, following the maximum-slope direction (profile curvature) to detect prevailing water flow and mass transport processes. Second, perpendicular

to the slope (plan curvature) to measure convergence or divergence and, thus, the concentration of water in a local neighbourhood (see Moore et al. 1993). Whereas positive values are associated to convex curvature indicating dispersion, negative values correspond to concave curvature highlighting flow accumulation (Olaya 2004). As curvature values were usually small, they have been multiplied by 100 prior to any statistical analysis (after Zevenbergen and Thorne 1987). However, another way to represent convergence (positive values) and divergence (negative values) behaviour rather than using curvature values is based on a separate index, which is calculated from gradients.

Hydrological parameters or topographic indices are more sophisticated, process-based indicators of sediment transport in landscapes. Flow-accumulation parameters calculated in this work are topographic and SAGA wetness index, stream power index as well as length-slope factor. The latter is known from the universal soil loss equation (USLE) and helps to identify erosion and deposition processes (Moore et al. 1993). Moreover, stream power index stands for the strength of overland flow causing net erosion (Olaya 2004). The topographic wetness index (TWI) goes back to Beven and Kirkby (1979) and is one of the most widely used terrain attributes in digital soil mapping. The TWI is strongly related to soil moisture and reveals how much a given pixel contributes to overland flow. The analogous SAGA wetness index, however, varies from the traditional TWI by using a slightly different algorithm for catchment area calculation. It is supposed to perform better, if pixels are located in valley floors in vertical proximity to a channel (see Brenning 2008). Despite the differences, all four compound indices are variants from mathematical combinations of slope and specific catchment area (see Moore et al. 1993).

Potential incoming solar radiation and its direct and diffuse components are computed from the group of land-surface influenced climatic quantities. Each radiation grid is calculated on an annual basis using the SAGA lighting module with a lumped atmospheric transmittance of 70 %.

More detailed information regarding the exact terrain analysis procedure of SAGA GIS can be found in Olaya (2004), Conrad (2006) and Cimmery (2007). Among

others, Moore et al. (1993), Florinsky et al. (2002) and Duttmann and Sumfleth (2007) provide an extensive description of terrain attributes for spatial prediction purposes in soil science.

4.3.2 Geological categories

The complex geological setting of the Rio di Costara catchment has already been discussed in chapter 3.2. Figure 3.4 on page 18 shows the distribution of the main lithostratigraphic units at the given test site. The resource behind this map – originally created by the local geological agency PROGEMISA – was downloaded from the Regional Administration of Sardinia (http://www.sardegnageoportale.it, IODL 2.0). A total of six geological categories were considered as explanatory variables for digital soil mapping at the landscape scale.

Table 4.2: Main geological units of the Rio di Costara catchment

Description	Period/Epoch	Abbreviations/Legends	
		GEOPPR[1]	GERARCHIA[2]
Eluvial-colluvial deposits	Holocene	12	A221
Alluvial deposits	Holocene	18	A222
Terraced alluvial deposits	Holocene	22	A222
Sandstones of Serra Longa	Oligo-Miocene	470	B232
Granites of Gerrei	Carboniferous-Perm	1233	D231
Sandstones of San Vito	Cambrian-Ordovician	1465	E222

[1] PPR = Piano Paesaggistico Regionale (regional landscape plan), attribute key to associated legend file
[2] GERARCHIA = Hierarchy, from the corresponding final report at http://www.sardegnageoportale.it

Note that unlike in soil sampling (see chapter 4.2) quaternary deposits are not summarised for modelling purposes but treated as three distinct categories.

4.3.3 Geophysical measurements

Geological maps, as well as electro-magnetical or gamma radiometric data are often incorporated into digital soil mapping projects as a proxy for parent material or lithology (see Gray et al. 2014). Being one of the five (or seven) factors of soil formation as described by the clorpt (or scorpan) model (Jenny 1941; McBratney

et al. 2003), parent material basically provides the raw components of any soil development and, thus, strongly determines the recent distribution of physical and chemical soil properties such as texture.

Table 4.3: Covariates from geophysical measurements

Geophysical parameter	Abbreviation	Dimension
- Electromagnetic induction -		
ECa in horizontal coil orientation	EMIH	mS/m
ECa in vertical dipole mode	EMIV	mS/m
- Gamma-ray spectrometry -		
Potassium ^{40}K	GAMMAK	%
Thorium ^{232}Th	GAMMATH	ppm
Uranium ^{238}U	GAMMAU	ppm
Dose rate	GAMMADR	nGy/h
Th/K ratio	THKratio	-
Th/U ratio	THUratio	-
U/K ratio	UKratio	-

ECa = apparent electrical conductivity

While geological units are considered at the landscape scale as listed in table 4.2, lithology of field 21 and 33 is represented by numerical data from two different proximal sensing information sources: electromagnetic induction (EMI) and gamma-ray (γ-ray) spectrometry. All geophysical measurements were performed by Dr. Ulrike Werban and her team from the Department Monitoring and Exploration Technologies (MET) of the UFZ Helmholtz Centre for Environmental Research. Their field experiments at the Azienda San Michele were carried out from 29 September to 2 October 2010 as part of a cross-cooperation between the EU-FP7 research projects iSOIL (Werban et al. 2010) and CLIMB (Ludwig et al. 2010). Results from this collaboration are published, for instance, by Cassiani et al. (2012) focusing on soil-vegetation interactions and their effect on soil water balances.

4.3.3.1 Electromagnetic induction

Electromagnetic induction techniques applied in soil science measure apparent electrical conductivity (ECa) of a certain soil volume in mS/m (Corwin and

Lesch 2005). Recorded ECa values are related to a variety of soil properties including salinity, clay content and mineralogy, cation exchange capacity, bulk density, organic matter, as well as water content and temperature (Clay et al. 2001; Sudduth et al. 2005). Thus, ECa measurements serve as a promising covariate for spatial prediction of manifold soil-related variables, especially, if the relationship between ECa and the desired property is predominant in a given case study (see De Benedetto et al. 2012). However, due to the complexity of influencing factors, reliable interpretation of collected ECa data is difficult and usually non-unique and site-specific. Successful applications of EMI data in soil scientific projects are listed, for instance, in the review of Kuang et al. (2012). Details on how to work with common EM equipment can be found in McNeill (1980), while theoretical aspects of ECa measurements are summarised by Corwin and Lesch (2005).

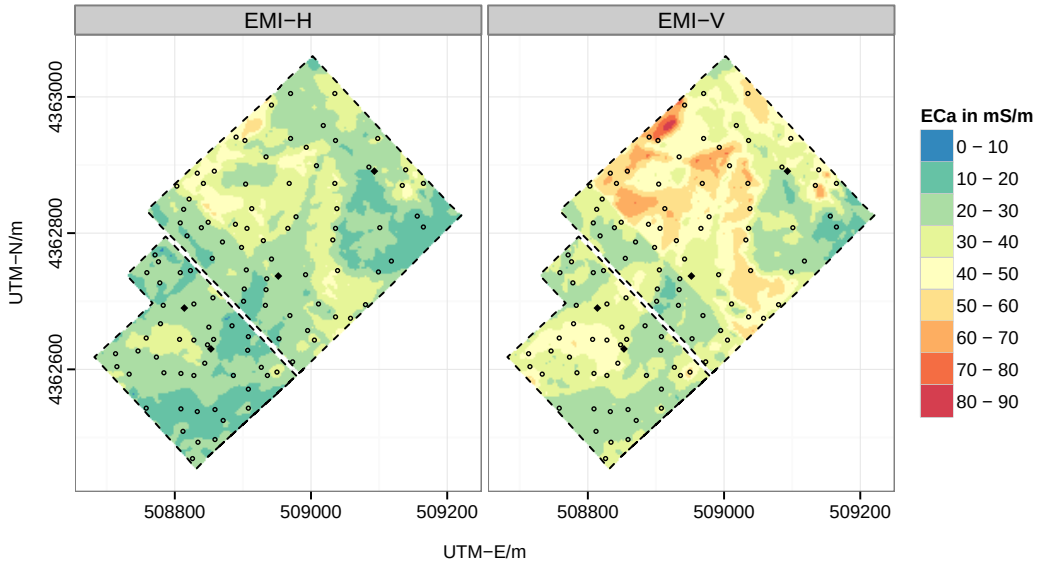


Figure 4.5: Maps of apparent soil electrical conductivity. ECa measured with an EM38-DD sensor in horizontal (left) and vertical (right) mode. Interpolated to a grid with a pixel size of 2.5 m using ordinary kriging. Dots represent sample locations. Rhombuses indicate the position of surveyed soil profiles.

In the frame of this thesis, soil ECa measurements at fields 21 and 33 were obtained by a Geonics EM38-DD (Geonics Ltd., Mississauga, ON, Canada) sensor. This particular instrument simultaneously measures in horizontal (EMIH) and

vertical (EMIV) modes of operation with exploration depths of up to 0.75 m and 1.5 m, respectively (see Callegary et al. 2007). As a consequence, it provides two measurements at all times: the horizontal coil orientation focusing on response from the topsoil and the vertical mode being dominated by subsoil properties (Cockx et al. 2009). The EM38-DD sensor was used in a mobile configuration (see Lausch et al. 2013), measuring within-line approximately every 0.3 m. The gap between two adjacent lines of measurements was about 10 m on average. A GPS device was connected to georeference any ECa observations as well as a field computer for data-logging purposes. The survey took place in early October 2010 on bare fields under dry weather conditions.

The collected ECa data was subsequently post-corrected by the UFZ Helmholtz Centre for Environmental Research. Preprocessing steps included noise removal, elimination of external influences like temperature, deletion of outliers caused by power supply lines as well as correction for instrumental drift and spatial offset. Finally, delivered ECa values were then interpolated to continuous predictors with a pixel size of 2.5 m using ordinary kriging (see 4.6.4). Resulting conductivity maps and corresponding variogram characteristics are presented in figure 4.5 and table 4.4, respectively.

4.3.3.2 Gamma-ray spectrometry

Gamma-ray spectrometry is another non-invasive as well as time and cost-effective geophysical method for measuring (co)variates that are associated with different physical and chemical soil properties. A γ-detector indirectly determines the concentrations of potassium (GAMMAK), uranium (GAMMAU) and thorium (GAMMATH) of an underlying soil volume. More precisely, it counts the decay rate of emitting nuclides ^{40}K and the decay series of ^{238}U and ^{232}Th (Dierke and Werban 2013). These intensities can then be converted into concentrations (% for K, ppm for U and Th) or dose rate (nGy/h), which is defined as a sum of GAMMAK, U and Th in the present study (see IAEA 2003). Varying concentrations of gamma-ray nuclides in a surveyed soil result from differences in the mineralogy and geochemistry of parent material, clay content and type of clay

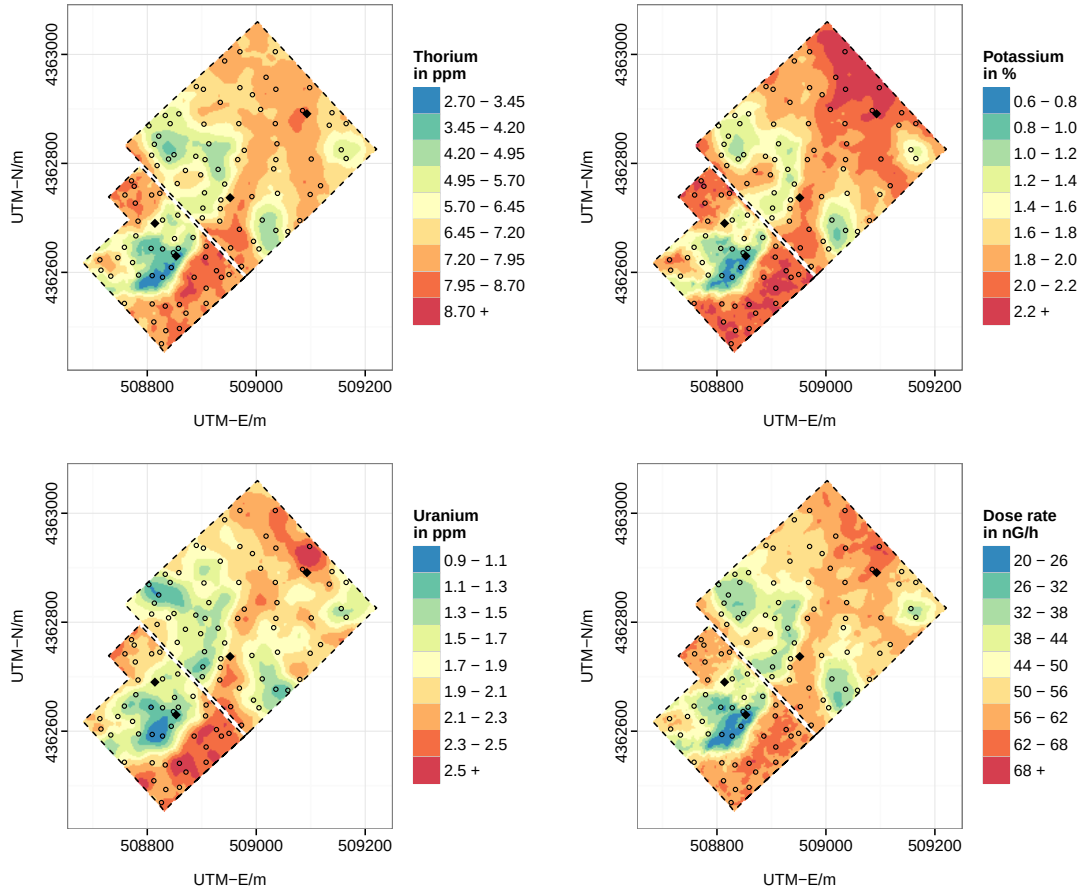


Figure 4.6: Spatial estimates of gamma-ray nuclides and total dose rate. Interpolated to a grid with a pixel size of 2.5 m using ordinary kriging. Dots represent sample locations. Rhombuses indicate the position of surveyed soil profiles.

minerals, as well as organic matter content (see Taylor et al. 2002; Dierke and Werban 2013). Additionally, radiation measurements are influenced by soil water content and active biomass (roots or plant cover), which significantly weaken the measured signal (Dierke and Werban 2013). With respect to soil texture, Megumi and Mamuro (1977) already concluded from their laboratory experiments that an increase in nuclide concentration corresponds with decreasing particle size. Note that 90 % of the gamma-ray signal originates from the top 30 cm of the surveyed soil volume (Cook et al. 1996). Unlike soil ECa measured by EMI devices, observed gamma-ray concentrations are therefore limited to the surface layer and cannot be sensibly used for predictions of subsoil quantities. For details on the

basic principles of gamma-ray spectrometry refer to the IAEA guidelines (IAEA 2003) or the textbook by Gilmore (2008).

In the present case study, radioactive elements at fields 21 and 33 were measured using a portable 512-channel gamma-ray spectrometer (4 l NaI(Tl)-crystal, automatic peak-stabilization) by GF Instruments with an energy range between 100 keV and 3 MeV. The detector was mounted on a GPS-positioned sledge and pulled by a four-wheel vehicle at a speed of 5 km/h. Its footprint was about 3 m in diameter and each individual counting interval lasted 5 s. The survey was conducted in the same period as the EMI measurements and, thus, under the same climate and land use conditions.

All three gamma-ray nuclides and the dose rate were interpolated to continuous maps with a grid resolution of 2.5 m using ordinary kriging (see 4.6.4). Final predictor maps and associated variogram characteristics are shown in figure 4.6 and table 4.4, respectively. Due to the fact that each radiant exhibits slightly different relationships with soil properties, pairwise nuclide ratios were calculated after interpolation.

Table 4.4: Interpolation of covariates from geophysical sensor data: variogram characteristics and accuracy measures

Parameter	Model	Fit.method	Nugget	Sill	Range	R^2	RMSE
- for field 21 -							
EMIH	Sph	WLS	21.22	50.3	103.6	0.55	4.53
EMIV	Sph	WLS	7.34	69.4	104.2	0.89	2.43
GAMMATH	Sph	WLS	1.04	4.4	140.8	0.63	1.05
GAMMAK	Sph	WLS	0.02	0.3	143.4	0.83	0.18
GAMMAU	Sph	OLS	0.37	0.6	156.3	0.30	0.64
GAMMADR	Sph	WLS	5.11	209.3	145.0	0.88	4.03
- for field 33 -							
EMIH	Sph + Sph	OLS	4.72	66.4	229.3	0.98	0.92
EMIV	Sph + Sph	OLS	2.52	162.5	202.2	0.99	0.79
GAMMATH	Sph	WLS	1.11	2.1	81.9	0.44	1.06
GAMMAK	Sph	WLS	0.03	0.1	83.3	0.70	0.20
GAMMAU	Sph	WLS	0.33	0.5	60.8	0.16	0.64
GAMMADR	Sph	WLS	7.21	69.5	80.3	0.80	3.89

Sph = Spherical model, OLS / WLS = Ordinary / Weighted least squares

R^2 = Coefficient of determination, RMSE = Root mean squared error: based on hold-out sample validation (1/3)

4.4 Exploratory data analysis

This section describes an approach to data analysis that is meant for characterisation of datasets prior to any (geo)statistical modelling. Kanevski and Maignan (2004) emphasise the importance of a priori understanding of data especially if data-driven techniques like artificial neural networks are involved in subsequent modelling steps. Introduced by John W. Tukey, exploratory data analysis (EDA) provides insight into a given dataset through a variety of numerical and graphical techniques (Tukey 1977). In the frame of this thesis, these methods are subdivided into univariate (one variable at a time) assessments regarding the target variables, bivariate analysis based on the concept of correlation as well as techniques to account for the spatial aspect of data.

A detailed introduction into EDA, focusing on environmental data, can be found in Goovaerts (1997) and Kanevski and Maignan (2004). Plant (2012) provides comprehensive examples regarding EDA in the fields of ecology and agriculture including reproducible R code.

4.4.1 Exploratory graphics and summary statistics

The univariate distributions of raw target variables are visualised in terms of histograms and box-and-whisker plots (see Tukey 1977; Dalgaard 2008; Burt et al. 2009). Probability densities are used instead of frequencies whenever histograms from datasets with varying numbers of observations are compared. The selection of histogram bin widths is guided by Scott's rule, depending only on the sample size and an estimate of the standard deviation (Scott 1979). For visual inspection of normality, curves of corresponding Gaussian distributions are fitted to every density plot. In order to numerically support this graphical evaluation, a Shapiro-Wilk test for normality is performed (Shapiro and Wilk 1965). The box-and-whisker plots, used at the landscape scale, show conditional distributions based on subsets according to categorical attributes. This procedure allows to investigate the relation between continuous target soil properties and potential categorical covariates considered for spatial interpolation such as geological units.

Outliers are specially marked and represent values outside one and a half times the interquartile range as proposed by Tukey (1977) and replicated, for instance, in Sachs and Hedderich (2006) or Burt et al. (2009).

Descriptive statistics of target variables are computed according to Tukey's five-number summary including the extremes (minimum and maximum value), a middle value (median) and the quartiles ("hinges") (Tukey 1977). For completeness, mean and standard deviation are also calculated as measures of central tendency and spread, respectively. The shape of distributions is expressed by the standardised moments skewness and kurtosis.

In order to investigate whether two samples such as validation and calibration sets at the landscape scale are representative for each other, two (simultaneous) statistical tests were conducted. The Hotelling T^2-test checks for the difference in multivariate means of two samples, while Bartlett's test examines to which extent the (two) datasets have common variance-covariance matrices. Refer to the papers by Jouan-Rimbaud et al. (1997) or Borovicka et al. (2012) for details on calculation and interpretation.

4.4.2 Correlation and factor analysis

Joint distributions of pairwise attributes are modelled taking into account all continuous potential covariates at a given scale. Numerically, bivariate relation is expressed by the covariance σ_{12} and its standardised form, the Pearson correlation coefficient ρ_{12}

$$\sigma_{12} = \frac{1}{n}\sum_{i=1}^{n}(z_1(i) - \mu_1) \cdot (z_2(i) - \mu_2) \tag{4.1}$$

$$\rho_{12} = \frac{\sigma_{12}}{\sigma_1 \cdot \sigma_2} \quad \in [-1, 1] \tag{4.2}$$

with μ_1, σ_1 and μ_2, σ_2 being arithmetic mean values and standard deviations of variables Z_1 and Z_2, respectively (Goovaerts 1997). Useful graphics to examine

the relation between two variables are bivariate scatterplots. Both, correlation coefficients and scatterplots of target variables and continuous covariates are summarised in the lower and upper panels of a scatterplot matrix.

To account for multicollinearity among continuous explanatory variables at field scale, a principal component analysis (PCA) was performed prior to regression modelling. Being a technique that basically transforms interrelated variables into uncorrelated, standardised principal components (PCs), PCA has a considerable tradition in digital soil mapping projects (e.g. Hengl et al. 2004). Besides eliminating multicollinearity, PCA is frequently used for dimensionality reduction (Jolliffe 2002), which is particularly important in case studies with relatively small sample sizes but an extensive amount of (co-)variables. In the frame of this thesis, PCA was implemented in R as a singular value decomposition on the mean centred data matrix. In addition, variables were scaled to have unit variance (see .4.13). Biplots are presented to examine the first two PCs of each analysis, who retain the largest part of the variation among all considered predictors. In order to identify the most influential variables for each selected PC, loading coefficients are listed. The right number of factors to be considered for further analysis is determined by those eigenvalues that exceed a critical value, calculated after Karlis et al. (2003). This selection is additionally verified by scree plots visualising explained variances corresponding to ordered principal components. For mathematical insights and geometrical interpretations related to PCA refer to the textbook by Jolliffe (2002). Greenacre (2010) provides a comprehensive introduction into biplots, while Gabriel (1971) focuses on biplots with application to PCA.

A factor analysis for mixed data (FAMD) was considered to reduce dimensionality and to avoid multicollinearity among the predictors at the landscape scale. FAMD combines elements of principal component and correspondence analysis for analysing quantitative and qualitative variables at the same time (Pagès 2014). However, the resulting factors were found to be inadequate to be used as covariates in spatial prediction (results not shown, see .4.13 for calculation using the FactoMineR package in R). Thus, variable selection for spatial interpolation at that particular scale was instead based on correlation analysis results. Only (terrain) attributes, which are significantly ($P < 0.05$) related to the (transformed)

target variables, are considered as covariates. To control multicollinearity, (land-surface) parameters are excluded if they exceed a critical correlation level of 0.65 among each other (see .4.13 for implementation details).

4.4.3 Trend detection and variography

Spatial aspects are crucial in environmental modelling and several exploratory techniques exist to inspect the spatial features of a dataset. For global trend evaluation and the detection of spatial outliers, proportional symbol postplots are created. Each observation is plotted according to its coordinates with circle sizes representing the proportion of the particular target variable.

Local spatial dependence among raw and transformed target variables is analysed through exploratory variography. The variogram $2\gamma(\cdot)$ is a measure of spatial relation defined as

$$2\gamma(\mathbf{h}) = var[Z(\mathbf{s}) - Z(\mathbf{s}+\mathbf{h})] \tag{4.3}$$

where $Z(\mathbf{s})$ and $Z(\mathbf{s}+\mathbf{h})$ are regionalised random variables at locations $\mathbf{s}$ and $\mathbf{s}+\mathbf{h}$. For interpolation purposes such as kriging, it is sufficient to know $\gamma(\cdot)$, often referred to as semivariogram. The term "semi" points to the fact that half of the variance of differences is all that is needed, because each point pair is considered twice during calculation (Webster and Oliver 2007). However, the notions of variogram and semivariogram are frequently mixed up in the geostatistical literature. Bachmaier and Backes (2011) discussed this confusion and recommended to interpret the values of a variogram as "entire variances of observations at a given spatial separation (lag)". Following their suggestion to avoid the term "semi", empirical variances $\hat{\gamma}(\mathbf{h})$ are estimated under the constant-mean assumption based on the method-of-moments after Matheron (1971),

$$\hat{\gamma}(\mathbf{h}) = \frac{1}{2} \cdot \frac{1}{N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} [z(\mathbf{s}_i) - z(\mathbf{s}_i+\mathbf{h})]^2 \tag{4.4}$$

where $z(\mathbf{s})$ and $z(\mathbf{s}+\mathbf{h})$ are measured values of Z at locations $\mathbf{s}$ and $\mathbf{s}+\mathbf{h}$ with $N(\mathbf{h})$ being the number of point pairs separated by a particular lag $\mathbf{h}$. In practice the calculation of averaged squared differences requires choices about the set of lags similar to the binning in a histogram. Journel and Huijbregts (1978) provided two practical rules regarding the grouping of individual distances for variogram estimation: First, the distance of reliability (cutoff), i. e. the separation distance up to which point pairs are considered for calculation, is restricted to a value smaller than approximately half of the maximum distance between point pairs in the area of interest. Second, the bin width is defined in a way that the number of point pairs for each individual lag is greater than 30–50. In accordance with these recommendations, a maximum cutoff distance of 3625 m and a constant bin width of 125 m are chosen for variogram estimation in the frame of this thesis. For visual inspection, the resulting variances are finally plotted against the lag distances.

The sample variogram uncovers spatial variation in the given dataset, but is, for instance, not sufficient for the prediction at unknown locations. Thus, theoretical functions, that continuously reflect variances for all distances and fulfil certain mathematical conditions such as being conditionally negative-definite (Cressie 1993), need to be fitted. Since this constraint is relatively difficult to prove, the user is prone to choose among authorised models for which non-zero variances are ensured. In practice the most commonly used variogram function is the spherical model, which is characterised by an increase in the region of small distances and a constant development after reaching a threshold value

$$\gamma(h) = \begin{cases} c \cdot \left(\frac{3}{2}\frac{h}{a} - \frac{1}{2}\frac{h^3}{a^3} \right) & \text{for } h \leq a \\ c & \text{for } h > a \end{cases} \tag{4.5}$$

where c is the sill variance and a is the range (Clark 1979). Other variogram models used in this work, are the exponential model

$$\gamma(h) = c \cdot \left(1 - \exp\left(-\frac{h}{r} \right) \right) \tag{4.6}$$

with practical range r as well as the linear model

$$\gamma(h) = wh^{\alpha} \quad \text{for } \alpha = 1 \tag{4.7}$$

with gradient w describing the intensity of variation (Webster and Oliver 2007).

As evident from equations 4.5 to 4.7 variogram models are defined by several parameters, namely the nugget, sill and range. The nugget variance refers to an infinitesimally small separation distance representing measurement error or small-scale variability that is not covered by the given sampling intervals. The range parameter (if it exists) denotes the distance at which the variogram model flattens out and point pairs beyond this separation distance are said to be spatially uncorrelated. The variance corresponding to the range parameter is called the sill, which is also the total variance of the underlying process. Note that the exponential model does not have a constant range, unlike the spherical model, for instance. However, an effective range of approximately $3r$ is commonly reported for this type of model, representing the distance where 95 % of the sill variance are reached (see Cressie 1993; Webster and Oliver 2007).

Fitting a variogram model following the classical approach requires an initial guess on the parameters taken from the sample variogram (Hengl 2009, p. 130). Afterwards, more "optimal" parameters are obtained by model-fitting based on weighted least squares using a method which is proportional to the number of point-pairs in each lag and inversely proportional to the square of distance. Alternative fitting procedures that, for instance, avoid the rather subjective initial guess are based on maximum likelihood estimation (Lark 2000).

It is also common practice to combine theoretical models to account for more complex structures in the sample variogram (Webster and Oliver 2007). Such a nested variogram was used, for instance, with regard to clay content in the Rio di Costara test site. Up to three components were considered, resulting in a model with equation

$$\gamma(h) = \begin{cases} c_0 & \text{for } h = 0 \\ c_0 + c_1 \cdot \left(\frac{3}{2}\frac{h}{a_1} - \frac{1}{2}\frac{h^3}{a_1^3}\right) + c_2 \cdot \left(\frac{3}{2}\frac{h}{a_2} - \frac{1}{2}\frac{h^3}{a_2^3}\right) & \text{for } 0 < h \leq a_1 \\ c_0 + c_1 + c_2 \cdot \left(\frac{3}{2}\frac{h}{a_2} - \frac{1}{2}\frac{h^3}{a_2^3}\right) & \text{for } a_1 < h \leq a_2 \\ c_0 + c_1 + c_2 & \text{for } h > a_2 \end{cases} \tag{4.8}$$

where c_1 and a_1 are the sill and range of a component representing short-range variability, while c_2 and a_2 are the parameters associated to the long-range part of variation and c_0 accounts for some nugget variance.

Due to the fact that variogram modelling is based on a separation vector, direction of spatial dependence is important, as well. However, sample size was limited and not significantly greater than 100 to 150 samples that are needed to properly estimate a variogram according to Webster and Oliver (1992) and Kerry and Oliver (2008). Thus, only isotropic variograms are considered in the frame of this thesis and anisotropy issues are neglected. Consequently, authorised variogram models 4.5 to 4.8 are shown in terms of a scalar measure h representing distance only instead of the full lag vector $\mathbf{h}$.

For mathematical details with regard to variogram theory refer to Christensen (1991) or Cressie (1993). Webster and Oliver (2007) provide a detailed description and interpretation of different authorised variogram models.

4.5 Data preparation

Working with (spatial) environmental data from different sources usually requires considerable preprocessing before (geo)statistical analysis can be carried out. In the frame of this thesis, early data processing included conversion from file-based formats such as spreadsheets, Esri shapefiles and GeoTIFFs into R-objects. After detecting and eliminating irregularities among those objects and performing an exploratory data analysis, three preparation steps were of particular interest:

First, transformations were performed to meet certain assumptions about the model and the underlying data. Second, data splitting methods were applied for the production of representative samples with respect to model selection. Third, the pixel size of the target grids was determined.

R packages used regarding data import and handling are rgdal (Bivand et al. 2014) and sp (Pebesma and Bivand 2005; Bivand et al. 2013), respectively. The coordinate reference system of all spatial data applied or produced in the frame of this thesis is WGS84/UTM Zone 32 N (EPSG: 32632).

4.5.1 Data transformation

Compositional (or closed) data like soil texture usually quantify parts of some whole, carrying only relative information (Pawlowsky-Glahn and Egozcue 2006) and resulting in additional constraints with respect to statistical modelling. Taking into account point locations $\mathbf{s}_i$ from a spatial region D, a regionalised composition is formally defined as a vector random function $\mathbf{P}(s_i)$ with realisation

$$\mathbf{p}(\mathbf{s}_i) = [p_1(\mathbf{s}_i), p_2(\mathbf{s}_i), \ldots, p_k(\mathbf{s}_i)]^\top \tag{4.9}$$

where superscript $\top$ stands for transposition (Pawlowsky et al. 1995). The k elements of any composition are positive

$$p_j(\mathbf{s}_i) > 0 \quad \text{for } j = 1, 2, \ldots, k \tag{4.10}$$

and sum to a constant that is 100 (%) in the context of soil textural fractions:

$$\sum_{j=1}^{k} p_j(\mathbf{s}_i) = 100. \tag{4.11}$$

Both constraints strongly work against the assumption of an underlying unbounded random process such as the multivariate Gaussian, since the elements

of compositions are not free to vary in real space $\mathbb{R}^k$. In addition, the latter constraint on the constant sum also causes spatial dependence between the components of a regionalised composition which is referred to as spurious spatial correlation (Pawlowsky et al. 1995). As a consequence, the results obtained from standard (spatial) prediction methods, applied to raw compositional data, must be interpreted with care and are not necessarily prime. To overcome these issues Aitchison (1982) introduced alternative approaches for statistical analysis of compositions based on log-ratio transformations. Pawlowsky et al. (1995) applied this idea to spatial prediction problems. Formally, the additive log-ratio (alr) transformation of $\mathbf{p}(\mathbf{s}_i)$ leads to variate $\mathbf{w}(\mathbf{s}_i)$:

$$\mathbf{w}(\mathbf{s}_i) = alr(\mathbf{p}(\mathbf{s}_i)) = \left(ln\frac{p_1(\mathbf{s}_i)}{p_k(\mathbf{s}_i)}, ln\frac{p_2(\mathbf{s}_i)}{p_k(\mathbf{s}_i)}, \ldots, ln\frac{p_{k-1}(\mathbf{s}_i)}{p_k(\mathbf{s}_i)} \right). \tag{4.12}$$

Its inverse, the additive generalised logistic (agl) transform is then defined as

$$\mathbf{p}(\mathbf{s}_i) = alr^{-1}(\mathbf{w}(\mathbf{s}_i)) = agl(\mathbf{w}(\mathbf{s}_i)) = \frac{100 \cdot [exp(\mathbf{w}(\mathbf{s}_i)), 1]}{1 + \sum_{j=1}^{k-1} exp(w_j(\mathbf{s}_i))} \tag{4.13}$$

with 100 being the constant of the composition at hand (Pawlowsky et al. 1995). Regarding the soil textural fractions of this thesis, sand was chosen as divisor in equation 4.12. Additive log-ratios were preferred to other possible variants suggested, for instance, by Aitchison (1986) and Egozcue et al. (2003), because they are relatively simple to interpret. Additionally, alr coordinates have some beneficial characteristics with respect to (geo)statistical modelling such as non-singular covariance matrices (see Pawlowsky-Glahn and Olea 2004).

To summarise, alr transformation allows for the use of standard (geo)statistical techniques, since it basically transfers compositions from k-dimensional, restricted space $\mathcal{S}^k$ (known as the simplex) to unconstrained real space $\mathbb{R}^{k-1}$. One common way to visualise and explore a three-part simplex in soil sciences, is by plotting a triangular ternary diagram. In the context of particle size fractions Lark and Bishop (2007) suggest to add contours of equal compositional Mahalanobis distances to the ternary plot. Assuming the alr coordinates to be multivariate

normally distributed, these contour lines are ellipses in real plane representing equal probability. Projected onto the simplex, however, they exhibit distortion especially near the edges and vertices caused by the distributional constraints of the data. Here, the Mahalanobis distance δ_m at any composition $\mathbf{p}$ on the simplex is calculated from the mean vector $\bar{\mathbf{x}}$ and covariance matrix $\mathbf{\Sigma}$ of the alr-transformed data

$$\delta_m(\mathbf{p}, \bar{\mathbf{x}}) = \left([alr(\mathbf{p}) - \bar{\mathbf{x}}]^\top \mathbf{\Sigma}^{-1} [alr(\mathbf{p}) - \bar{\mathbf{x}}]\right)^{\frac{1}{2}} . \tag{4.14}$$

Lark and Bishop (2007) propose using a ternary plot superimposed on contours of δ_m as a diagnostic tool to investigate to which extent single compositions statistically suffer from inherent constraints. It, therefore, provides a general impression whether (geo)statistical analysis based on alr coordinates is not only theoretically sound, but also practically more promising compared to standard procedures on raw data. Several functions summarised in the R-soiltexture package (Moeys 2014) are used to plot the ternary diagrams and corresponding Mahalanobis distances within this work (see .4.12).

For further insights into compositional theory, characteristics and operations, consider the monographs by Aitchison (1986) or Pawlowsky-Glahn and Olea (2004). Van den Boogaart and Tolosana-Delgado (2013) provide comprehensive examples of statistical calculations in the presence of compositional data using R. In the frame of this thesis, functions from the R-compositions package are used to transform between soil textural fractions and alr coordinates (van den Boogaart et al. 2014).

4.5.2 Data splitting

Before building a network-based model that is meant to accurately predict at unknown locations inside the Rio di Costara catchment, calibration data is subdivided into training and validation sets. These two distinct sets are required for model selection purposes including parameter estimation and the determination

of an appropriate network architecture. The independent test set mentioned in section 4.2 is reserved for assessing the generalisation ability of the final, fully-trained model.

A common technique to subset a given data collection is random sampling. However, this very simple approach does not account for the data itself and, thus, neither ensures representative sets, nor produces unique results. It is, therefore, often preferred to use subset selection methods that uniformly cover the multi-dimensional space among a combination of variables (Daszykowski et al. 2002). Such a uniform design has been introduced by Kennard and Stone (1969) who based their sequential algorithm (KS) on Euclidean distances between the independent variables of an experimental region. Galvão et al. (2005) eventually extended the KS algorithm to cover not only the response vector but also the dependent variable of a particular sample. In the context of neural network modelling, Saptoro et al. (2012) recently based their dataset partitioning on a KS version that uses Mahalanobis distances instead of the Euclidean ones.

According to this methodological progress, the Matlab code published by Galvão et al. (2005) was implemented in R, replacing Euclidean by Mahalanobis distances as multivariate measure of variability among continuous variables in the given sample (see .4.21). As suggested by Saptoro et al. (2012), 20 % of the calibration data were selected for validation. Note, however, that the outlined procedure is only relevant at the landscape scale, since model building and evaluation at field level is based on resampling techniques (see section 4.7). For formal descriptions of the partitioning procedure, refer to Kennard and Stone (1969), Galvão et al. (2005) and Saptoro et al. (2012).

4.5.3 Defining target grid size

Mapping projects in soil science usually end in raster images of some desired property. An important decision prior to spatial prediction is, therefore, related to the pixel size of the output maps if not specified by subsequent applications. In a much-noticed paper, Hengl (2006) based the selection of an appropriate grid

resolution on simple cartographic and statistical concepts. Considering inspection density and working scale aspects, output grid size px was determined by

$$px = \sqrt{obs \cdot A \cdot \frac{10^6}{n}} \cdot 100 \cdot 0.0005 \tag{4.15}$$

where A is the study site area in km^2, n the given sample size and obs equals 2.5 observations per cm^2 on the map (Hengl 2006). The latter value balances the relationship between grid size and observation density, while the term 0.0005 relates grid resolution with scale number. Both values are recommendations following certain cartographic rules of thumb which are discussed in detail by Hengl (2006). Applied to the Rio di Costara catchment with an area of 16.44 km^2 and 197 samples, representing a working scale of approximately 1:50,000, a target grid size of 20 m was deemed optimal. This choice is also coherent with both, the resolution or scale of considered covariates (10 m DEM and 1:25,000 geological map) and the positional accuracy of the point measurements (Trimble Juno SB handheld device, 2–5 m). In the scope of field-related mapping, calculations according to equation 4.15 led to a recommended grid resolution of 2.5 m.

4.6 Spatial interpolation

Spatial interpolation or mapping outlines the procedure of estimating one or more target quantities at unknown locations in an entire region of interest. For this purpose, various deterministic and (geo)statistical methods have been applied successfully on different scales and in different types of landscape (see chapter 2 from page 7). This section explains the principles behind those techniques considered for soil texture mapping at the Sardinian test sites. First, multiple linear regression and artificial neural networks are introduced for modelling trends and large scale structures among the target variables. Both non-spatial prediction techniques build upon auxiliary information about soil forming factors derived and described in section 4.3. Second, inverse distance weighting is briefly sketched, representing a simple deterministic interpolation technique used as a

benchmark model in the frame of this thesis. Third, (multivariate) geostatistical kriging variants are portrayed providing stochastic solutions to model spatial variation based on auto-correlation functions. Finally, hybrid methods are described. These combine the benefits of a flexible, non-spatial representation of the trend and a geostatistical model by jointly using multiple linear regression (field scale) or artificial neural networks (landscape scale) with kriging techniques.

All calculations with respect to spatial prediction were done focusing on functions from the R-gstat package (Pebesma 2004). In addition, RSNNS (Bergmeir and Benítez 2012) was used for neural network modelling.

4.6.1 Multiple linear regression

Multiple linear regression (MLR) is recognised as the most widespread statistical tool for predicting a certain target variable Z, for instance, at a location in space with coordinates vector $\mathbf{s}_0$ that depends in a linear way on q independent covariates X

$$Z(\mathbf{s}_0) = \beta_0 + \beta_1 X_1(\mathbf{s}_0) + \cdots + \beta_q X_q(\mathbf{s}_0) + \epsilon \tag{4.16}$$

where β_0 is the intercept, $\beta_{1,\ldots,q}$ are the regression coefficients for the explanatory factors and ϵ represents an error term. For a set of n (observation) points and a considerably high number of covariates, matrix notation is more convenient and the multiple linear regression model can be written as

$$\mathbf{Z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{4.17}$$

with $\mathbf{X}$ being the design matrix, $\boldsymbol{\beta}$ representing the regression coefficients including the intercept and error $\boldsymbol{\epsilon} \sim^{iid} N(\mathbf{0}, \sigma^2\mathbf{I}_{n\times n})$. The latter expression summarises three important prerequisites for a successful application of MLR, namely statistical independence and homoscedasticity (constant variance) of the errors as

well as normality of the error distribution. A common approach to estimate the regression coefficients is the method of ordinary least squares (OLS)

$$\hat{\boldsymbol{\beta}}_{OLS} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{z} \tag{4.18}$$

where $\hat{\boldsymbol{\beta}}_{OLS}$ is the vector of estimated regression coefficients, $\mathbf{X}$ is a matrix of explanatory factors at known locations and $\mathbf{z}$ is the vector of measured target values. In spatial prediction problems, regression coefficients are often fitted by using generalised least squares (GLS)

$$\hat{\boldsymbol{\beta}}_{GLS} = (\mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\Sigma}^{-1} \mathbf{z} \tag{4.19}$$

with $\boldsymbol{\Sigma}$ being the covariance matrix of the residuals usually estimated from an initial OLS regression. The big advantage of using GLS instead of OLS is that it formally allows the regression residuals to be (spatially) correlated and, thus, accounts for one of the major assumptions of MLR often violated by natural phenomena. Another issue which is frequently found in mapping case studies based on environmental correlation is related to multicollinearity, that is, severe direct relationships between explanatory variables potentially leading to unstable estimates of the regression coefficients (see Neter et al. 1996, ch. 7.6). In the frame of this thesis, multicollinearity effects are avoided by using principal components instead of raw covariates. These factors are orthogonal to each other, and thus, uncorrelated by definition. The number of initially selected principal components is then reduced through stepwise regression in order to identify models which are both effective and parsimonious (see Hastie et al. 2009). Effectiveness is thereby judged in terms of the Akaike Information Criterion (AIC), which is a goodness-of-fit measure that considers the number of involved parameters (Akaike 1998). The model with the lowest AIC is finally used for prediction.

All linear models fitted by OLS are examined using several accepted summary statistics and regression diagnostics. The statistical significance of every model and each single regression coefficient is analysed through traditional F- and t-

statistics, respectively. Coefficients of (multiple) determination are used to assess the overall performance of final regressions. A few hypothesis tests are applied to evaluate, whether important MLR assumptions are actually met by the chosen models. These numerical quantities include Shapiro-Wilk normality tests (Shapiro and Wilk 1965), Breusch-Pagan tests for equal variance (Breusch and Pagan 1979; Cook and Weisberg 1983) and the calculation of Moran's I in a global setting to check for remaining spatial auto-correlation among the MLR-residuals (Moran 1950). The latter has been performed using the R-spdep package (Bivand 2014) and requires the determination of a spatial weights matrix. In the context of field 21 and 33, a binary weighting strategy without row standardisation was implemented using a fixed set of four nearest neighbours to properly depict any spatial relationships. For each hypothesis test, p-values are calculated to interpret the test results. If a certain probability value falls below the common significance level of 0.05, the null hypothesis of the given test is rejected. In addition to statistical significance tests to assess MLR assumptions, diagnostic methods for the detection of influential observations are applied. Identifying anomalies among the regression residuals is referred to as outlier detection and based on (externally) studentized residuals. On the contrary, leverage points are solely limited to extreme observations of explanatory factors commonly indicated by unusually high hat values. Points that affect both the coefficient estimates and fitted values are finally classified with respect to the concept of Cook's distance (Cook 1977).

For more details on statistical inference including hypothesis testing refer to any standard textbook (e.g. Rao 2002; Sachs and Hedderich 2006). A comprehensive introduction into linear regression is given, for instance, by Kutner et al. (2004) or Weisberg (2014), while Faraway (2004) and Fox and Weisberg (2011) focus on the implementation of linear models with R.

4.6.2 Artificial neural networks

Artificial neural networks, NNs for short, are an extensive collection of rather flexible, non-linear tools for data analysis (Sarle 1994). Originally developed to simulate the human brain (McCulloch and Pitts 1943), NNs became relevant as

a powerful statistical modelling alternative in the 1980s and 1990s due to rapidly growing computational resources. With regards to (spatial) prediction, the most widely used NN model is the multi-layer perceptron (MLP) (Bishop 1995). This particular network type is characterised by a specific set-up and learning routine described in the following paragraphs.

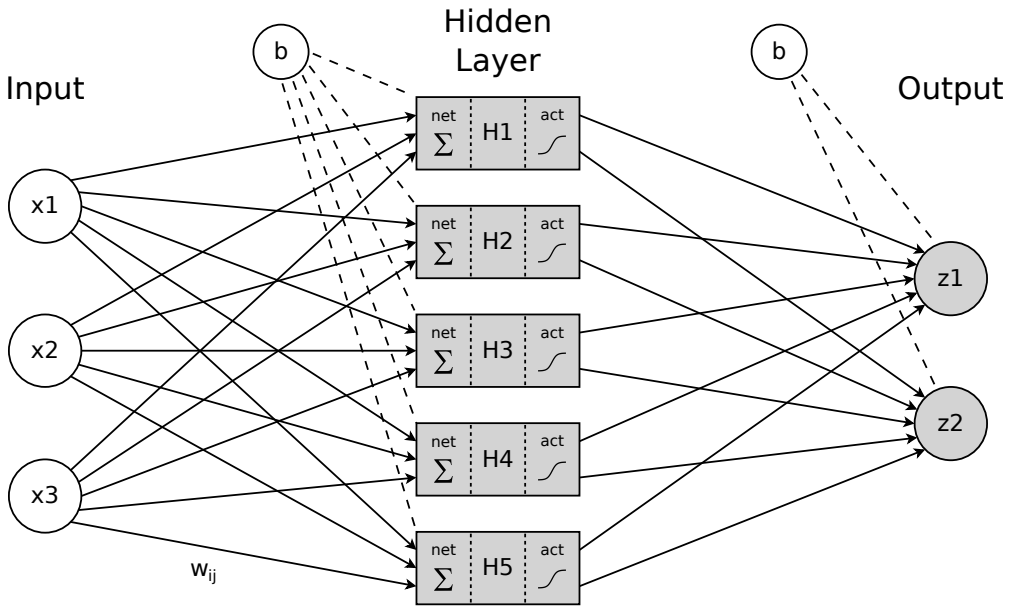


Figure 4.7: The structure of a multi-layer perceptron

Regarding structure, a perceptron basically denotes the key feature of any network, that is, the elementary processing unit also known as node or neuron. These units are grouped into layers that are linked together with (synaptic) weights assigned to each connection reflecting its importance (Warner and Misra 1996). In particular, any MLP comprises exactly one input and one output layer with one or more so-called hidden layers in between. All units of a specific layer are, thereby, fully connected to those of the next layer, while no direct links are established between input and output nodes. Thus, MLPs are feed-forward networks where information only propagates into one direction, starting with the independent variables as input unit values and finishing with the predicted target quantities as network output. In the given case study, attention is paid to

MLPs with only a single hidden layer. This is sufficient, since NNs are here intended as a de-trending method and do not need to capture any variation detail. Consequently, an increased model complexity through another hidden layer was considered unnecessary. The described structure of a typical multi-layer perceptron is additionally shown in figure 4.7.

Focusing on computation, the net input net_i of any unit i is calculated as a weighted average of the response from n_o preceding nodes j

$$\text{net}_i = \theta_i + \sum_{j=1}^{n_o} w_{ij} o_j \tag{4.20}$$

where w_{ij} denotes the weight of the corresponding connection and θ_i represents some node-specific bias term that may be interpreted as an intercept in linear regression. For convenience, θ_i is treated as an additional weight to a bias node with output value 1 and is henceforth included into the weight vector. The determined net input is subsequently transformed using a so-called activation function, for instance, the hyperbolic tangent (tanh):

$$o_i = f_{\tanh}(\text{net}_i) = \frac{e^{\text{net}_i} - e^{-\text{net}_i}}{e^{\text{net}_i} + e^{-\text{net}_i}} \tag{4.21}$$

Applying sigmoidal activation functions such as the tanh eventually introduces non-linearity into the chosen MLP network. These functions are favoured among a variety of unit transformation options and have positive influences on common training algorithms such as fast convergence and differentiability (Bishop 1995). In this study, all hidden units are activated by tanh, whereas input and output units remain untransformed. Using the identity function for activation of output units ensures that the prediction values can freely vary in real space $\mathbb{R}$. This is highly desirable, since alr-transformed clay and silt contents are continuous-valued target quantities with both, positive and negative expected output.

The number of hidden layers and units, their degree of connectivity through weighted links as well as the choice of appropriate activation functions are key

elements of the network architecture that must be defined by the user with respect to the studied problem. Once the structure is determined, the learning or training phase follows, being simply a gradual adaptation of the connection weights, which are often randomly chosen at the beginning. MLP-related learning is usually supervised referring to the presence of a training set with measured target quantities. Changing the weights is then based on differences between these known values and predicted network output measured by some error function (Riedmiller 1994). Minimising this error function to find the best NN parameter set is a non-linear optimisation problem for which several algorithms are at hand, such as gradient descent, (scaled) conjugate gradients, quasi-Newton methods or the Levenberg-Marquardt algorithm (Bishop 1995).

In the frame of this thesis, MLPs are trained using resilient back-propagation (Rprop) as proposed by Riedmiller and Braun (1993). The Rprop algorithm has been used in many soil-scientific applications (see Behrens et al. 2005; Zhao et al. 2009) and is said to be one of the most successful learning rules that are available in the area of NN modelling (Igel and Hüsken 2000). It is declared as a local adaptive learning scheme working in a supervised batch training mode. Local adaptation means that weight updates are exclusively based on weight-specific information. Batch training (or learning by epoch) refers to the decision that weights are adjusted only once per iteration, i. e. after the complete training set is processed (Rojas 1996).

From a computational point of view, the first stage of (resilient) back-propagation is to determine the derivative of an error function E with respect to each individual weight in the network (Riedmiller and Braun 1993). This can be achieved by applying the chain rule for partial derivatives leading to

$$\frac{\partial E}{\partial w_{ij}} = \frac{\partial E}{\partial o_i} \frac{\partial o_i}{\partial w_{ij}} \tag{4.22}$$

For mathematical details on how to proceed from equation 4.22 refer to Rumelhart et al. (1986), Riedmiller (1994) or Bishop (1995). Once the partial derivative is known for each network parameter (weights and biases), it can be used to fi-

nally minimise the error function by updating the weights. In its original version of error back-propagation, Rumelhart et al. (1986) minimised a sum-of-squares error function using the gradient descent method. Riedmiller and Braun (1993) improved this procedure by proposing to rely on the sign of the gradient information rather than on its absolute value for adjusting any weight. Formally, they introduced a parameter-specific update-value Δ_{ij}

$$\Delta_{ij}^{(t)} = \begin{cases} \eta^+ \cdot \Delta_{ij}^{(t-1)} & \text{if } \frac{\partial E}{\partial w_{ij}}^{(t-1)} \cdot \frac{\partial E}{\partial w_{ij}}^{(t)} > 0 \\ \eta^- \cdot \Delta_{ij}^{(t-1)} & \text{if } \frac{\partial E}{\partial w_{ij}}^{(t-1)} \cdot \frac{\partial E}{\partial w_{ij}}^{(t)} < 0 \\ \Delta_{ij}^{(t-1)} & \text{else} \end{cases} \tag{4.23}$$

where $0 < \eta^- < 1 < \eta^+$, index t tags the current iteration step and $\frac{\partial E}{\partial w_{ij}}^{(t)}$ denotes the gradient information summed over all observations of the training set. The increase (η^+) and decrease factor (η^-) are constantly set to 1.2 and 0.5, respectively, following recommendations given by Riedmiller and Braun (1993). Knowing the magnitude of the update-value, the actual weight change is computed from:

$$\Delta w_{ij}^{(t)} = \begin{cases} -\Delta_{ij}^{(t)} & \text{if } \frac{\partial E}{\partial w_{ij}}^{(t)} > 0 \\ +\Delta_{ij}^{(t)} & \text{if } \frac{\partial E}{\partial w_{ij}}^{(t)} < 0 \\ 0 & \text{else} \end{cases} \tag{4.24}$$

$$w_{ij}(t+1) = w_{ij}^{(t)} + \Delta w_{ij}^{(t)} \tag{4.25}$$

Applying the Rprop algorithm to a real-world problem using the mlp-function from the R-package RSNNS (Bergmeir and Benítez 2012), requires some choice with respect to three arbitrary parameters. The initial update-value Δ_0 at the beginning of the MLP training was set to 0.1, while the maximum weight-step size $\Delta_{\max}$ was limited to 30. Note, however, that both selections usually have only little influence on the training quality (Riedmiller 1994). The third quantity refers to some weight-decay term intended to prevent the MLP system from overfitting to peculiarities (noise) in the training data (see Bishop 1995). The latter is

harmful, because the over-trained network looses its generalisation (interpolation) ability, which is the actual aim of NNs/MLPs used in prediction problems. In the given implementation, the common sum-of-squares error function is augmented by another sum-of-squares regarding all weights and biases w_{ij} of the network

$$E = \sum_n \sum_i (z_i^n - o_i^n)^2 + 10^{-\alpha} \sum w_{ij}^2 \tag{4.26}$$

where indices n and i range over the set of patterns (observations) and output units, respectively. Variable z_i^n is the measured target quantity for a certain pattern and o_i^n represents the output calculated from trained weights with regards to the input vector of the same particular pattern. The exponent α defines to what extent the size of the weights are reduced by the penalty (for complexity) term and is set to 5 in the present application (Zell et al. 1998). Possible alternatives to regularisation using simple weight-decay are pruning algorithms or early stopping techniques (Bishop 1995).

Besides defining these Rprop-specific parameters, an exhaustive MLP application requires a few more decisions to make. One important practical issue is related to the network architecture and deals with the determination of an appropriate number of hidden units in each hidden layer. This decision considerably influences the complexity of the modelled MLP and must be taken with care to neither over-fit to the noise of the training data nor substantially miss any underlying process. As mentioned earlier, only one hidden layer is considered from start and its optimal number of units is then identified by trial and error. Another crucial concern that is tackled by testing addresses the question, when to stop the applied iterative Rprop training procedure. Thus, a set of MLPs with 5, 7, 9, 11, 13 and 15 hidden units was repeatedly (10 times) trained with 20, 40 or 60 iterations. Based on prediction errors calculated from an independent validation set, the network with the best performance was finally selected for prediction. Repeating the MLP training for each unit and iteration number combination is strongly advised, because the weight set in the network is randomly initialised between -2 and 2 prior to any training run.

The determination of a proper network architecture and a suitable iteration number based on trial and error depends on two distinct datasets. Thus, splitting the calibration data into training and validation sets (see chapter 4.5.2) is an important prerequisite to NN modelling. Another important pre-processing step is the (co-)variable selection, which is based here on correlation analysis as outlined in the exploratory data analysis section 4.4.2. Once the final set of network input data is identified, all continuous-valued predictors are standardised, i. e. re-calculated to have zero mean and unit variance. Normalising the independent variables is especially helpful in cases where initial weights are randomly set, since it eliminates the difficulties arising from different variable ranges. Note that mean and standard deviation used for normalisation are determined from training data and must be used as well to standardise validation and test data. Categorical input is transferred to the NN system as dummy variables using 1-of-c coding, whereas target quantities entered the network without any changes.

Although being a data-driven approach that almost entirely focuses on prediction, the fully trained network is also briefly analysed from an explanatory perspective. In order to study the contribution of each independent variable on the network outputs, MLP weights are examined in two different ways. First, a neural interpretation diagram (NID) is constructed in which the line width of any connection is proportional to its associated weight, while color denotes the direction of the weights (Özesmi and Özesmi 1999). Second, the relative importance of any input variable is computed by partitioning the connection weights according to Garson's algorithm (Garson 1991). Olden and Jackson (2002) and Gevrey et al. (2003) provide some comprehensive reviews on methods to gain insight into neural network models applied to ecological problems.

More comprehensive descriptions of the theory behind neural networks can be found, for instance, in the textbooks by Bishop (1995) or Rojas (1996). For practical tips with respect to application refer to the online FAQ-collection by Sarle (2002). Implementation details are provided by the manual related to the Stuttgart Neural Network Simulator (SNNS) (Zell et al. 1998), which has been used here in its R-adaptation (Bergmeir and Benítez 2012). Note, however, that the presented neural network terminology can also be expressed in statistical

terms. Sarle (1994) provides the translations and shows that multi-layer perceptrons are equivalent to multivariate, multiple non-linear regressions. Moreover, traditional back-propagation methods can also be seen as variations of maximum likelihood estimation.

4.6.3 Inverse distance weighting

Since its application is comparatively simple, inverse distance weighting (IDW) is a very popular and traditional spatial interpolation technique (Shepard 1968). It is exact and strictly reflects the idea of Tobler's first law of geography that "everything is related to everything else, but near things are more related than distant things" (Tobler 1970, p. 236) by inversely relating the similarity of attribute values to the distance between point pairs. As with many (spatial) prediction methods, IDW is based on a weighted average of measured values $z(\mathbf{s}_i)$

$$\hat{z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i^{IDW}(\mathbf{s}_0) z(\mathbf{s}_i) \tag{4.27}$$

with λ_i^{IDW} being the weight for an adjacent point at location i and n representing the number of samples involved. All weights together necessarily sum to one, while each individual coefficient is determined by

$$\lambda_i^{IDW}(\mathbf{s}_0) = \frac{1/d^{\beta}(\mathbf{s}_0, \mathbf{s}_i)}{\sum_{i=1}^{n} (1/d^{\beta}(\mathbf{s}_0, \mathbf{s}_i))} \tag{4.28}$$

where $d(\mathbf{s}_0, \mathbf{s}_i)$ is the distance between an unknown point and a surveyed location (Hengl 2009). The calculation of weights does not depend on stochastic determinants but only on the arbitrary choice of power coefficient b and considered neighbouring radius. In the present thesis, interpolation using IDW is based on all observations and weighting by inverse squared distances ($\beta = 2$), which is a common choice in environmental modelling reported, for instance, in the textbooks by Kanevski and Maignan (2004, p. 54) or Webster and Oliver (2007, p. 40).

4.6.4 Geostatistical techniques

Due to inaccurate measurements and limited process knowledge the modelling of spatial data is often based on stochastic considerations. Within the field of geostatistics, estimation, therefore, follows the idea that the value of a variable z at location with coordinates vector $\mathbf{s}$ is a realisation of a random function $Z(\mathbf{s})$

$$Z(\mathbf{s}) = m(\mathbf{s}) + e(\mathbf{s}) \tag{4.29}$$

which can be decomposed as a deterministic mean structure $m(\mathbf{s})$ and a spatially correlated error process $e(\mathbf{s})$. The latter term denotes a random field with zero mean and covariance function $\sigma(\mathbf{h})$ or variogram $2\gamma(\mathbf{h})$ depending only on the lag vector $\mathbf{h}$ between two points $\mathbf{s}_i$ and $\mathbf{s}_j$ (Christensen 1991; Cressie 1993). Since second-order properties of $e(\mathbf{s})$ are rarely known a priori, they have to be modelled in terms of some characteristic parameters during interpolation. This essential step of estimating a continuous variogram can only be achieved, if the stochastic process under study satisfies Matheron's intrinsic hypothesis (Matheron 1973). It is a relief from more restrictive second-order (or weak) stationarity assuming the mean to be constant for small increments $|\mathbf{h}|$ only, which is more realistic with regards to most environmental applications. Additionally, Matheron introduced the variogram as a substitute for the covariance function to mathematically describe the spatial dependence between point pairs (Webster and Oliver 2007). If the intrinsic hypothesis holds, variogram parameter estimates are determined from the one available realisation of $Z(\mathbf{s})$, that is, the given set of spatially indexed measurement values. Matheron (1971) calls the sample of observations a regionalized variable and very similar to IDW it is this set of actual quantities that geostatistical prediction is greatly based upon.

Geostatistical prediction is concerned with the estimation at unknown locations, a procedure known as kriging in the setting discussed below. Rising from ore reserve estimation, kriging was developed rather in isolation from mainstream statistics starting with D. G. Krige and G. Matheron in the 1950s and 1960s (see Cressie 1990). Nevertheless, several authors (e.g. Christensen 1991; Stein 1999)

have shown that kriging provides identical estimates for intrinsically stationary Gaussian processes as best linear unbiased prediction (BLUP), a well-defined statistical concept derived from linear (mixed) model theory by Goldberger (1962) and Robinson (1991). Assuming model 4.29, the kriging estimate or BLUP at some unobserved point $\mathbf{s}_0$ is

$$\hat{z}(\mathbf{s}_0) = \sum_{i=1}^{n} \lambda_i z(\mathbf{s}_i) = \boldsymbol{\lambda}^\top \mathbf{z} \tag{4.30}$$

with λ_i being the kriging weight, $z(\mathbf{s}_i)$ representing the measured target quantity of a neighbouring site i and mean squared prediction error (kriging variance)

$$\sigma_k^2(\mathbf{s}_0) = 2\boldsymbol{\lambda}^\top \boldsymbol{\gamma}_0 - \boldsymbol{\lambda}^\top \boldsymbol{\Gamma} \boldsymbol{\lambda} \tag{4.31}$$

where $\boldsymbol{\gamma}_0 = (\gamma(\mathbf{s}_0 - \mathbf{s}_1), \gamma(\mathbf{s}_0 - \mathbf{s}_2), \ldots, \gamma(\mathbf{s}_0 - \mathbf{s}_n))^\top$ and $\boldsymbol{\Gamma}$ is a symmetric $n \times n$ matrix with (semi)variances between any two points $\gamma(\mathbf{s}_i - \mathbf{s}_j), i, j = 1, \ldots, n$. Minimising 4.31 subject to the unbiasedness constraint, i.e. to ensure the expected error to be zero, requires the introduction of Lagrange multipliers $\boldsymbol{\psi}$ and the optimisation criterion 4.31 changes to:

$$\sigma_k^2(\mathbf{s}_0) = 2\boldsymbol{\lambda}^\top \boldsymbol{\gamma}_0 - \boldsymbol{\lambda}^\top \boldsymbol{\Gamma} \boldsymbol{\lambda} - 2\boldsymbol{\psi}^\top (\mathbf{X}^\top \boldsymbol{\lambda} - \mathbf{x}_0) \tag{4.32}$$

Differentiating 4.32 with respect to Lagrange multipliers $\boldsymbol{\psi}$ and kriging weights $\boldsymbol{\lambda}$ yields the general formulation of the kriging system in matrix notation after equating to zero

$$\begin{bmatrix} \boldsymbol{\Gamma} & \mathbf{X} \\ \mathbf{X}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ \boldsymbol{\psi} \end{bmatrix} = \begin{bmatrix} \boldsymbol{\gamma}_0 \\ \mathbf{x}_0 \end{bmatrix} \tag{4.33}$$

with design matrix $\mathbf{X}$ and zero matrix $\mathbf{0}$ (see Christensen 1991; Minasny and McBratney 2007). Inverting the left-hand side of equation 4.33 and solving for $\boldsymbol{\lambda}$ gives the optimal set of weights required for the BLUP (4.30). The exact

formula to obtain $\boldsymbol{\lambda}$ changes according to the kriging variant used for spatial prediction. However, all kriging approaches have in common that calculations are heavily dependent on variogram models that are usually unknown. Thus, the key practical issue applying kriging techniques in its traditional sense is to estimate the parameters of any variogram model from sample values. Go back to chapter 4.4.3 on page 50 for details on variogram analysis in the univariate case. The following sections describe the specific kriging variants that are relevant in the frame of this dissertation. Note that their presentation consistently uses variogram terminology and focuses on prediction based on point-support. In addition, only isotropic processes are considered, i. e. the second-order properties are assumed to solely depend on distance, denoted by h, omitting any possible directional influences.

For a detailed derivation of the kriging equations refer to the textbooks by Webster and Oliver (2007), Chiles and Delfiner (2012) or the more mathematically demanding reference work from Cressie (1993). Christensen (1991) gives insight into spatial data modelling by explicitly relating the kriging system to the concept of linear mixed models and BLUP theory. Alternative textbooks that focus on linear mixed model frameworks for geostatistical prediction are provided by Stein (1999) and Diggle and Ribeiro Jr. (2007).

4.6.4.1 Ordinary kriging

Ordinary kriging (OK) is the most frequently used member of the kriging family and a special case of the random function decomposition defined in 4.29. Following the notation introduced in section 4.6.4, the OK model is

$$Z(\mathbf{s}) = \mu + e(\mathbf{s}) \tag{4.34}$$

where μ represents an unknown spatial mean that is assumed to be constant throughout the entire study area or some local neighbourhood. With respect to the kriging system (4.33), $\mathbf{X}$ reduces to a vector of ones and $\mathbf{x}_0$ to 1:

$$\begin{bmatrix} \boldsymbol{\Gamma} & \mathbf{1} \\ \mathbf{1}^\top & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda} \\ \psi \end{bmatrix} = \begin{bmatrix} \boldsymbol{\gamma}_0 \\ 1 \end{bmatrix} \tag{4.35}$$

Solving equations 4.35 for $\boldsymbol{\lambda}$ yields the best linear unbiased predictor of $Z(\mathbf{s}_0)$

$$\hat{z}(\mathbf{s}_0) = \boldsymbol{\lambda}^\top \mathbf{z} = \boldsymbol{\gamma}_0^\top \boldsymbol{\Gamma}^{-1} \mathbf{z} + [1 - \boldsymbol{\gamma}_0^\top \boldsymbol{\Gamma}^{-1} \mathbf{1}](\mathbf{1}^\top \boldsymbol{\Gamma}^{-1} \mathbf{1})^{-1} \mathbf{1}^\top \boldsymbol{\Gamma}^{-1} \mathbf{z} \tag{4.36}$$

satisfying $\sum_{i=1}^{n} \lambda_i = 1$ to guarantee unbiasedness. The corresponding OK prediction variance is estimated using:

$$\sigma_{ok}^2(\mathbf{s}_0) = \boldsymbol{\gamma}_0^\top \boldsymbol{\Gamma}^{-1} \boldsymbol{\gamma}_0 - [1 - \boldsymbol{\gamma}_0^\top \boldsymbol{\Gamma}^{-1} \mathbf{1}](\mathbf{1}^\top \boldsymbol{\Gamma}^{-1} \mathbf{1})^{-1} [1 - \boldsymbol{\gamma}_0^\top \boldsymbol{\Gamma}^{-1} \mathbf{1}]^\top \tag{4.37}$$

The possibility to calculate prediction variances at each unsampled location provides a rough measure of uncertainty related to every particular estimate or grid cell, respectively. This is one of the major advantages of (ordinary) kriging over deterministic interpolation techniques. Linear regression or inverse distance weighting, which were formerly discussed, do not come with such an internal quantification of model quality.

The given derivation of kriging weights and prediction variance again underlines the outstanding importance of the variogram in geostatistics. In the frame of this thesis, OK prediction of soil textural fractions at the landscape scale make use of variogram models obtained from variography as part of the exploratory spatial data analysis described in chapter 4.4.3.

4.6.4.2 Cokriging

Cokriging is an interpolation technique that extends the kriging methods to the multivariate case. In most applications cokriging aims at predicting just a single target quantity, called the principal or primary variable, which is modelled using a set of more densely sampled, but not continuously available, correlated attributes (Goovaerts 1997; Webster and Oliver 2007). However, in the frame of this thesis,

cokriging is considered in a slightly different way to simultaneously predict a vector of two variables which are equally sampled. The corresponding linear model in its ordinary form is

$$\mathbf{Z}(\mathbf{s}) = \boldsymbol{\mu} + \mathbf{e}(\mathbf{s}) \tag{4.38}$$

where $\boldsymbol{\mu}$ represents a vector of unknown, overall spatial means and $\mathbf{e}(\mathbf{s})$ is an intrinsic random vector with zero mean. Following the nonnegative-definiteness criterion given by Ver Hoef and Cressie (1993) and using the matrix formulation from section 4.6.4, the ordinary cokriging system for the bivariate case can be derived as

$$\begin{bmatrix} \boldsymbol{\Gamma}_{11} & \boldsymbol{\Gamma}_{12} & \mathbf{1}_1 & \mathbf{0} \\ \boldsymbol{\Gamma}_{21} & \boldsymbol{\Gamma}_{22} & \mathbf{0} & \mathbf{1}_2 \\ \mathbf{1}_1^\top & \mathbf{0}^\top & 0 & 0 \\ \mathbf{0}^\top & \mathbf{1}_2^\top & 0 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{\lambda}_1 \\ \boldsymbol{\lambda}_2 \\ \psi_1 \\ \psi_2 \end{bmatrix} = \begin{bmatrix} \mathbf{b}_{11} & \mathbf{b}_{12} \\ \mathbf{b}_{21} & \mathbf{b}_{22} \\ 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{4.39}$$

where $\mathbf{b}_{11} = (\gamma_{11}(\mathbf{s}_0 - \mathbf{s}_1), \gamma_{11}(\mathbf{s}_0 - \mathbf{s}_2), \ldots, \gamma_{11}(\mathbf{s}_0 - \mathbf{s}_n))^\top$, $\mathbf{b}_{22} = (\gamma_{22}(\mathbf{s}_0 - \mathbf{s}_1), \gamma_{22}(\mathbf{s}_0 - \mathbf{s}_2), \ldots, \gamma_{22}(\mathbf{s}_0 - \mathbf{s}_n))^\top$, $\mathbf{b}_{12} = \mathbf{b}_{21} = (\gamma_{12}(\mathbf{s}_0 - \mathbf{s}_1), \gamma_{12}(\mathbf{s}_0 - \mathbf{s}_2), \ldots, \gamma_{12}(\mathbf{s}_0 - \mathbf{s}_n))^\top$ and $\boldsymbol{\Gamma}_{uv}$ are symmetric $n \times n$ matrices representing direct ($u = v = 1$, $u = v = 2$) or cross ($u \neq v$) semivariances between any two points $\gamma_{uv}(\mathbf{s}_i - \mathbf{s}_j), i, j = 1, \ldots, n$. More concisely, the cokriging equations (4.39) in terms of the cross-variogram are written as

$$\begin{bmatrix} \mathbf{G} & \mathbf{X} \\ \mathbf{X}^\top & \mathbf{0} \end{bmatrix} \begin{bmatrix} \boldsymbol{\Lambda} \\ \boldsymbol{\Psi} \end{bmatrix} = \begin{bmatrix} \mathbf{b} \\ \mathbf{x}_c \end{bmatrix} \tag{4.40}$$

from which the best linear unbiased predictor is received assuming $\mathbf{G}$ to be invertible and solving for the weights $\boldsymbol{\Lambda}$:

$$\hat{\mathbf{z}}(\mathbf{s}_0) = \boldsymbol{\Lambda}^\top \mathbf{z} = \mathbf{b}^\top \mathbf{G}^{-1} \mathbf{z} + [\mathbf{x}_c - \mathbf{b}^\top \mathbf{G}^{-1} \mathbf{X}](\mathbf{X}^\top \mathbf{G}^{-1} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{G}^{-1} \mathbf{z} \tag{4.41}$$

The 2×2 mean squared prediction error (kriging variance) matrix is accordingly defined as:

$$\boldsymbol{\sigma}^2_{ck}(\mathbf{s}_0) = \mathbf{b}^\top \mathbf{G}^{-1}\mathbf{b} - [\mathbf{x}_c - \mathbf{b}^\top \mathbf{G}^{-1}\mathbf{X}](\mathbf{X}^\top \mathbf{G}^{-1}\mathbf{X})^{-1}[\mathbf{x}_c - \mathbf{b}^\top \mathbf{G}^{-1}\mathbf{X}]^\top \quad (4.42)$$

The interpolation algorithm of the cokriging predictor requires not only direct variograms of all involved target quantities but also cross-(semi)variances between each variable pair for any lag distance h. These continuous direct and cross-variogram models reflecting the spatial dependence within and between two or more regionalised variables are commonly addressed by assuming a linear model of coregionalisation (LMC). Formally, the LMC is a linear combination of one particular set of L authorised variogram functions $g_l(h)$

$$\mathbf{G}(h) = \sum_{l=1}^{L} \mathbf{B}_l g_l(h) \quad (4.43)$$

with $\mathbf{B}_l$ being the 2×2 (in the bivariate case) coregionalisation matrix that summarises the partial sills of some corresponding variogram model $g_l(h)$ (Goovaerts 1997). In this thesis, only spherical variogram models (4.5) are considered as introduced in chapter 4.4.3 including a nugget effect.

From a practical point of view, as the LMC is unknown, it is established from multivariate spatial data. Whereas experimental direct variograms are obtained from sample data following Matheron's method-of-moments estimator (4.4), calculating empirical cross-variograms is more difficult as discussed, for instance, by Papritz et al. (1993) or Cressie and Wikle (1998). Different estimators are proposed, from which the classic (covariance-based) cross-variogram is used

$$\hat{\gamma}_{uv}(\mathbf{h}) = \frac{1}{2} \cdot \frac{1}{N(\mathbf{h})} \sum_{i=1}^{N(\mathbf{h})} [z_u(\mathbf{s}_i) - z_u(\mathbf{s}_i + \mathbf{h})][z_v(\mathbf{s}_i) - z_v(\mathbf{s}_i + \mathbf{h})] \quad (4.44)$$

for $u \neq v$ where $z_u(\mathbf{s})$, $z_v(\mathbf{s})$ and $z_u(\mathbf{s}+\mathbf{h})$, $z_v(\mathbf{s}+\mathbf{h})$ are measured values of Z at locations $\mathbf{s}$ and $\mathbf{s}+\mathbf{h}$ with $N(\mathbf{h})$ being the number of point pairs separated by a particular lag $\mathbf{h}$. Restricting to isotropic processes and the bivariate case, $\mathbf{h}$ reduces to a scalar and $u, v = 1, 2$. Note that this type of estimator can only be used if the target quantities are measured at coincident sites.

The fitting procedure to end up with continuous variogram functions to calculate $\mathbf{G}$ and $\mathbf{b}$ in equation 4.41 for prediction purposes is divided into three steps. First, empirical direct and cross-variograms are inspected for identical range and model structure as well as an initial guess on the nugget and partial sill coefficients. Second, optimal estimates of the nugget and partial sills are calculated based on a weighted least squares ($\mathrm{N/h^2}$) approach. Third, the eigenvalues of the resulting sill matrices are checked for being non-negative. The latter step is adequate to verify, whether the built LMC is permissible and, thus, ensures the positive definiteness of the cokriging system. Note, however, that the described procedure is often considered as being sub-optimal (see Goovaerts 1997) and can easily become cumbersome if the number of coregionalised variables increases. Thus, more sophisticated ways of fitting a LMC directly addressing the positive definiteness condition are available. For instance, Goulard and Voltz (1992) suggest an iterative least-squares like technique, Pelletier et al. (2004) compares different weighted and generalised least squares procedures and Marchant and Lark (2007) proposes a restricted maximum likelihood based approach. However, the simple three-step strategy appeared sufficient in the given bivariate case studies and more advanced fitting methods were disregarded.

Ordinary cokriging as previously presented was applied merely on alr-transformed soil textural fractions and their residuals. Pearson's correlation coefficients were considered to justify the appropriateness of using this multivariate extension of kriging. In cases where equally sampled variables are hardly correlated, the differences between kriging and cokriging results are likely to be negligible (Webster and Oliver 2007).

For an exhaustive matrix formulation of cokriging, refer to Myers (1982) or Ver Hoef and Cressie (1993). Goovaerts (1997) comprehensively presents the theory

on LMC. A very recent review written by Fanshawe and Diggle (2012) focuses on bivariate models, discussing alternative strategies to the common LMC approach such as kernel convolution techniques, latent dimensions or copulas.

4.6.5 Hybrid methods

The previous chapters introduced linear regression (4.6.1) and neural networks (4.6.2) as non-spatial tools for predictions based on explanatory factors correlated with the target variables and available for each pixel of the entire study area. However, possible auto- and cross-correlations among the modelled quantities are neglected by these techniques. On the contrary, (ordinary) geostatistical methods (4.6.4) were described that explicitly focus on spatial dependence within and between regionalised variables, while omitting any deterministic structure by assuming unknown and fixed spatial means. Since soil variation often cannot be fully addressed by either deterministic or stochastic components, hybrid methods are applied, combining the benefits of those two groups of interpolation techniques. Note that ordinary cokriging is not classified as hybrid, as it is used for simultaneous kriging of two target variables rather than modelling an undersampled single quantity in terms of other correlated properties.

4.6.5.1 Regression (co)kriging

Regression kriging (RK) is an univariate interpolation technique that additively mixes the regression of some target quantity on a set of independent variables with (ordinary) kriging of the regression residuals (Odeh et al. 1995; Hengl et al. 2004). Accordingly, some RK point estimate at $\mathbf{s}_0$ is formally described by

$$\hat{z}(\mathbf{s}_0) = \hat{m}(\mathbf{s}_0) + \hat{e}(\mathbf{s}_0) \tag{4.45}$$

with $\hat{m}(\mathbf{s}_0)$ being the fitted trend estimate and $\hat{e}(\mathbf{s}_0)$ representing the kriged regression residuals. Note the similarity between equation 4.45 and the general

decomposition of the stochastic process (4.29) leading to the BLUP derivation from linear mixed model theory. This is the reason why some authors state that spatial prediction based on MLR and the kriging variants are nothing but special cases of the BLUP (see Christensen 1991; Stein 1999), which is in matrix notation formulated as

$$\hat{z}(\mathbf{s}_0) = \mathbf{x}_0^\top \hat{\boldsymbol{\beta}}_{GLS} + \boldsymbol{\lambda}_0^\top (\mathbf{z} - \mathbf{X}\hat{\boldsymbol{\beta}}_{GLS}) \tag{4.46}$$

where $\mathbf{X}$ is the design matrix of predictors at measured locations, $\mathbf{x}_0$ is the vector of predictors at the new point $\mathbf{s}_0$ and $(\mathbf{z}-\mathbf{X}\hat{\boldsymbol{\beta}}_{GLS})$ representing the column vector of GLS residuals. The regression coefficients $\hat{\boldsymbol{\beta}}_{GLS}$ and kriging weights $\boldsymbol{\lambda}_0$ are estimated independently following the principles outlined in section 4.6.1 and 4.6.4, respectively. Alternatively, predictions can be made directly by solving for fixed and random effects in conjunction following methods known as universal kriging (UK) or kriging with external drift (KED) that are mathematically equivalent to RK (see Hengl et al. 2007).

Extending RK formally to the multivariate case is straight forward. Instead of repeating mathematical derivations, required practical steps to perform regression cokriging (RCOK) as applied in the context of this thesis are summarised in 4.8.

Practical issues related to stepwise regression modelling and ordinary cokriging can be understood from sections 4.6.1 and 4.6.4.2, respectively. Note, however, that GLS estimates only occur in the regression part or rather trend analysis. No recalculations of the auto- and cross-variogram parameters were considered based on GLS residuals. Such an iteratively re-weighted GLS procedure was suggested by Schabenberger and Gotway (2004) to overcome the RK drawback that estimated empirical variances from OLS residuals are biased especially with regard to larger lags (Cressie 1993; Lark et al. 2006). However, among others, Kitanidis (1993) found out that there is often not a big difference in practice between the iterative procedure and OLS-based estimation of second-order properties (Hengl 2009). Accordingly, any iteratively re-weighting was disregarded in the RK and RCOK applications of this dissertation.

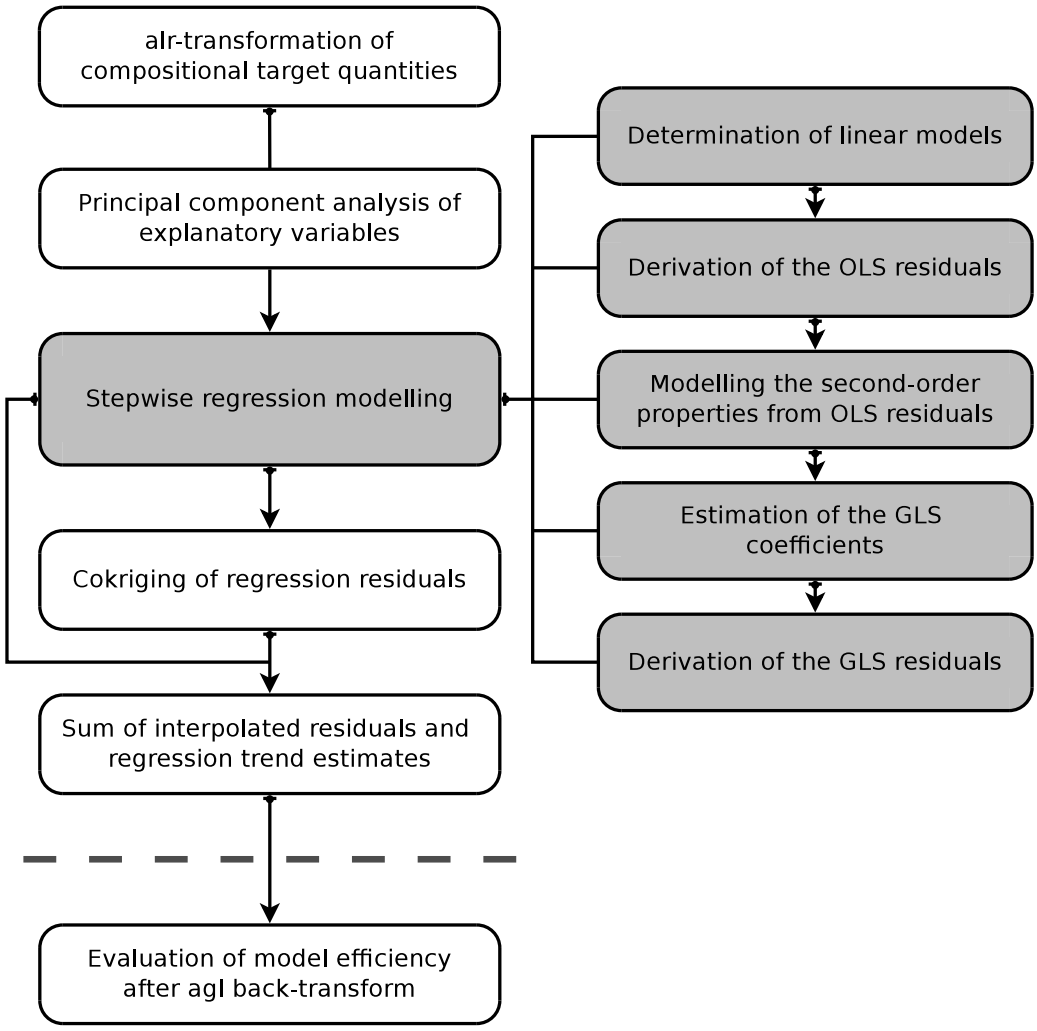


Figure 4.8: Overview of the regression cokriging framework

With respect to digital soil mapping at the field scale, regression cokriging is applied in a global setting. RCOK was favoured over other hybrid techniques, because it allows for separate investigation of the regression modelling part and the geostatistical interpolation component. At the landscape scale, uni- and multivariate versions of RK were used as a benchmark model to evaluate potentially more accurate neural network approaches for substituting the regression part of the analysis as described in the subsequent section.

4.6.5.2 Neural network residual cokriging

A big advantage of the hybrid regression (co)kriging method is the independent treatment of deterministic and random components. This flexibility also allows for the incorporation of more complex techniques that open up the spatial modelling procedures beyond the traditional linear case. Such an approach is shown in figure 4.9 illustrating the steps needed to combine an artificial neural network (multi-layer perceptron) with cokriging for soil-related mapping. Remarkable issues of the neural network training phase are shown in gray with individual decisions added in brackets. For practical details on the neural network part consider section 4.6.2, while subsection 4.6.4.2 refers to cokriging theory. Since the required amount of measured data from which to train any multi-layer perceptron is rather large, neural network residual cokriging (NNRCK) is only applied to the Rio di Costara catchment.

4.7 Validation and model comparison

Validation provides information on how well a particular model performed in practice and is a crucial task associated with any digital soil mapping project. This quality check can be based on either graphical analysis or numerical indices quantifying the error of the evaluated model. A frequently used measure of prediction accuracy is the root mean squared error (RMSE) given by

$$\mathrm{RMSE} = \left[\frac{1}{n} \sum_{i=1}^{n} (\hat{z}(\mathbf{s}_i) - z(\mathbf{s}_i))^2 \right]^{\frac{1}{2}} \tag{4.47}$$

with n being the number of observations in the test set, $\hat{z}(\mathbf{s}_i)$ representing the model estimate at position $\mathbf{s}_i$ and $z(\mathbf{s}_i)$ denoting the corresponding true value. In contrast to many other deviation-based statistics often considered for validation purposes (Bellocchi et al. 2010), RMSE equally judges positive and negative model errors. By taking the square root, it remains in the same dimension as the

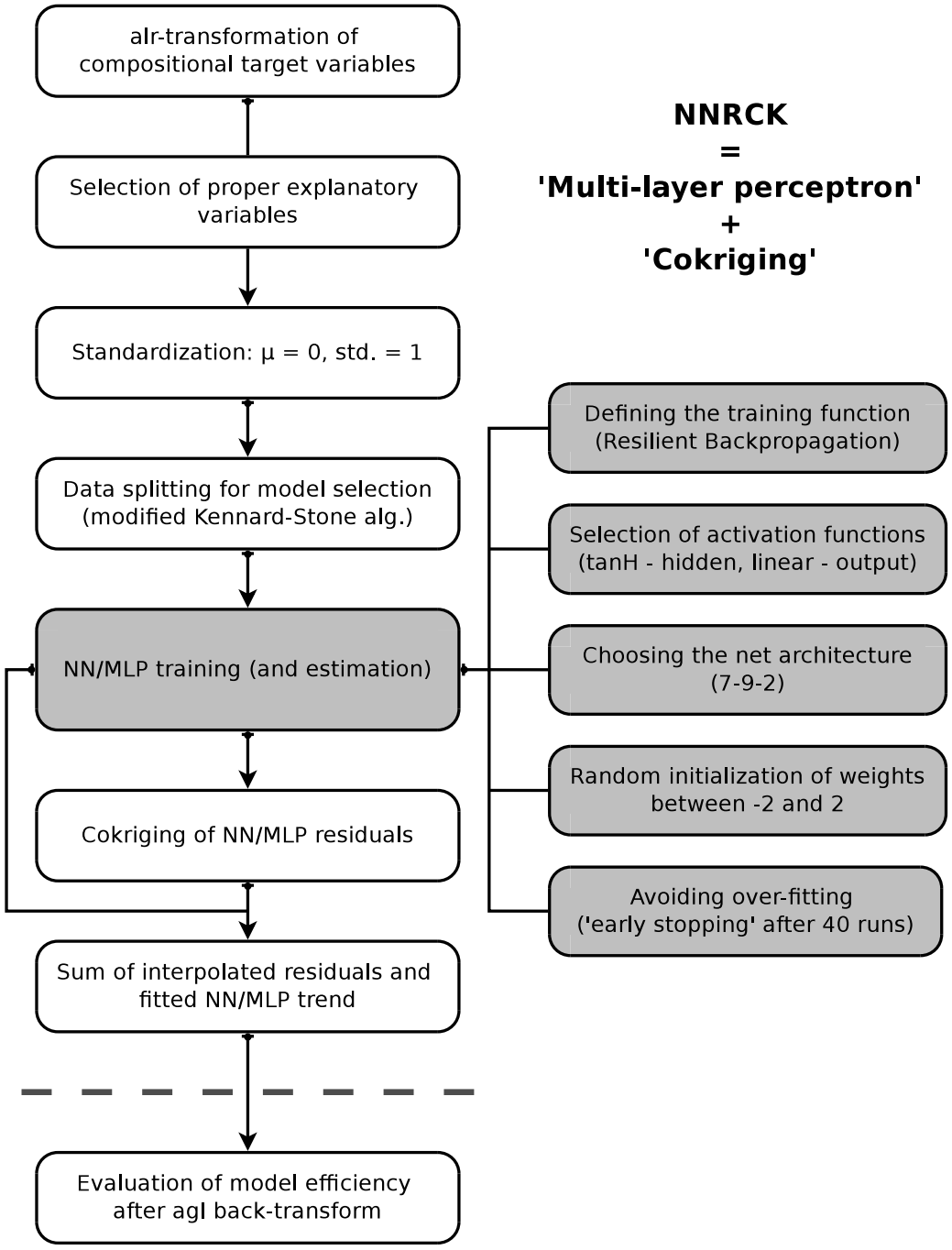


Figure 4.9: Overview of the neural network residual cokriging framework

studied target variable, which makes interpretation quite intuitive. Representing a measure of distance between predictions and measurements, it is inferred that smaller RMSE values point to better model performances. Similarly widespread association-based validation measures like coefficients of determination from regressions between estimated and observed values are also considered in the frame of this thesis. Note, however, that several authors discussed the drawbacks of model assessment with regards to this particular category (e.g. Mayer et al. 1994; Kobayashi and Salam 2000; Gauch et al. 2003). As a third numerical evaluation quantity, model efficiency (EF) is computed

$$\mathrm{EF} = 1 - \frac{\sum_{i=1}^{n}(z(\mathbf{s}_i) - \hat{z}(\mathbf{s}_i))^2}{\sum_{i=1}^{n}(z(\mathbf{s}_i) - \bar{z}(\mathbf{s}_i))^2} \tag{4.48}$$

where $\bar{z}(\mathbf{s}_i)$ is the mean of the measurement values considered for validation. EF can be interpreted as the explained variance level of the inspected model and is also known as Nash-Sutcliffe coefficient from hydrological modelling applications.

RMSE, EF as well as univariate summary statistics of model residuals basically treat the target variables as independent quantities. However, soil textural classes are compositional data and were modelled in a multivariate log-ratio approach. To be consistent, a combined quality estimate is added to the model evaluation procedure of this work. The standardised residual sum-of-squares (STRESS) as described by Martín-Fernández et al. (2001) and applied, for instance, in Lark and Bishop (2007) and Ward and Mueller (2012) is defined as

$$\mathrm{STRESS} = \left[\frac{\sum_{i<j}(\delta_{ij} - \delta_{ij}^*)^2}{\sum_{i<j}(\delta_{ij})^2}\right]^{\frac{1}{2}} \tag{4.49}$$

where δ_{ij} and δ_{ij}^* are Aitchison distances of two compositions i and j regarding observed and estimated (*) particle-size fractions, respectively. From this definition, STRESS can be interpreted as the overall similarity between multivariate predictions and measurements. Accordingly, lower STRESS values refer to more accurate prediction models (Lark and Bishop 2007).

In addition to common average error metrics, Shapiro-Wilk tests for normality and global Moran's I statistics to verify the absence of remaining spatial correlation are calculated based on the model residuals. To investigate model performances with respect to individual samples, bubble plots are created to unravel any systematic over- and underestimation in the study area. Note that hypothesis testing and the detection of (spatial) error clustering and outliers is only done for the final, hybrid interpolation techniques, not for applied reference methods.

In order to successfully validate any model, it is at best tested against data not used at any stage of the model calibration process (Snee 1977; McBratney et al. 2003; Esbensen and Geladi 2010). Such an independent set of reference data was collected at the landscape scale, but was not available for fields 21 and 33 due to time and budget limitations. As a consequence, the generalisation ability of any field-related prediction methods is examined using leave-one-out cross-validation (LOOCV). Following the LOOCV procedure, one sample at a time is removed from the model and the target quantities are estimated for this particular point. The leave-one-out re-sampling approach is proceeded until each measured location was omitted once and average validation measures can finally be determined (see Davis 1987; Hengl 2009).

Up to this point, performances of spatial prediction methods were evaluated individually. However, often more than one interpolation technique is applied to the same problem and the user is supposed to find a way to choose between competing methods. In this thesis, model comparison starts with simple visual inspection of selected prediction maps presenting the output of each model built in the corresponding case studies. This rather qualitative approach is then supplemented by measurements of spatial agreement between (soil texture) maps which are computed based on Cohen's κ statistics (see Sterlacchini et al. 2011). In addition, distributions of the residuals from the various methods are compared and overall error measures as defined for validation purposes are considered to select the best model at hand. In a last step, all sites are visited one by one simply counting best predictions in terms of absolute errors. A likewise procedure is applied, for instance, by Vašát et al. (2013).

4.8 Web-based delivery of digital soil mapping data

This section provides a brief introduction into the free and open-source based geoportal solution deployed with regards to the international research project CLIMB. As a major outcome of work package 2, the CLIMB Geoportal is primarily intended for storing and providing spatially distributed data about the current state and future changes of the hydrological conditions within the seven CLIMB test sites around the Mediterranean (see http://www.climb-fp7.eu). Besides, it also disseminates the final soil maps produced in the frame of this thesis in forms of GIS-related web services. Thus, the presented web-platform can be seen as the final step of the methodological workflow of this thesis providing long-term visibility and access to the produced soil mapping outputs.

The above-mentioned (geo)services are supposed to meet certain standards and specifications in order to properly operate inside a significant number of clients. Well-established standards and GIS-related specifications on spatial data, services and their describing meta information are defined by the Open Geospatial Consortium (OGC) and the ISO Technical Committee 211 (see Mitchell 2005; Nolde et al. 2010). In Europe, another important framework is given by the INSPIRE directive 2007/2/EC of the European Parliament aiming at a continental spatial data infrastructure (EC 2007). As a major component, this directive explicitly claims the dissemination of spatial information in forms of standardised web services that are mainly grouped into view, download and discovery services. The latter are meant for searching and querying spatial resources, while view services provide an image of some mapped geographic layer and download services grant access to its underlying data. The steps required to create INSPIRE-compliant (view) services are defined more precisely in the corresponding regulation document (EC 2009) and its technical guidance for implementation (EC 2011). There are also additional guidance documents on national level, e.g. the ones from the GDI-DE coordination unit in Germany (GDI-DE 2011). The same applies to the standard-compliant collection of metadata which is thoroughly specified, for instance, in EC (2008), EC (2013) and GDI-DE (2008).

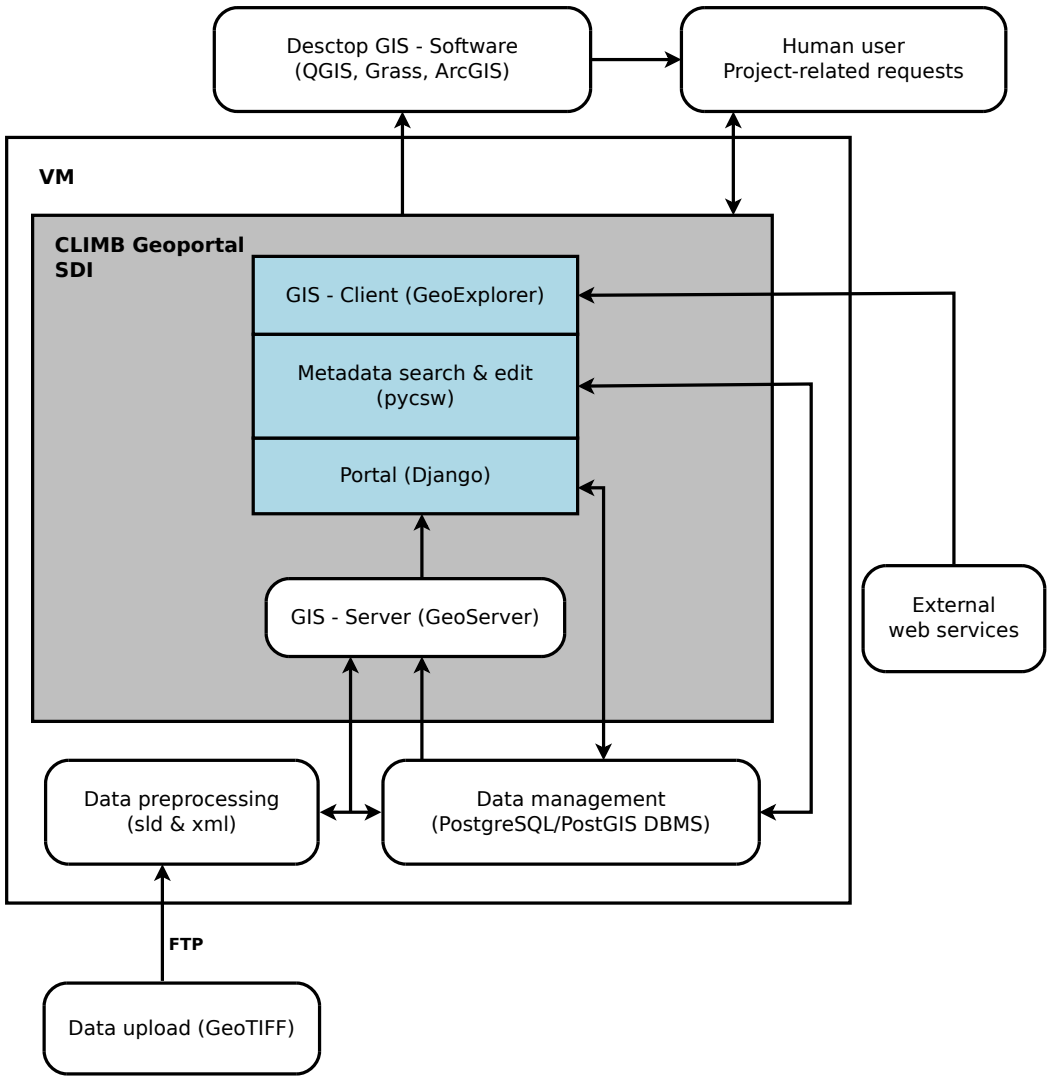


Figure 4.10: Architecture of the CLIMB-specific SDI with geoportal solution

Geoportals and their consolidated web services are key elements of spatial data infrastructures (SDIs), since they virtually represent their connection to the outside world. The CLIMB-SDI clearly focuses on the technical implementation issues of any SDI definition (see GSDI 2004), namely the distribution, documentation and visualisation of spatial data, whereas legal and organisational aspects are of minor interest. Figure 4.10 provides an overview of the central (software) components used to set-up the CLIMB-SDI and informs about their interactions.

The implementation of the CLIMB Geoportal is based on version 2.0c5 of the open-source geospatial content management system GeoNode (for details refer to http://geonode.org). It mainly consists of the Python web development framework Django and includes a GeoServer instance for providing OGC-compliant web services. In addition, GeoNode comes with an OGC CSW server implementation (pycsw) as well as a built-in WebGIS-client (GeoExplorer). PostgreSQL enhanced by PostGIS in versions 9.2.1/2.0.1 serves as database backend for relevant information on maps, layers and other elements of the final portal product. Note that soil mapping results in raster-format are stored file-based as GeoTIFFs, since the used GeoServer instance could not by default work on PostGIS raster. Nevertheless, GeoServer is an essential SDI component responsible for the creation of view and download services. In the context of view services, the OGC Web Map Service (WMS) standard, eventually defined in ISO 19128, is used in its current version 1.3.0. A download option is provided by means of a Web Coverage Service (WCS) which is, however, only available to registered users.

Once the portal has been successfully configured, adding a new resource requires the preparation and upload of up to three different files. Alongside the image itself, describing information can be provided through an additional XML-document containing metadata elements that follow the ISO 19115/19139 standards. Furthermore, graphical rendering instructions associated to each pre-rendered map need to be defined in accordance with the OGC Styled Layer Descriptor (SLD) standard. All supplementary files must have the same name as the desired image so that GeoNode identifies them correctly during the upload process. The creation of metadata records (XML) and styling information (SLD) was done using functions from the R-package plotKML (Hengl 2014).

The presented web-platform is described in greater detail within corresponding deliverable reports related to work package 2 (geospatial data management) of the EU-FP7-project CLIMB. Most of these reports can be freely downloaded, for instance, from the project's website at http://www.climb-fp7.eu. The CLIMB Geoportal itself is available under http://lgi-climbsrv.geographie.uni-kiel.de. Appendix .2 at page 204 provides some portal screenshots focusing on resources related to soil mapping outputs produced in this work.

Chapter 5

Results

This chapter presents the results of soil spatial prediction at the Sardinian test sites, based upon the data and methods introduced in the previous sections. Part 5.1 distinguishes between the field 21 and field 33, and mapping focuses on the integration of covariates from geophysical measurements into regression-based interpolation techniques. Part 5.2 elaborates the potential of a hybrid soil mapping approach combining an artificial neural network model with bivariate kriging at the landscape scale.

5.1 Interpolation results at the field scale: San Michele farm

The current section starts with an extensive exploratory data analysis of fields 21 and 33 located at the San Michele farm. Subsequently, spatial prediction results are presented following a hybrid interpolation technique based on regression and cokriging. In particular, digital soil mapping at the field scale focuses on the integration of relatively novel predictors from geophysical measurements combined with traditional land-surface parameters and point coordinates. Model efficiency is quantified by validation measures based on leave-one-out cross-validation.

5.1.1 Exploratory data analysis

Exploratory data analysis is helpful to familiarise with any new dataset. For this purpose, basic statistical parameters, as summarised in table 5.1, provide a first impression on raw target variables and their distribution. Focusing on field 21, sand content is predominant on average among the three soil textural fractions with a mean value of 40.5 %. Its variation portrayed by a standard deviation of 7.3 % is also slightly higher compared to clay and silt content. In addition, a positive skewness value of 0.6 indicates some small right-skewed behaviour for that particular variable. On the contrary, clay and silt content are rather symmetric and normally distributed. With respect to field 33, clay and sand content are almost of the same size on average, each having a mean value of about 39 %. Both fractions also vary to a similar extent with standard deviations of approximately 6.5 %. Note, however, that they differ in symmetry as evidenced by a negative skewness value of 0.6 for clay and a positive one of 0.4 for sand.

Table 5.1: Summary statistics of soil textural fractions, San Michele farm

Targets[1]	Min	Max	1st[2]	2nd[2]	3rd[2]	Mean	Std.[3]	Skewness	Kurtosis
- for field 21 (N = 43) -									
CLAY	21.43	44.52	28.26	34.43	37.90	33.50	5.76	−0.19	−0.84
SILT	19.08	34.87	23.25	25.55	28.36	26.00	3.89	0.35	−0.35
SAND	31.02	55.31	35.36	36.70	46.61	40.51	7.26	0.62	−1.06
- for field 33 (N = 64) -									
CLAY	20.94	51.63	36.09	40.62	43.92	39.69	6.49	−0.64	−0.01
SILT	10.79	35.86	18.30	21.95	23.72	21.27	4.91	0.33	0.56
SAND	26.35	55.31	34.20	37.24	43.57	39.04	6.62	0.40	−0.70

[1] Clay, silt and sand content in %, [2] Quartiles (2nd = Median), [3] Standard deviation

Referring to field 21, own measured mean values are compared to previous surveys by AGRIS Sardegna and UFZ Leipzig. Table 5.2 enumerates the centre of each measured composition and roughly underlines the similarity of the different case studies. Both, UFZ and CAU/LMU found an average clay loam texture, whereas a more sandy soil type was observed by AGRIS. This slight discrepancy in sand content between the AGRIS results and own measurements might be due to different limits used to distinguish silt and sand particles (50 µm instead of 63 µm).

Note, however, that exact field and laboratory conditions of the previous surveys are unknown. In addition, sample sizes are significantly smaller for the two former case studies.

Table 5.2: Mean values of soil textural fractions of different surveys at field 21

Survey	Clay	Silt	Sand	Sample size
AGRIS 2005 by M. Melis	32.6	20.7	46.7	9
UFZ 2010 by U. Werban	27.9	29.3	42.8	5
CAU/LMU 2010/2011 by M. Blaschek and S. Meyer	33.4	26.1	40.5	43

Clay, silt and sand content in %

Linking the numerical summaries from both investigated fields uncovers some shift concerning clay and silt contents. Regarding the mean value, field 21 exhibits a lower clay content than field 33, but is on average enriched in silty material. The impression of soil textural differences between the two adjacent fields is further confirmed taking the complete composition into account by performing a Hotelling T^2-test. Its F-statistic of 15.3 (df1 = 3, df2 = 103) with an associated p-value close to zero (2.6e-08) clearly implies that the null hypothesis of equal multivariate means must be rejected for the two samples. Thus, with regard to obvious differences in summary statistics and the significant statistical test result, both fields are not representative for each other in terms of soil texture. Mixing the two datasets to proceed with one larger sample is, therefore, unjustified.

Graphical representations of the distributions of the target variables are presented in form of histograms (see figure 5.1) and box-and-whisker-plots (see figure 5.2). These visualisations basically acknowledge the findings from descriptive statistics, highlighting the clay and silt differences between field 21 and 33. It is also striking that clay content of field 33 is slightly skewed to the left with one potential outlier. Furthermore, clay and silt measurements of field 21 spread less compared to observations from field 33.

Figure 5.3 displays the three-part simplex of (compositional) soil textural fractions in form of an equilateral triangle. It incorporates all sample locations from the two surveyed fields of the San Michele farm. The centres of both compositions are located inside the clay loam class according to the USDA soil taxonomy and

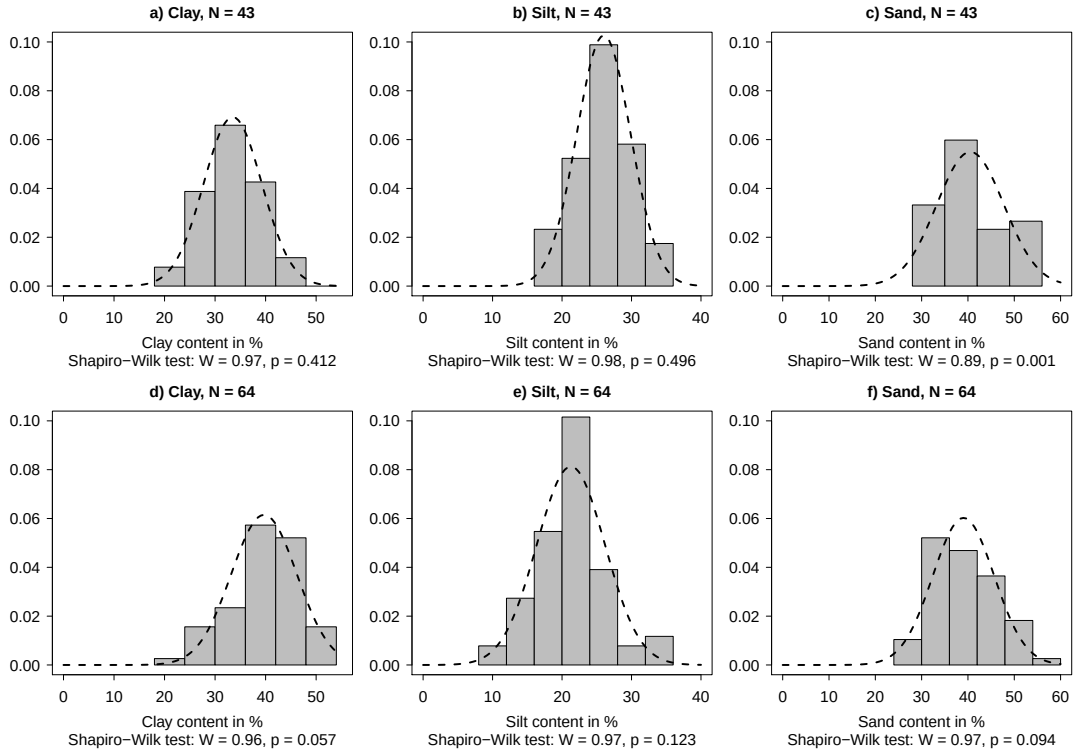


Figure 5.1: Density histograms of soil textural fractions at the San Michele farm. Figures a) – c) refer to field 21, while the graphs in d) – f) are determined from field 33. The dashed lines represent the corresponding normal distribution based on sample mean values and standard deviations.

FAO guidelines. In terms of absolute counts, clay loam remains the predominant class with regards to field 21, while field 33 is also characterised by a high number of points belonging to the clay texture class.

Until now, univariate analysis steps were examined, focusing on distributional concerns of each single target variable. However, knowledge about the target variables is not sufficient for subsequent modelling procedures, since exhaustive auxiliary information is available for the two fields of interest serving as potential explanatory factors. In order to investigate the bivariate relations between target variables and continuous covariates, scatterplot matrices are prepared (see figures 5.4 and 5.5 on pages 93–94, respectively). In the lower panel of each scatterplot matrix, Pearson's correlation coefficients (ρ) are shown.

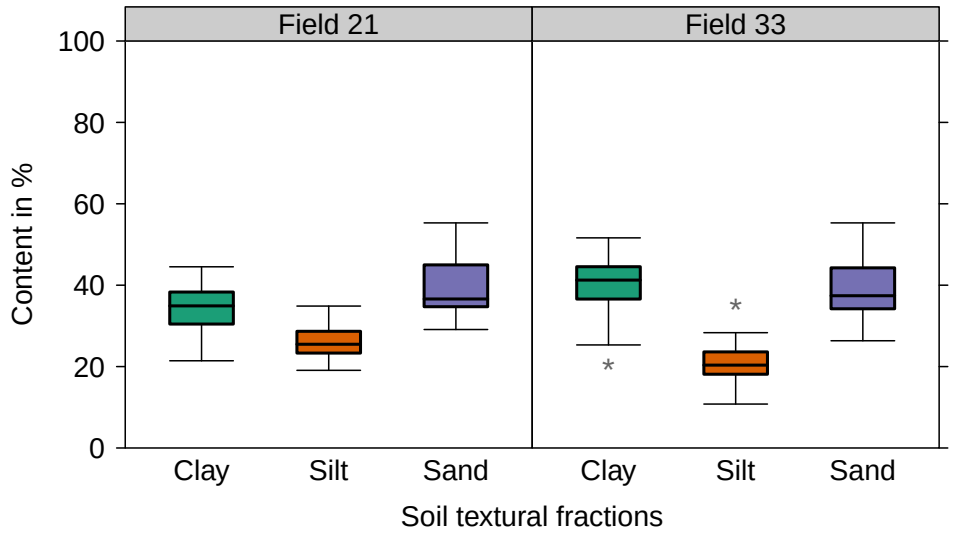


Figure 5.2: Contents of soil textural fractions by agricultural fields at the San Michele farm. The number of observations at field 21 and 33 are 43 and 64, respectively. Potential outliers are marked by an asterisk and represent values outside 1.5 times the interquartile range.

With respect to field 21, soil textural fractions are significantly related to gamma-ray nuclides. The highest correlation coefficient of 0.58 is measured between sand and uranium, while finer texture classes exhibit a slightly weaker and negative correlation with nuclide concentrations. Opposed signs are observed for ratios of pairwise nuclides such as Th/K, being positively related to clay and silt content, but negatively in case of sand. However, these correlations are not statistically significant ($P < 0.05$) for field 21. Slightly different correlations between soil textural fractions and radiometric data emerge for field 33. While clay content correlates to a higher extent compared to field 21, silt is not significantly related to nuclide concentrations. Again the sign is negative for the finer particle size fractions and positive for sand content. Apparent electrical conductivity (ECa) measures from electromagnetic induction are only significantly related to silt ($\rho = 0.38$) at field 21, no matter which operation mode is considered. At field 33, no significant correlations exist between ECa observations and soil textural fractions. Terrain attributes such as elevation, slope and wetness index show partly significant, yet diverse relations to soil texture in terms of direction, comparing the two fields.

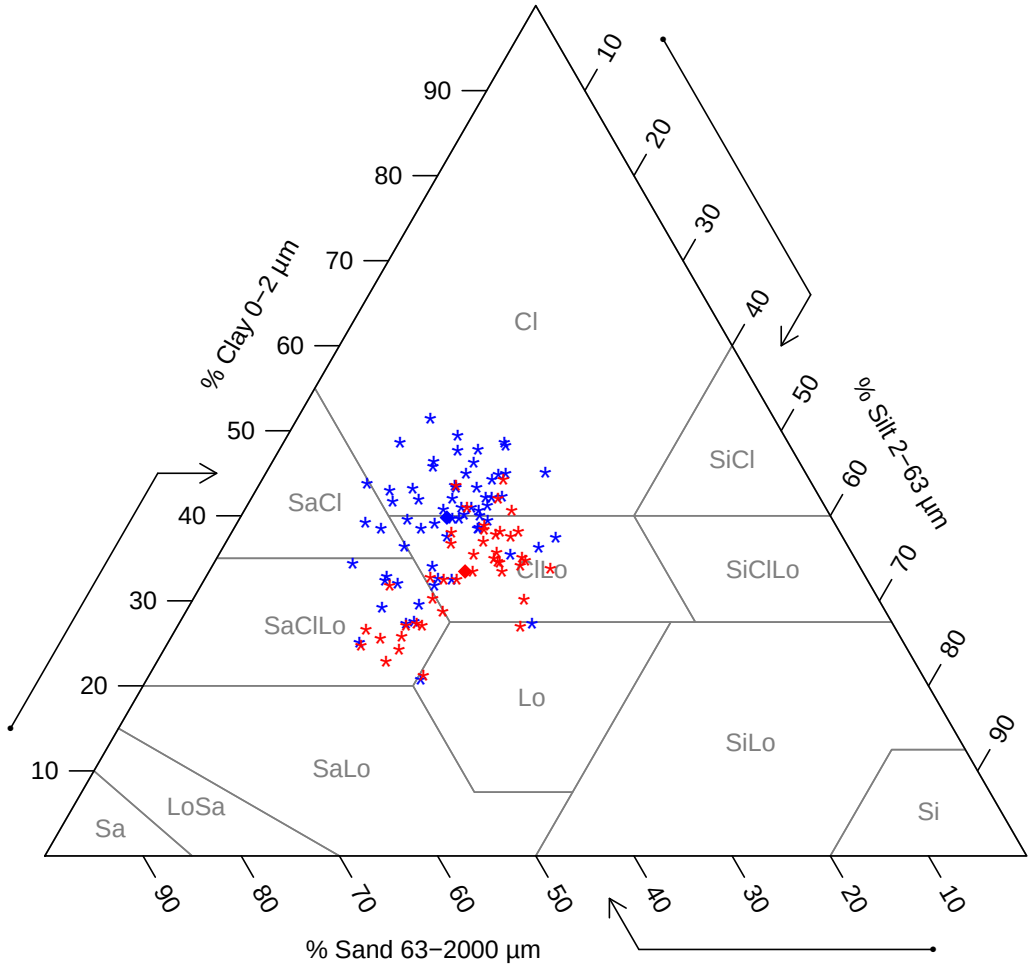


Figure 5.3: Ternary diagram of sampled soil texture classes at the field scale, San Michele farm. Grey lines indicate the transition between different textural classes in accordance with the USDA soil taxonomy (Soil Survey Staff 1999). The red-coloured symbols are related to field 21, whereas blue asterisks refer to field 33. The rhombuses represent the centre of the given compositions.

Both scatterplot matrices shown in figure 5.4 and 5.5, present merely selections of all available covariates. For example, the additional dose rate parameter only represents a combination of all three gamma-ray nuclides and does not reveal any new relations between radiometric data and soil texture. It is, therefore, excluded from further analysis steps. Th/U and U/K ratios exhibit very similar, yet not identical, correlations with the target variables as already introduced by

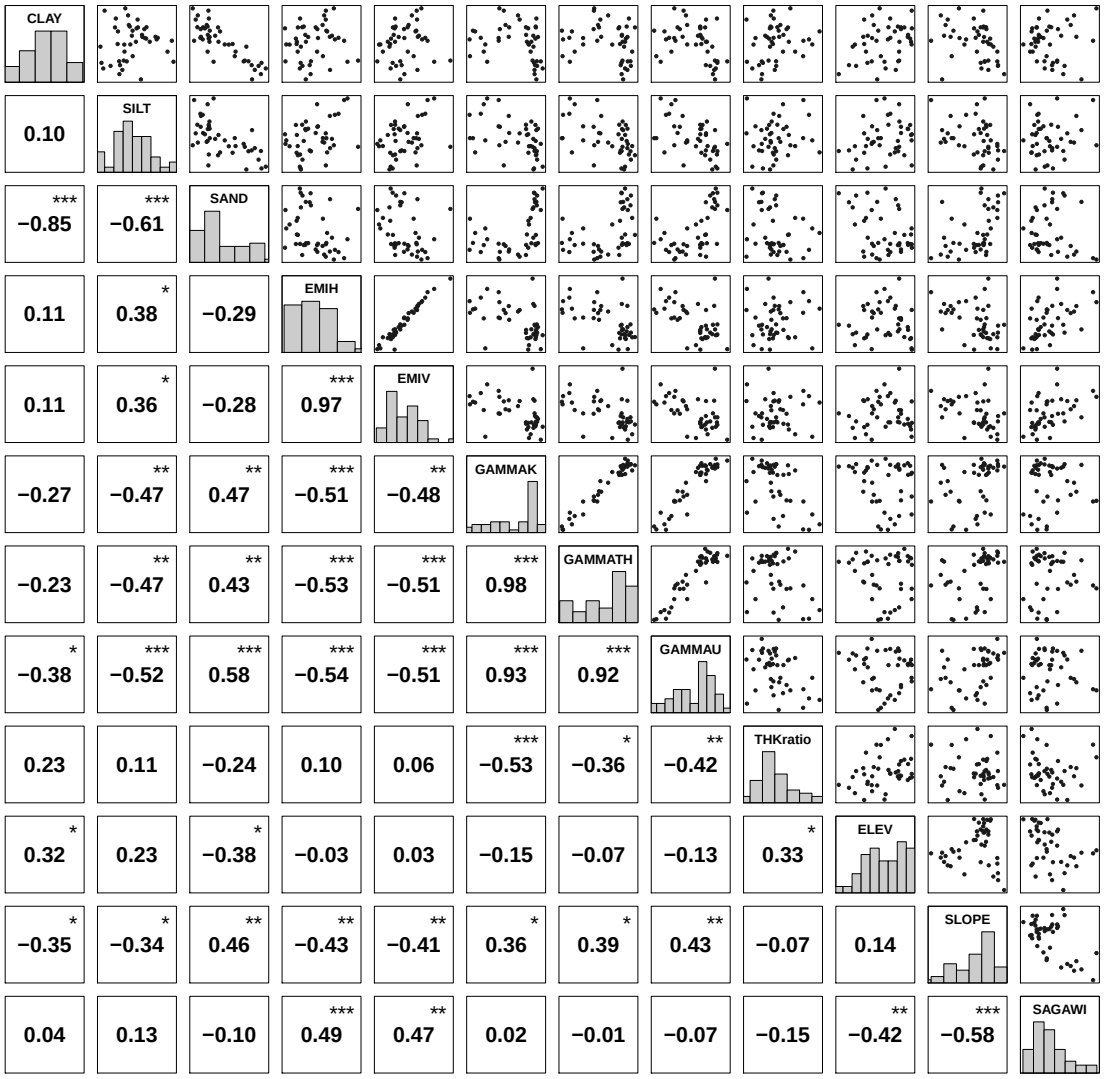


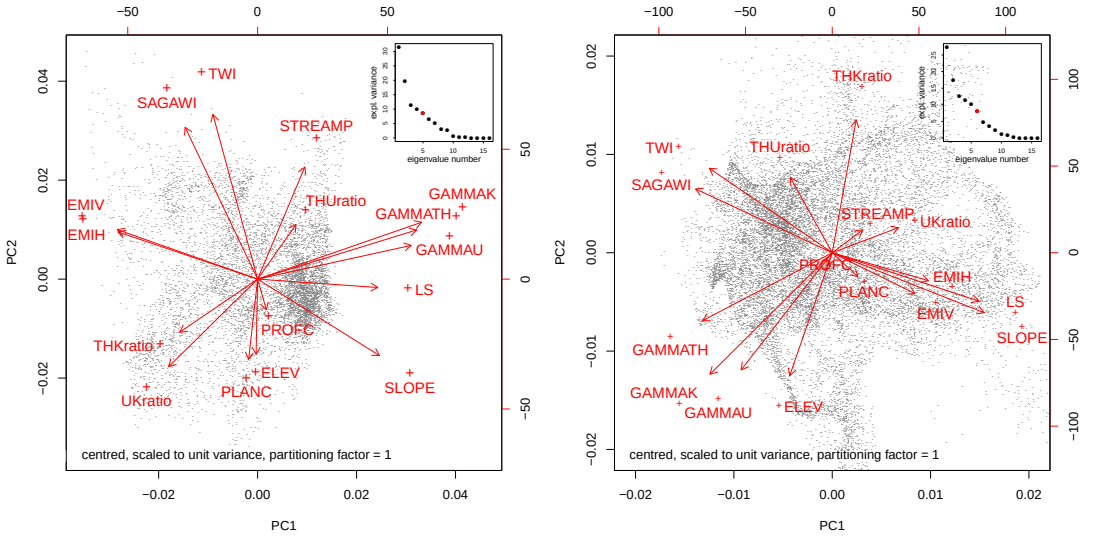
Figure 5.4: Scatterplot matrix related to field 21 at the San Michele farm. The lower triangle displays Pearson's correlation coefficients of each pairwise variable combination. Statistical significance is indicated as follows: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Diagonal boxes show the histograms of each parameter and bivariate scatterplots are drawn in the upper panel.

Th/K. For the sake of clarity, these parameters are not shown in the scatterplot matrices, but are still considered for subsequent modelling. The same applies to additional land-surface parameters such as topographic wetness index and length-slope factor. By contrast, aspect (or exposition) contains several no-data

Figure 5.5: Scatterplot matrix related to field 33 at the San Michele farm

pixels and is omitted from the group of possible covariates for numerical reasons, although being significantly related to soil texture at field 21.

Focusing on Pearson's correlation coefficients between independent variables as shown in the lower triangle of figures 5.4 and 5.5, clearly indicates that most of the given covariates are strongly correlated with each other. Considering this high level of multicollinearity and the relatively small sample sizes at both sites,



(a) field 21 (b) field 33

Figure 5.6: Covariance biplots at the field scale. Both biplots present scores of the observations (black dots) and coefficients of the variables (red vectors) on the first two PCs. Note that the red plus-signs actually represent the length of the vectors and not the arrow heads. Scree plots are shown in the right-upper corner, indicating the considered number of factors as red points.

a principal component analysis (PCA) is highly beneficial for reducing the dimensions among the numerous explanatory factors prior to statistical modelling. PCA results for both investigated fields are illustrated in form of biplots in figure 5.6. They basically consider the first two principal components (PCs) of each analysis and focus on inter-variable covariance structure. The latter implies that the cosines of apparent angles between vectors roughly reflect the correlations between the corresponding covariates. Since all co-variables were normalised prior to decomposition, the lengths of the vectors again approximate their representation by the two considered PCs (see Greenacre 2010, ch. 6). Thus, for instance, ECa from electromagnetic induction (EMI) seems to have more influence on the first two PCs at field 21 compared to field 33. In addition, vectors of EMI measures and gamma-ray nuclides are almost orthogonal in figure 5.6b and the involved variables, therefore, be treated as uncorrelated. This observation implies that the two geophysical sensors seem to be affected by different soil properties at that

particular field, and thus may complement each other for soil (texture) mapping. At field 21, by contrast, these two groups of covariates exhibit some negative correlation expressed by opposed vector directions.

Table 5.3: Loading coefficients of selected principal components at the field scale

Covariate	Abbreviation	Field 21		Field 33		
		PC1	PC2	PC1	PC2	PC3
Elevation	ELEV	0.00	−0.23	−0.11	−0.40	0.15
SAGA wetness index	SAGAWI	−0.18	0.47	−0.36	0.21	−0.12
Slope	SLOPE	0.30	−0.23	0.40	−0.20	0.00
Profile curvature	PROFC	0.02	−0.09	−0.01	−0.04	0.33
Plan curvature	PLANC	−0.02	−0.25	0.07	−0.08	0.35
Topographic wetness index	TWI	−0.11	0.50	−0.32	0.28	−0.27
Stream power index	STREAMP	0.12	0.34	0.08	0.07	−0.37
Length-slope factor	LS	0.29	−0.03	0.39	−0.16	−0.09
ECa in horizontal mode	EMIH	−0.34	0.14	0.25	−0.09	−0.12
ECa in vertical mode	EMIV	−0.34	0.15	0.22	−0.13	−0.13
Potassium ^{40}K	GAMMAK	0.40	0.17	−0.32	−0.40	−0.04
Thorium ^{232}Th	GAMMATH	0.39	0.15	−0.34	−0.22	−0.07
Uranium ^{238}U	GAMMAU	0.37	0.10	−0.24	−0.38	−0.33
Th/K ratio	THKratio	−0.19	−0.16	0.06	0.43	−0.07
Th/U ratio	THUratio	0.09	0.17	−0.11	0.25	0.40
U/K ratio	UKratio	−0.22	−0.27	0.17	0.08	−0.45
Eigenvalue	-	5.03	3.17	4.37	2.81	2.01
Explained variance	-	31.46	19.79	27.34	17.56	12.56

To identify the most influencing variables of the principal components, loading coefficients are analysed. Table 5.3 summarises the loading values of those PCs that cumulatively account for more than 50 % of the total spatial variation. With regard to field 21, gamma-ray nuclides and ECa measurements from electromagnetic induction produce the largest loading values and are considered as most influencing in the context of PC1. The second PC is better represented by the two wetness indices. Land-surface parameter are the dominant covariate class for the first PC at field 33, whereas ratios of nuclide concentrations influence the second and third PCs. EMI measures are only relevant with respect to PC4 accounting for 11.3 % of the total spatial variation. Selecting the optimal set of factors utilised as predictors in further (geo)statistical modelling based on the eigenvalues, i. e. the squared singular values of the decomposition. PCs, whose eigenvalues exceed 1.05 and 1.03 at field 21 and 33, respectively (calculated after

Karlis et al. 2003), remain in the analysis. Thus, the first five PCs are deemed relevant for field 21, while the first six PCs are chosen at field 33 according to the given critical values. These selections are supported by visual inspection of scree plots shown in the upper right corner of figure 5.6, where the so-called elbow of the curve roughly indicates the number of factors to keep.

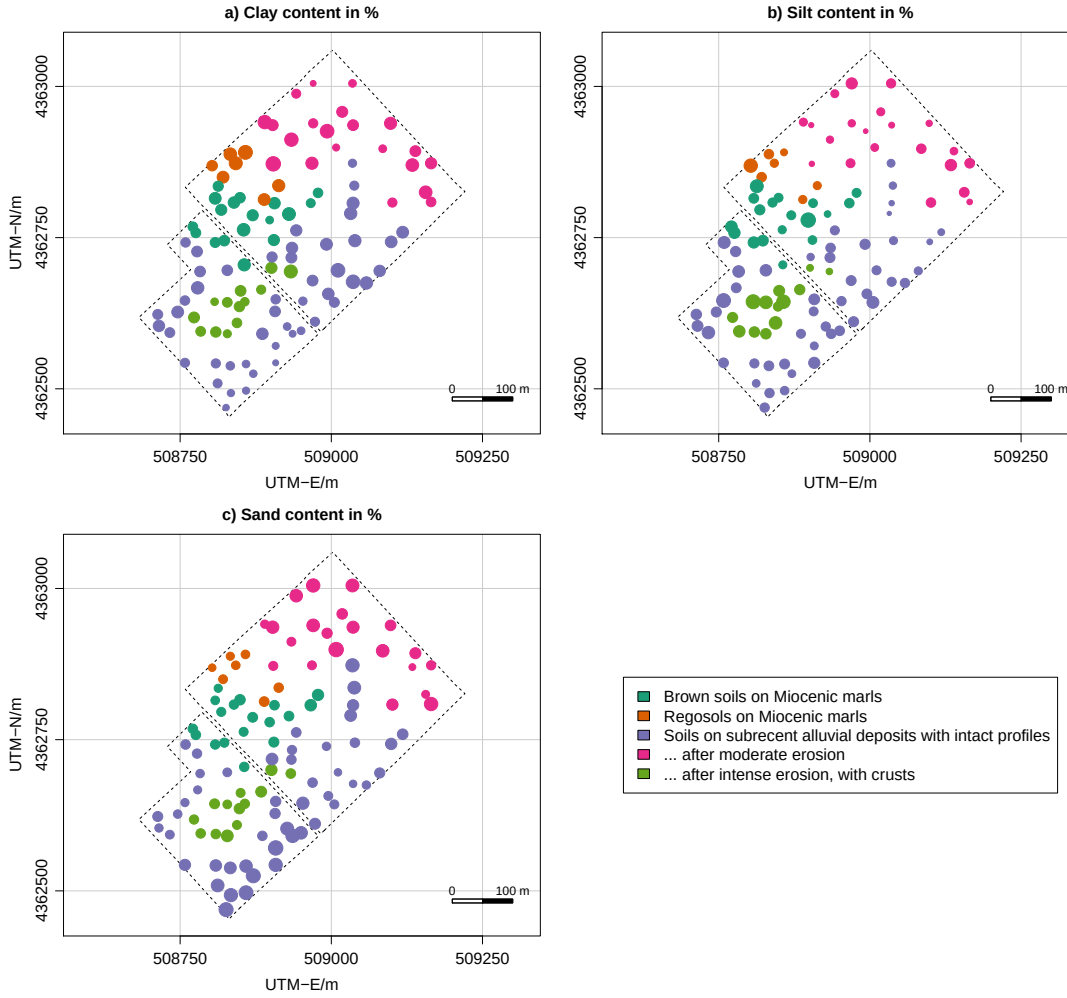


Figure 5.7: Postplots of measured clay, silt and sand content by soil unit after Aru (1966) at field 21 and 33. The size of the circles represents the proportion of the respective textural fractions.

After focusing on the values of involved datasets, proportional symbol postplots are examined to familiarise with spatial peculiarities among the target variables

at both fields. Figure 5.7 shows the measurement locations and proportions of the textural fractions, and highlights possible differences due to certain soil units as defined by Aru (1966). It can be seen in the maps that the previously surveyed soil zones do not explain much of the current spatial variation among clay, silt and sand content. This resource is, therefore, inappropriate to address any global trend of the target soil properties and no longer considered for spatial modelling. Focusing on global spatial structures, silt exhibits some tendency towards higher percentages in the central part of field 21. Another apparent clustering of high values occurs for sand contents sampled close to the southern edge of field 21. In contrast to global trend observations from field 21, spatial distribution of measured soil textural fractions is more subtle regarding field 33 with higher and lower values being rather dispersed.

Local spatial dependence representing the second key factor of spatial structure is best examined by plotting empirical variances against distance. From variograms shown in figure 5.8 it is recognisable that there is spatial auto-correlation for all three particle size fractions at both fields. Differences between two points located close to each other are on average lower than those for more widely spaced point pairs. Focusing on field 33, distinctions for clay content show an almost ideal spherical curve shape with a linear increase in dissimilarity up to 111 m and a nugget variance of 13 $\%^2$. Silt exhibits a longer range of 151 m and a closer nugget-to-sill ratio of 62 % indicating a larger amount of micro-scale variability and measurement error components. In addition, clay is overall more variable, expressed by a higher total sill variance of 43 $\%^2$ compared to 25 $\%^2$ regarding silt. Unlike clay and silt content, where spherical curves have been fitted, spatial variation of sand is better addressed by an exponential model ($c_0 = 8$, $c_1 = 55$, $a = 3r = 405$). Note the existence of an eye-catching outlier in the sand variogram at field 33 with an exceptional high variance for very short distances. However, only two point pairs contribute to that particular lag, neglecting its influence to the fitting procedure. Looking at variogram analysis results with respect to textural fractions at field 21 strong unbounded behaviour is revealed, starting from 135 m for clay and sand content. This observation clearly underlines the presence of a global trend already discussed in the paragraph related to postplot

analysis. In contrast to clay and sand content, only very little spatial autocorrelation could be detected for silt content at field 21.

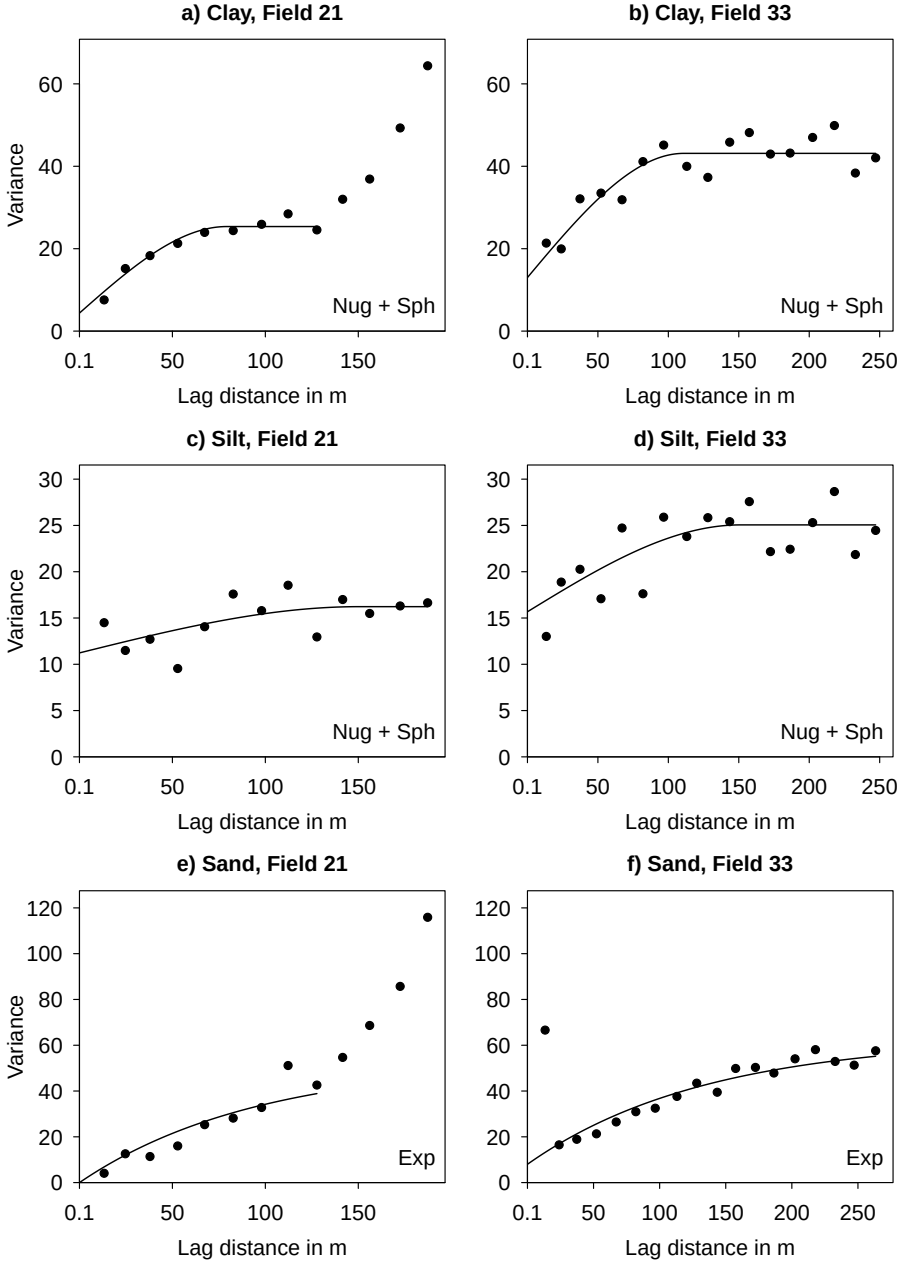


Figure 5.8: Variograms of soil textural fractions at field sites. Empirical variances are estimated by the method-of-moments and fitted using (weighted) least squares (N/h^2) or adjusted by eye (sand).

5.1.2 Regression modelling

Exploratory data analysis revealed high correlations among the covariates derived from three different sources of auxiliary information. In order to overcome multicollinearity effects and to simultaneously reduce the number of explanatory factors in the final (regression) models, principal components were used rather than raw covariates. Figure 5.9 shows the spatial distribution of the first four principal components at both fields 21 and 33. Focusing on the first PC at field 33, high values are predominant at hillside positions which corresponds to a relatively high positive loading value for slope in table 5.3 at page 96. At field 21, the map of PC1 appears very similar to the spatial estimates of gamma-ray nuclides and the total dose rate shown in figure 4.6 at page 45 due to a strong positive weighting of these variables. On the contrary, the second PC exhibits increased values in zones characterised by low gamma-ray emission and small heights at field 33 according to negative loading coefficients for nuclide concentrations and elevation, respectively. PC3 at field 33 is dominated by ratios of pairwise nuclides, but does not reveal a remarkable spatial pattern.

The number of initially selected principal components was based on critical (eigen)values calculated in accordance with Karlis et al. (2003). For effectiveness and parsimony, these initial subsets of predictors are further reduced using stepwise regression. In case of field 21, PC1 as well as the x- and y-coordinates remain in the final regression model for alr-transformed clay and silt content. The best subset of predictors in terms of the Akaike Information Criterion (AIC) at field 33 comprises the first three PCs in respect of clay content, whereas silt is most accurately modelled considering PC1 and PC2 only. Backward-elimination and forward-selection procedures led to the same final models.

Table 5.4 and 5.5 summarise the parameter estimates from ordinary least squares (OLS) regression at field 21 and 33, respectively. In addition, common summary statistics are provided to roughly evaluate the quality of the given models. With respect to field 21, all coefficients can be considered as highly significant following the F- and t-statistics results. Thus, each single predictor is likely to be a reasonable completion of the examined models. The only exception is PC1 for

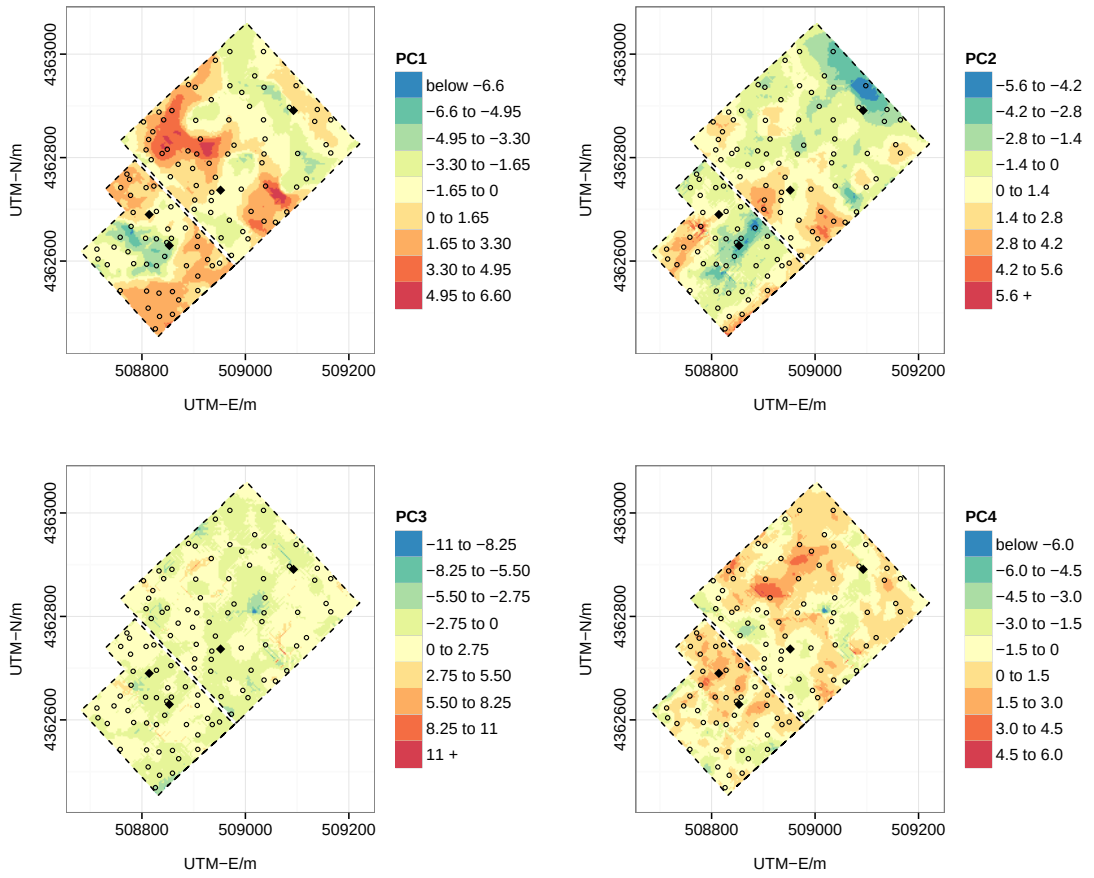


Figure 5.9: Spatial distribution of the first four principal components at the field scale, San Michele farm. The points represent sample locations and the rhombuses indicate the position of surveyed soil profiles.

alr-transformed clay content which is slightly less significant ($P < 0.05$) than the other parameters ($P < 0.001$). In terms of coefficients of (multiple) determination, both models reach comparatively high accuracy rates with explained variance levels of 67 % for clay and 74 % for silt content. The conspicuously large intercept values of both regressions are irrelevant with regards to interpretation and can be traced back to the metrics of incorporated x- and y-coordinates.

Referring to field 33, most coefficients are significant at the 0.05 significance level except for PC1 for alr Clay ($P < 0.001$). As with field 21, each chosen predictor proved to be meaningful for modelling alr-transformed clay and silt content at

Table 5.4: Regression parameter estimates and summary statistics at field 21

	alr Clay			alr Silt		
Variable	Estimate	Std. error[1]	t-value[2]	Estimate	Std. error[1]	t-value[2]
- Regression coefficients -						
Intercept	−9443.0720	1944.5050	−4.86***	−7739.3890	1484.9620	−5.21***
PC1	−0.0364	0.0137	−2.66*	−0.0506	0.0105	−4.83***
X	−0.0021	0.0005	−3.86***	−0.0016	0.0004	−3.80***
Y	0.0024	0.0004	5.65***	0.0020	0.0003	6.01***
- Summary statistics -						
R-squared		0.67			0.74	
F-statistic		26.654			36.718	
p-value		1.513×10^{-9}			1.921×10^{-11}	

[1] Standard error, [2] Statistical significance is indicated as follows: * P < 0.05, ** P < 0.01, *** P < 0.001

the given test sites. However, R-squared values are much less compared to those of field 21, indicating a considerably weaker overall prediction performance at field 33. Explained variance levels of 32 % and 17 % are reached by the regression models for clay and silt, respectively.

Table 5.5: Regression parameter estimates and summary statistics at field 33

	alr Clay			alr Silt		
Variable	Estimate	Std. error[1]	t-value[2]	Estimate	Std. error[1]	t-value[2]
- Regression coefficients -						
Intercept	−0.0206	0.0357	−0.58	−0.6630	0.0418	−15.87***
PC1	0.0649	0.0162	4.01***	0.0480	0.0191	2.51*
PC2	0.0439	0.0195	2.25*	0.0562	0.0228	2.46*
PC3	0.0434	0.0204	2.13*	-	-	-
- Summary statistics -						
R-squared		0.32			0.17	
F-statistic		9.57			6.12	
p-value		2.95×10^{-5}			3.79×10^{-3}	

[1] Standard error, [2] Statistical significance is indicated as follows: * P < 0.05, ** P < 0.01, *** P < 0.001

A proper (OLS) regression analysis requires the critical assessment of the model residuals. On the one hand, particular hypothesis tests are evaluated to approve important assumptions of the applied linear models such as normality, variance homogeneity and independence of the errors. On the other hand, outliers and

influential observations are examined through different diagnostic methods. The quantitative results of both analysis are outlined in table 5.6 and 5.7.

Focusing on field 21, the distributions of the residuals for both, alr-transformed clay and silt content, pass the Shapiro-Wilk normality tests with p-values of 0.42 and 0.13, respectively. Even higher p-values of 0.49 (alr Clay) and 0.73 (alr Silt) are obtained from Breusch-Pagan tests, clearly confirming a valid assumption of constant error variance for both regression models. Moran's I test statistics results additionally suggest the absence of spatial correlation among the alr Clay residuals (p-value = 0.23), whereas spatially uncorrelated errors are less likely but still significant for alr Silt (p-value = 0.08). Taking all testing results into consideration, it follows that major OLS regression assumptions are met. Thus, the outcome of the underlying linear models for field 21 seem trustworthy. In a second step, several influence measures are analysed to identify unusual observations that might disproportionately affect the OLS estimates. Starting with regression outliers, no studentized residual with a Bonferroni adjusted p-value less than 0.05 was found. Critical leverage points are judged by hat values exceeding three times the average hat value, that is, the number of parameters including the intercept divided by the sample size (Belsley et al. 2005). With regard to field 21, no such unusually high hat value exists. Thus, regression diagnostics at field 21 suggest that OLS estimates do not suffer from any outstanding measured response nor are they severely influenced by odd values among the predictors. Nevertheless, investigating both effects together using Cook's distance measure reveals at least two points (IDs 3 and 17) that exceed the common cut-off rule of thumb with a numerical threshold of $4/(N - k - 1)$. Hence, these two samples are possibly influential on the least squares fit for alr-transformed clay and silt content. This should be kept in mind while interpreting final prediction results. However, since there are no evident reasons that allow the exclusion of those points, the setting of discussed regression models remains unchanged.

Concerning field 33, Shapiro-Wilk normality tests with p-values of 0.93 and 0.75 for alr-transformed clay and silt, respectively, clearly confirm normally distributed residuals. In contrast to field 21, the Breusch-Pagan test results for alr Clay (p-value = 0.03) imply that the null hypothesis of constant error variance should

Table 5.6: Regression diagnostics and test statistics at field 21

	alr Clay				alr Silt			
	Min	Max	Mean	Std.[1]	Min	Max	Mean	Std.[1]
- Residual summary statistics -								
Residuals	−0.34	0.57	0.00	0.19	−0.33	0.26	0.00	0.15
Studentized residuals	−1.84	3.27	0.00	1.04	−2.33	1.77	−0.01	1.03
- Leverage points -								
Maximum hat value		0.1692				0.1692		
- Shapiro-Wilk normality test -								
W		0.9738				0.9588		
p-value		0.4247				0.1258		
- Breusch-Pagan test -								
χ^2		0.4735				0.1211		
p-value		0.4914				0.7278		
- Global Moran's I [2] -								
Moran I		0.0923				−0.1909		
z-score		1.2090				−1.7390		
p-value		0.2267				0.0820		

[1] Standard deviation, [2] based on k-nearest neighbour weights with k = 4

be rejected on the 0.05 significance level. Consequently, OLS estimates are not necessarily efficient and related F- and t-statistics might be inappropriate due to biased standard errors of the coefficients. With respect to Moran's I test results, alr Silt residuals seem to be spatially correlated as indicated by a p-value slightly below the 0.05 threshold. Thus, instead of assuming variance homogeneity and independence of the errors in an OLS regression framework, coefficients for final prediction are fitted using generalised least squares (GLS). In addition to the weak assumption violations, analysing influence measures also reveals a higher number of possible complications compared to regression modelling at field 21. Whereas no measured response values are suspicious, based on the Bonferroni outlier test, sample point 6 has leverage as evidenced by a noteworthy hat value. However, this particular point is associated with low studentized residuals and, therefore, is not seen as problematic. Following the more sensitive Cook's distance measure, samples 68, 69, 1, 92 and 37 are marked as potentially influential for the regression

Table 5.7: Regression diagnostics and test statistics at field 33

	alr Clay				alr Silt			
	Min	Max	Mean	Std.[1]	Min	Max	Mean	Std.[1]
- Residual summary statistics -								
Residuals	−0.68	0.64	0.00	0.26	−0.82	0.83	0.00	0.31
Studentized residuals	−2.70	2.54	0.00	1.02	−2.73	2.80	0.00	1.03
- Leverage points -								
Maximum hat value		0.1538				0.1450		
- Shapiro-Wilk normality test -								
W		0.9912				0.9871		
p-value		0.9309				0.7466		
- Breusch-Pagan test -								
χ^2		4.6729				0.0001		
p-value		0.0306				0.9927		
- Global Moran's I [2] *-*								
Moran I		0.0838				0.1510		
z-score		1.2364				2.0707		
p-value		0.2163				0.0384		

[1] Standard deviation, [2] based on k-nearest neighbour weights with k = 4

fit of alr-transformed clay content. With respect to linear modelling of alr Silt, points 68, 3, 77 and 92 exceed the critical threshold value. The order of listing reflects the magnitude of the respective Cook's distance values. Considering the relevant field records, point 68 was especially difficult to sample due to highly porous soil material, so that observational errors are likely in this particular case. The same applies to points 77 and 92, where a high amount of soil skeleton severely hampered the sampling. Moreover, repeated OLS regression analysis for field 33 without the influential samples 68, 77 and 92 remarkably improved model performance. Increased R-squared values of 0.41 for alr Clay and 0.28 for alr Silt are obtained after point removal compared to accuracy measures of 0.32 and 0.17 observed regarding the complete set of measured locations. As a consequence, samples 68, 77 and 92 are omitted from final regression modelling at field 33.

It follows from regression diagnostics discussed in the previous paragraphs that final coefficients used for linear modelling of the trend at the field scale are fitted

using generalised least squares. However, due to the fact that true covariance matrices of the errors are unknown in the present case studies, OLS estimates are used to derive regression residuals from which correlation structures are determined. These are then used again to recalculate the regression coefficients leading to the following final linear trend models:

- field 21 -

$$\text{alr Clay} = -10392.3 - 0.0287(\text{PC1}) - 0.0019(\text{X}) + 0.0026(\text{Y}) + \epsilon$$

$$\text{alr Silt} = -7906.9 - 0.0492(\text{PC1}) - 0.0016(\text{X}) + 0.0020(\text{Y}) + \epsilon$$

- field 33 -

$$\text{alr Clay} = -0.0407 + 0.0510(\text{PC1}) + 0.0562(\text{PC2}) + 0.0396(\text{PC3}) + \epsilon$$

$$\text{alr Silt} = -0.6270 + 0.0181(\text{PC1}) + 0.0640(\text{PC2}) + \epsilon$$

5.1.3 Geostatistical analysis of regression residuals

Regression modelling is meant to address any trend components among the alr-transformed target quantities using ancillary data (land-surface parameters, geophysical sensor data as well as x- and y-coordinates). In a next step, regression residuals are analysed for remaining auto- and cross-correlation. This section focuses on the (cross-)variogram modelling results based on ordinary least squares (OLS) regression residuals and the linear model of coregionalisation (LMC). Obtained parameter estimates are then used to calculate the auto- and cross-variogram matrices for cokriging-based interpolation of residuals from final GLS trend models.

As outlined in chapter 4.6.4.2 on page 72, an LMC is basically a linear combination of authorised variogram functions. Inspecting the residual auto- and cross-variograms shown in figure 5.10, reveals that only two basic structures are practically relevant at the given test sites at field level: pure nugget effects and spherical models. With respect to field 21, residuals from regressed alr-transformed clay content exhibit some auto-correlation, whereas no spatial dependence is observed

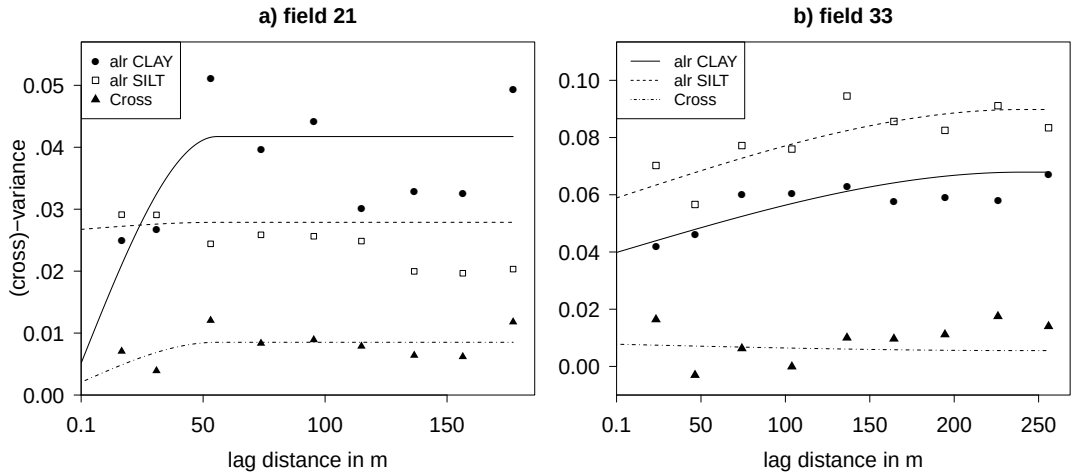


Figure 5.10: Residual (cross-)variograms of alr-transformed target quantities at the San Michele farm. The auto- and cross-(semi)variances come from method-of-moments estimation with a linear model of coregionalisation fitted by (weighted) least squares including a subsequent check for positive-definiteness. Note that cross-variogram estimates can be negative.

among the alr Silt residuals. Modelled overall range values of 56 m are relatively small indicating the presence of fine-scale variability, while no long-range structure can be detected. The latter observation implies that the preceding regression part successfully modelled large-scale variation among the log-ratio transformed target variables. At field 33, the LMC fits the empirical estimates fairly well. However, a linear increase in dissimilarity up to 240 m suggests some remarkable long-range variability among both, the alr-transformed clay and silt residuals. Obviously, removing any trend component through linear regression with secondary data was less successful compared to field 21. This confirms regression summary statistics shown in the previous section (see tables 5.4 and 5.5). Poor regression results were reported for alr Silt at field 33 and, accordingly, highest absolute variances occur among their residuals. Or, put another way, only very little variation in alr Silt at field 33 could actually be addressed by linear regression and most of it is still apparent in the corresponding residuals.

Table 5.8 emphasises the implications drawn from graphical representations by listing any important parameters of the fitted variogram functions. An Nugget-

Table 5.8: Residual (cross-)variogram characteristics at the San Michele farm

Parameter	Model	Fit. method	Nugget	p. sill	Range	np[1]	NSR
- for field 21 -							
alr Clay	Sph	LMC by WLS	0.0051	0.0366	55.7	10	12.3
alr Silt	Sph	LMC by WLS	0.0267	0.0011	55.7	10	95.9
Cross	Sph	LMC by WLS	0.0021	0.0065	55.7	20	-
- for field 33 -							
alr Clay	Sph	LMC by WLS	0.0398	0.0281	240.0	30	58.7
alr Silt	Sph	LMC by WLS	0.0589	0.0309	240.0	30	65.6
Cross	Sph	LMC by WLS	0.0078	−0.0022	240.0	60	-

Sph = Spherical model, LMC = Linear model of coregionalization, WLS = Weighted least squares

p. sill = partial sill, np[1] = number of point pairs within the first lag bin, NSR = Nugget-to-Sill-Ratio in %

to-Sill-Ratio (NSR) of 96 % indicates almost no spatial dependency among the alr-transformed silt residuals at field 21. By contrast, a strong auto-correlation behaviour is shown for alr Clay with a NSR of 12 %. However, taking into account the non-significant Moran's I test statistics obtained from regression diagnostics and considering that only 10 point pairs are available for lag one calculation, the true variogram function for alr Clay at field 21 could easily be different. Thus, the LMC fitted at field 21 is not as reliable as the one found for field 33, where NSR values jointly suggest the presence of moderate spatial dependencies.

5.1.4 Final prediction and model validation

Adding up the linear regression results and their cokriged residuals presented in the previous sections, yields the final estimates of soil textural fractions at unknown locations in respect of the studied agricultural fields 21 and 33.

Figure 5.11 shows modelled clay, silt and sand content after back-transformation. Resulting spatial patterns appear preponderantly plausible with regard to observations from site-specific exploratory (spatial) data analysis. In particular, this becomes obvious in the south-eastern part of field 21 where high sand (and low clay) contents are estimated. The slight increase of sandy material corresponds to the course of a former creek bed (see Cassiani et al. 2012) and also reflects

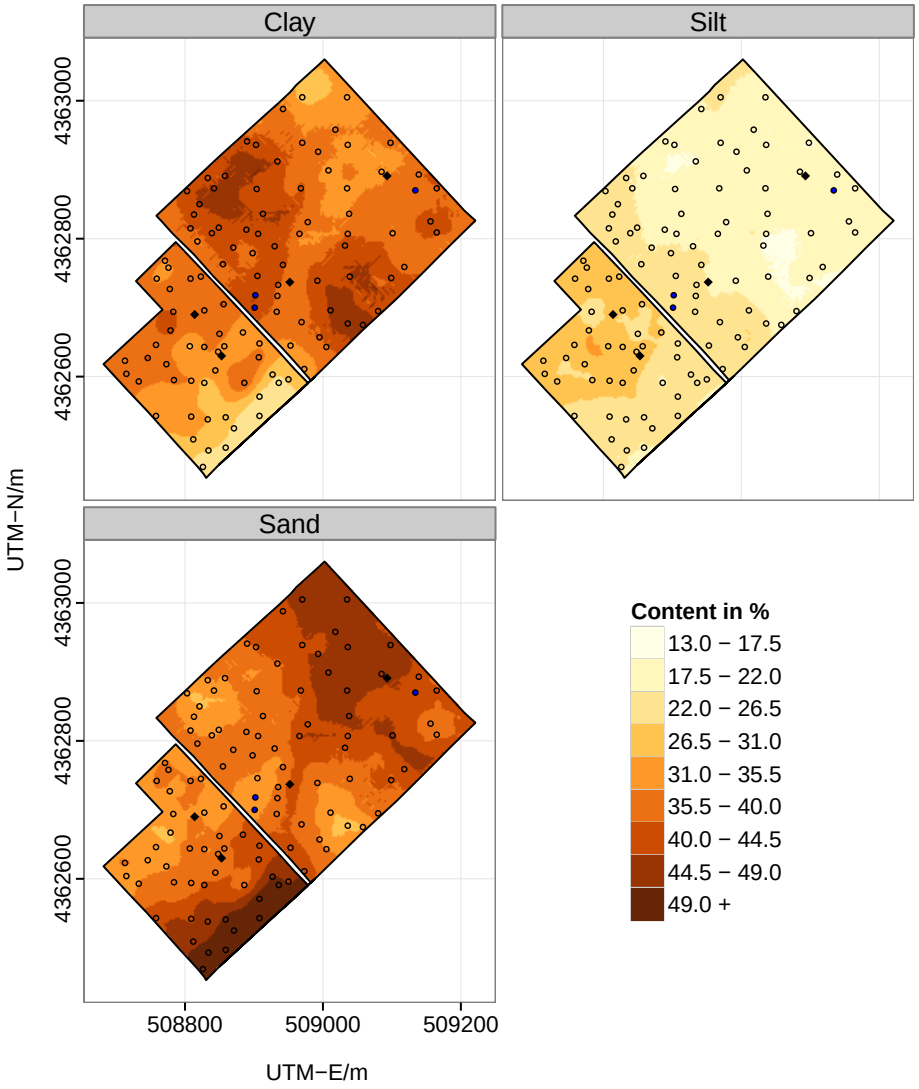


Figure 5.11: Regression cokriging estimates of soil textural fractions at the field scale. The prediction results are shown after additive generalised logistic back-transform. Hollow data points represent sample locations, blue-coloured dots are surveyed points omitted from final prediction. The black rhombuses indicate the position of analysed soil profiles.

the extraordinary significance of this particular part of field 21 as evidenced by different farming activities recorded during field campaigns. Focusing on field 33, higher and lower values of particle-size fractions occur rather dispersed. Instead of x- and y-coordinates, the first three PCs involved in the regression analysis part

dominate the look of resulting prediction maps. Since PC1 and PC2 were essentially influenced by land-surface parameters, high clay (and low sand) contents are primarily estimated at hillside positions and in lower elevated zones characterised by wetter conditions. Ratios of pairwise nuclides strongly contribute to PC3 causing band-shaped features especially easy to see in the clay content map. Since PC3 was irrelevant for modelling the alr-transformed silt target variable, stripes do not occur in silt-related outputs.

Although regression-induced appearances stand out, typical geostatistical features can be found in the prediction maps, as well. For instance, in the context of field 21, very short spatial correlation ranges fitted during residual auto- and cross-variogram modelling result in gentle bull's eye effects. On the contrary, a very strong smoothing impact occur with regard to silt content, and particularly at field 33. Despite the visually influencing equal colour scheme, this impression is in line with lower spreading of that particular fraction known from exploratory summary statistics.

Table 5.9: Summary statistics, normality tests and checks for remaining spatial correlation of model residuals at the field scale

	Min	Max	Mean	Std.	Skew	W	p	I	z	p
- field 21 -										
Clay	−11.63	10.17	−0.12	4.47	−0.28	0.99	0.960	−0.20	−1.85	0.064
Silt	−6.99	8.15	−0.01	3.30	0.00	0.99	0.888	−0.27	−2.55	0.011
Sand	−8.62	10.43	0.13	3.43	0.11	0.97	0.461	−0.12	−0.95	0.340
- field 33 -										
Clay	−14.51	14.06	−0.02	6.14	−0.06	0.99	0.744	−0.01	0.14	0.890
Silt	−15.07	8.91	−0.20	4.68	−0.34	0.96	0.077	−0.06	−0.58	0.563
Sand	−16.19	10.04	0.22	5.00	−0.53	0.98	0.270	−0.12	−1.22	0.223

Std. = standard deviation, W = Shapiro-Wilk W, I = Global Moran's I based on 4-nearest neighbour weights, z = z-score

Due to the relatively small number of samples available at each field, model evaluation was done on the basis of leave-one-out cross-validation instead of some split-sample method. Summary statistics and hypothesis test results computed from cross-validation residuals are summarised in table 5.9. It follows from absolute mean errors ranging between 0.01 and 0.22 that achieved clay, silt and sand predictions can be considered as unbiased estimates. Low-valued skewness

and significant Shapiro-Wilk test results ($P > 0.05$) also imply nicely symmetric cross-validation residual distributions. Comparing simple error metrics of both neighbouring sites, largest absolute discrepancies as well as average spreading are slightly higher for field 33 compared to field 21. With regard to remaining spatial correlation, global Moran's I test results are clearly negative for the target variables at field 33. A random spatial distribution is also observed for cross-validated sand residuals at field 21, but less likely regarding clay and silt content. Thus, spatial variation might still exist with respect to these two particle-size fractions and cannot be determined by the constructed RCOK model.

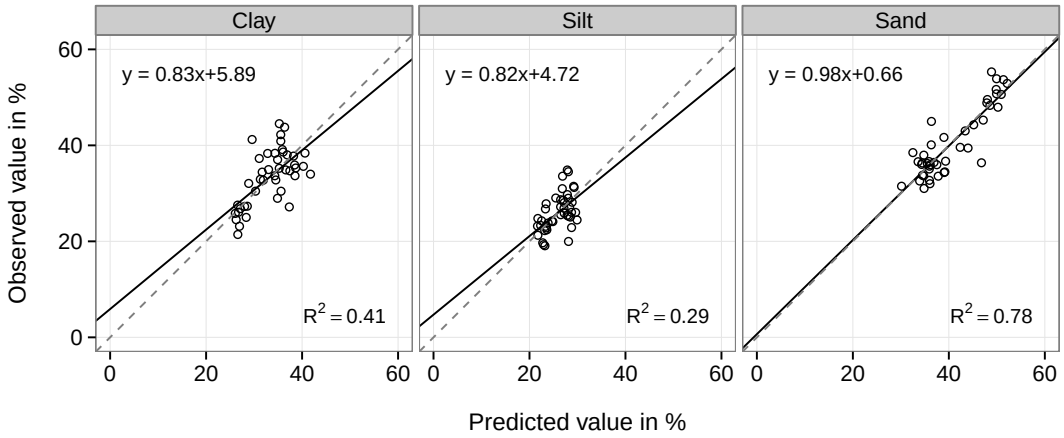
Following the given error analysis, predictions are found fairly adequate regarding both studied fields. This impression slightly weakens considering common validation measures listed in table 5.10. While satisfactory model efficiencies are reported from field 21, only 13 % and 9 % explained variance levels could be reached at field 33 for clay and silt content, respectively.

Table 5.10: Validation measures of soil spatial interpolation at the field scale

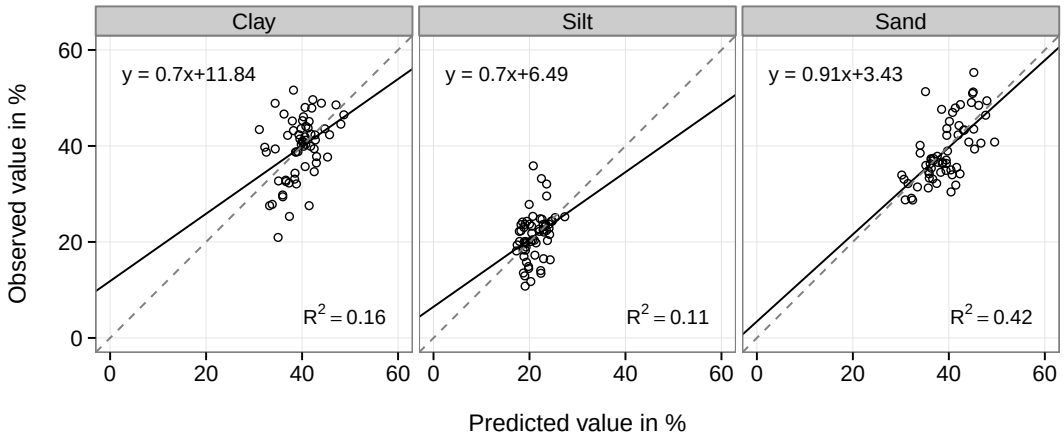
	Clay		Silt		Sand		
	RMSE	EF	RMSE	EF	RMSE	EF	STRESS
- field 21 -							
Regression cokriging	4.440	0.39	3.263	0.28	3.389	0.78	0.82
- field 33 -							
Regression cokriging	6.093	0.13	4.643	0.09	4.965	0.41	0.89

RMSE = root mean squared error, EF = model efficiency, STRESS = standardized residual sum of squares

Taking a closer look at individual estimates, scatterplots of actual versus predicted values are examined from figure 5.12. Focusing on field 21, sand content appears well-predicted as indicated by a regression line that is almost perfectly congruent with the desired 1:1 line and an associated determination coefficient of 0.78. Clay content regression reveals a slightly weaker performance with an R-squared value of 0.41, while silt content exhibits the worst model quality among the investigated target variables. In particular, the range of true silt contents is not really matched by the estimated values. This becomes even more obvious at field 33, where silt-related predictions are by far too smooth with an R-squared



(a) Field 21



(b) Field 33

Figure 5.12: Scatterplots of actual versus predicted values at the San Michele farm. The solid lines represent linear regression models of actual measurements on back-transformed regression cokriging estimates based on leave-one-out cross-validation. The dashed lines are the (desired) 1:1 lines.

value of 0.11, indicating a weak statistical model. The interpolation of clay content performed just slightly better, whereas the sand fraction is moderately well-estimated with an R-squared value of 0.42. Thus, in terms of association-based error metrics, visual inspection from scatterplots and model efficiencies, predictions at field 21 are much more accurate than those achieved from field 33.

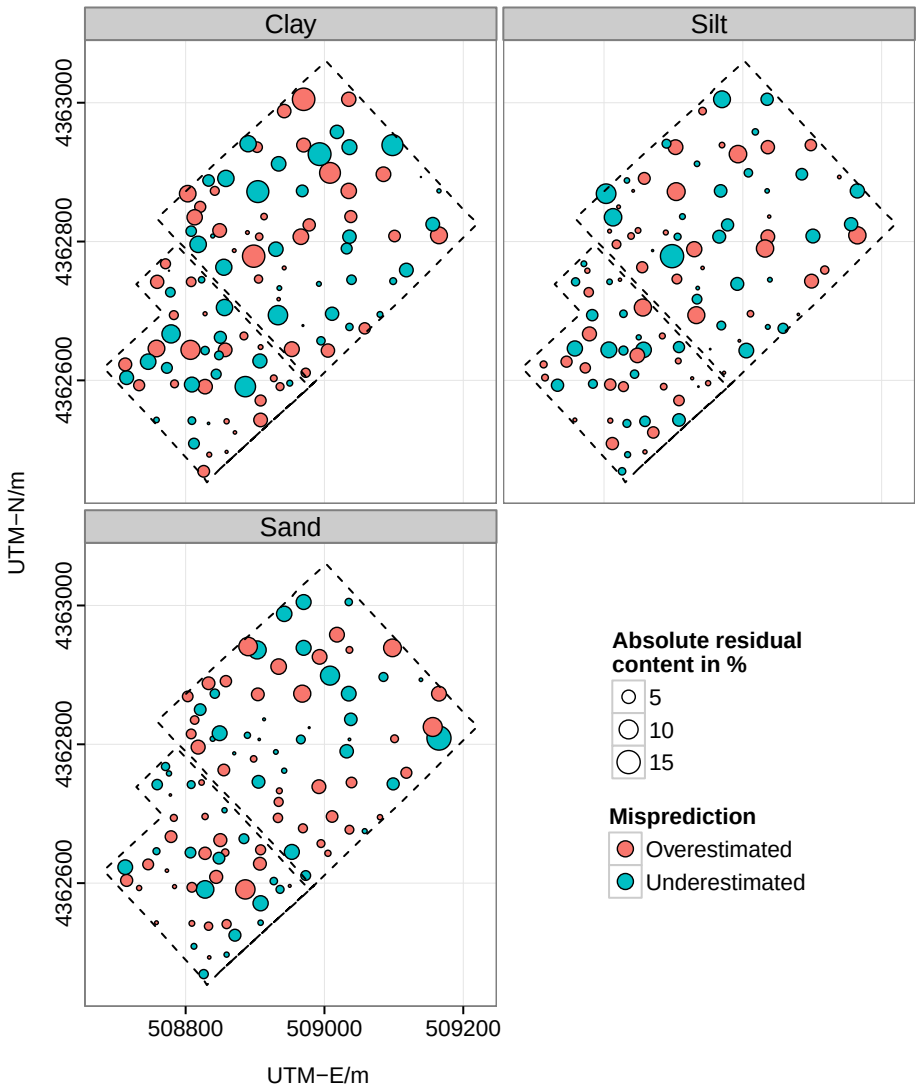


Figure 5.13: Bubble plots of regression cokriging residuals at the field scale

The latter finding corresponds to implications drawn from bubble plots, reflecting absolute residual contents by the size of the circles, and highlighting the direction of mismatch by colour. Figure 5.13 presents on average higher absolute cross-validation residuals for clay and silt content at field 33. Nevertheless, no systematic (spatial) misprediction can be observed from either map of the field and target variable combinations. Over- and underestimation seems independent from exact sample positions and, thus, no unpleasant spatial error clustering

exists. However, remarkably large oppositely coloured adjacent cross-validation results occur, for instance, close to the eastern corner of field 33. Since these two points are also rather isolated in space, their direct vicinity should be considered for any possible supplementary survey. The influential points 3 and 17 as identified based on Cook's distance measures from regression diagnostics at field 21 exhibit the largest absolute residual contents in the sand-related bubble plot. Regression cokriging models would certainly benefit from excluding these samples from calculation. However, this is not justified by any evident reason for incorrect sampling or erroneous lab-analysis.

5.1.5 Interim conclusion

Exploratory (spatial) data analysis revealed remarkable differences between the spatial distribution of soil textural fractions at fields 21 and 33. Regardless of geographical proximity, a fusion of both datasets was not advisable and the two adjacent fields were treated independently during subsequent modelling steps. Covariates from three different sources were considered as auxiliary information in field-scale soil texture mapping. However, due to significant multicollinearity, final predictors were derived from principal component analysis. Global spatial structures regarding clay, silt and sand content were tackled by stepwise linear regression procedures. In case of field 21, PC1 as well as the x- and y-coordinates remained in the final regression model. The best subset of predictors for modelling alr-transformed clay content at field 33 contained the first three PCs, whereas alr Silt was most accurately addressed considering PC1 and PC2 only. Examination of regression diagnostics led to the removal of three highly influential samples from interpolation data at field 33. Final regression parameter estimates were then developed using GLS regression. After trend removal, remaining spatial variation among the stationary regression residuals was modelled using cokriging, based on linear models of coregionalisation. Cross-validation of finally summed results suggest that the regression cokriging model predicts soil textural fractions well at field 21, but exhibit comparatively low performance values with regards to clay and silt content at field 33.

5.2 Interpolation results at the landscape scale: Rio di Costara catchment

Following an exploratory data analysis, results from neural network training and subsequent geostatistical modelling of residuals are of particular interest in this section. Unlike at the field scale, spatial prediction in the Rio di Costara catchment centres on methodological innovations. For this reason, different soil spatial interpolation models are compared to each other after mapping results from neural network residual cokriging (NNRCK) are shown. Model evaluation and comparison uses validation measures calculated from an independent test set.

5.2.1 Exploratory data analysis

Exploratory data analysis at the landscape scale investigates not only the complete sample of 197 point measurements but also the interpolation and validation subsets. Furthermore, it takes into account peculiarities due to different geological formations that are predominant at the given sample locations.

Table 5.11: Summary statistics of soil textural fractions, Rio di Costara

Targets[1]	Min	Max	1st[2]	2nd[2]	3rd[2]	Mean	Std.[3]	Skewness	Kurtosis
- for all samples (N = 197) -									
Clay	5.20	63.19	23.79	33.12	39.78	32.33	10.60	−0.05	−0.34
Silt	6.16	50.23	22.19	26.62	30.64	26.35	6.68	0.10	1.17
Sand	8.73	88.64	33.35	39.60	49.21	41.32	12.22	0.86	1.65
- for interpolation data (N = 156) -									
Clay	7.03	63.19	24.59	33.13	39.72	32.46	10.13	−0.02	−0.23
Silt	8.51	50.23	22.77	26.92	30.73	26.68	6.45	0.25	1.41
Sand	8.73	84.46	33.27	39.63	47.49	40.86	11.33	0.73	1.40
- for test data (N = 41) -									
Clay	5.20	57.33	21.82	33.12	42.01	31.84	12.35	−0.09	−0.87
Silt	6.16	42.59	20.87	25.70	28.57	25.09	7.46	−0.13	0.12
Sand	14.72	88.64	34.61	39.13	51.78	43.07	15.15	0.85	0.87

[1] Clay, silt and sand content in %, [2] Quartiles (2nd = Median), [3] Standard deviation

Table 5.11 summarises statistical properties of raw target variables for all samples, calibration and test data. Taking the whole dataset into account, sand content ranges up to 88.6 %, which is by far the highest percentage among the three particle-size fractions that occur in the data. Sand is also predominant on average with a mean value of 41.3 % and a median of 39.6 %. It varies more than the other fractions with a standard deviation of 12.2 % and exhibits some slightly right-skewed behaviour indicated by a positive skewness value close to 1. By contrast, the distributions of clay and silt content are rather symmetric. Silt content varies much less in the given catchment than clay and sand content with a standard deviation of 6.7 %.

Comparing the summary statistics for interpolation and test data reveals very similar proportions among the two datasets. Regarding the mean, test data is slightly lower on average for clay and silt content compared to the interpolation set, but a little higher with respect to sand. All three target variables tend to vary slightly stronger among the test points indicated by higher standard deviations. Note, however, the different numbers of observations in the two distinct samples. A Hotelling T^2-test statistic of 2.3 (df1 = 3, df2 = 193) and an associated p-value of 0.08 imply that the null hypothesis of equal multivariate means should not be rejected on the 0.05 significance level for the two samples. With regards to the variance-covariance matrices, Bartlett's test statistic of 6.4 (df = 6) with a p-value of 0.38 clearly indicates variance homogeneity among the two datasets. Thus, considering the relatively small differences in summary statistics and the results from two statistical tests it can be said that the sampling design described in section 4.2 led to representative calibration and test sets.

Similar conclusions about the distributions of the target variables, previously discussed in terms of numerical summaries, can be drawn from histograms displayed in figure 5.14. The sand content is slightly skewed towards higher values, whereas clay exhibits some bimodal behaviour especially in the test dataset. There are basically two peaks in the clay data at around 20 % and just below 40 %, both of which strongly correspond to median values of clay samples taken from Paleozoic basement areas (19.8 %) and quaternary deposits (36.9 %), respectively. Thus, the bimodal behaviour in clay can be explained to a large extent by different

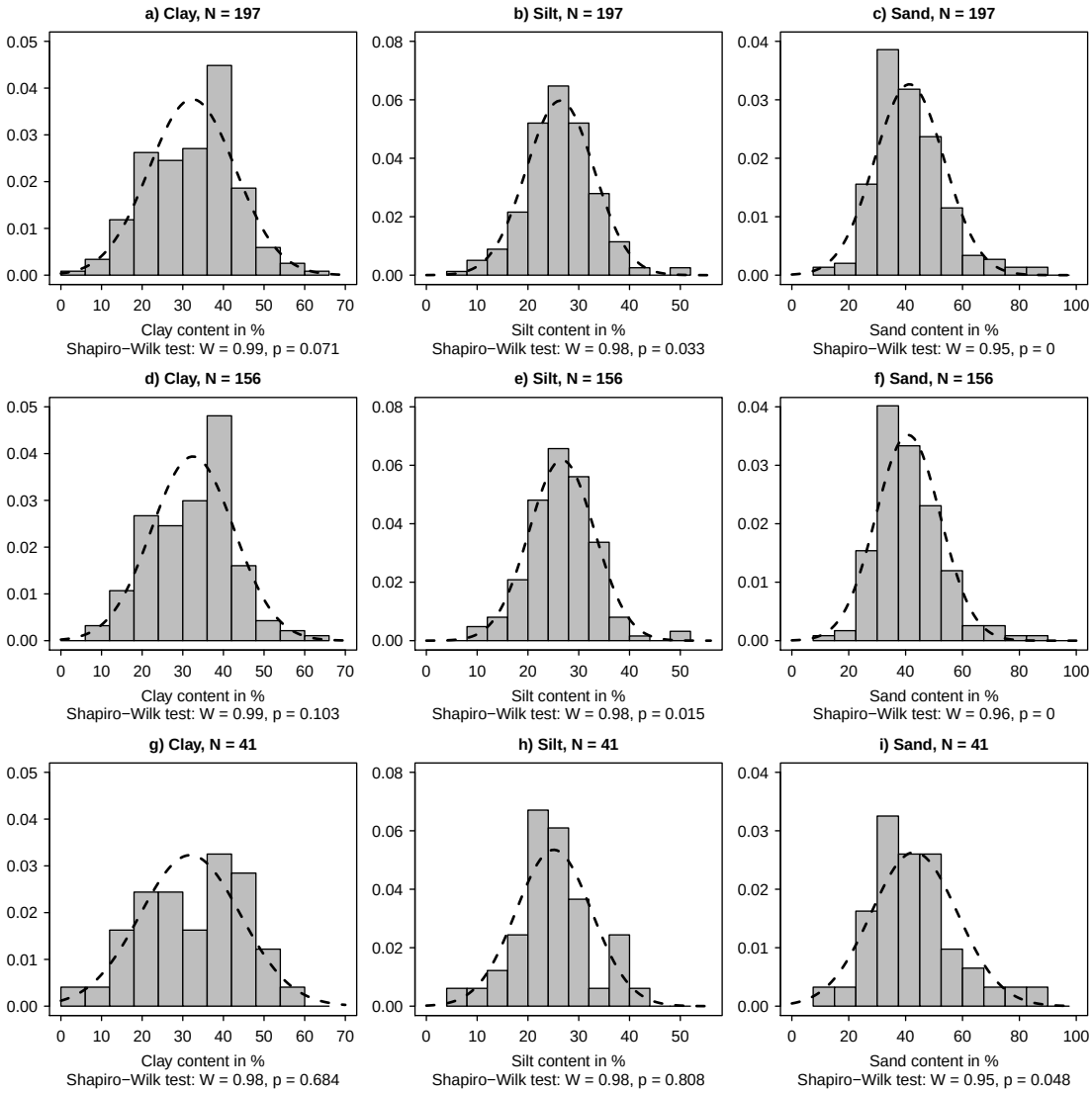


Figure 5.14: Density histograms of soil textural fractions, Rio di Costara catchment: a) – c) for all soil samples, d) – f) for interpolation data, g) – i) for test data. The dashed lines represent the corresponding normal distribution based on sample mean values and standard deviations.

geological units inside the given catchment. With regard to statistical modelling, however, this bimodal distribution is not relevant, as it diminishes after additive log-ratio (alr) transformation. Figure 5.15 illustrates the distributions of these alr coordinates focusing on the calibration data.

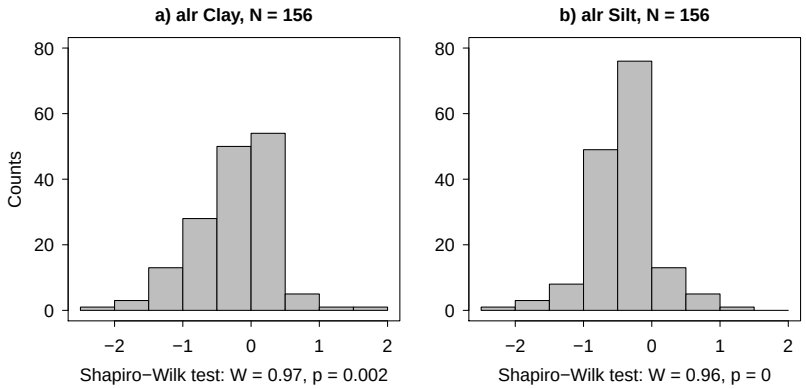


Figure 5.15: Histograms of alr-transformed clay and silt content

The grouped boxplots shown in figure 5.16 reveal considerable discrepancies between geological units with regard to clay and sand content. The highest sand values with a median of 51.1 % are found in soils developed from Paleozoic rocks, which are mainly characterised by micaceous and quartz-rich metasandstones (refer to section 3.2 for details on the geological setting of Sardinia). The lowest sand values with a median value of 36.8 % occur among the more recent alluvial deposits. It is also noteworthy that asymmetry in sand changes between Oligo-Miocene sediments, on the one hand, as well as Paleozoic rocks and quaternary deposits, on the other. The first is right-skewed with skewness of 1, whereas the latter two are more oriented to lower sand contents with skewness parameter of -0.8 and -0.13, respectively. Regarding potential outliers, sample point 735 is strikingly suspicious. At that particular site, the highest clay percentage as well as the lowest sand content value are measured from quaternary deposits. As a whole, the given observations indicate an apparent influence of geological differences on variations of soil texture in the Rio di Costara catchment. It seems, therefore, reasonable to incorporate geology as an explaining factor into spatial interpolation procedures discussed further on. Note, however, that silt content is not greatly affected by geological differences.

So far, each (target) variable is presented individually, ignoring the fact that they are parts of a whole or more precisely fractions of the soil texture. The ternary plot in figure 5.17 represents the three-part simplex of (compositional)

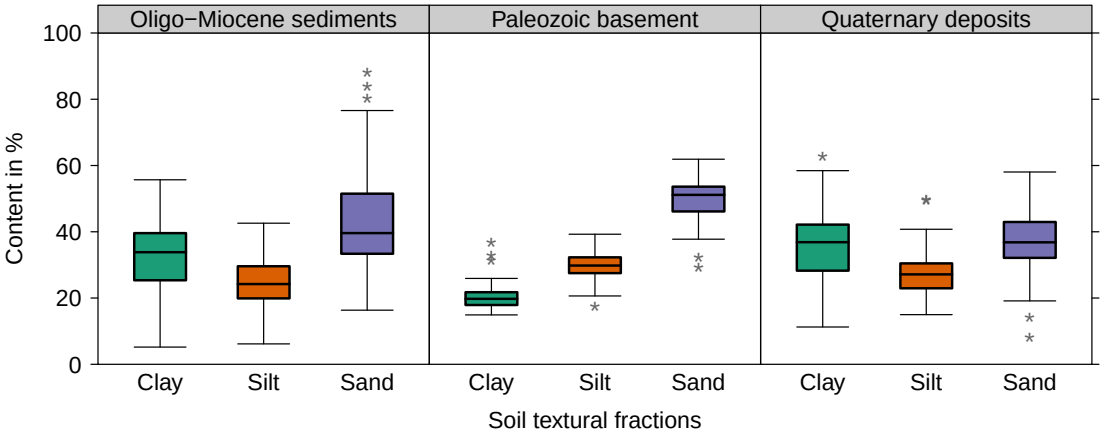


Figure 5.16: Contents of soil textural fractions by main geological units in the Rio di Costara catchment. For convenience, metamorphic and granitic rocks are grouped into Paleozoic basement. The number of observations per group are 77, 28 and 92 from left to right. Potential outliers are marked by an asterisk and represent values outside 1.5 times the interquartile range.

soil separates in form of an equilateral triangle visualising fine earth fractions at all 197 sample locations. The centre of the given composition, with clay, silt and sand contents of 31.8 %, 26.7 % and 41.5 %, respectively, is located inside the clay loam class as defined by the USDA soil taxonomy and FAO guidelines. Clay loam also dominates the investigated sites with respect to absolute frequency. 68 points of the given sample lie within that particular texture class. Interesting with regard to spatial interpolation, however, are implications drawn from the contours of compositional Mahalanobis distances added to the given ternary plot. The majority of data points are cumulated near the centre of the simplex and, thus, are far away from the vertices where distortion is most severe. Consequently, only a few observations of this specific sample (those with high sand content) are considerably affected by the compositional constraints.

After focusing on the distribution of single target variables, bivariate relations between soil textural fractions and terrain attributes are discussed. With respect to Pearson's correlation coefficients (ρ) listed in the lower panel of figure 5.18, clay content of the Rio di Costara test site is significantly ($P < 0.001$) related to elevation ($\rho = -0.41$), slope ($\rho = -0.28$) and SAGA wetness index ($\rho = 0.30$).

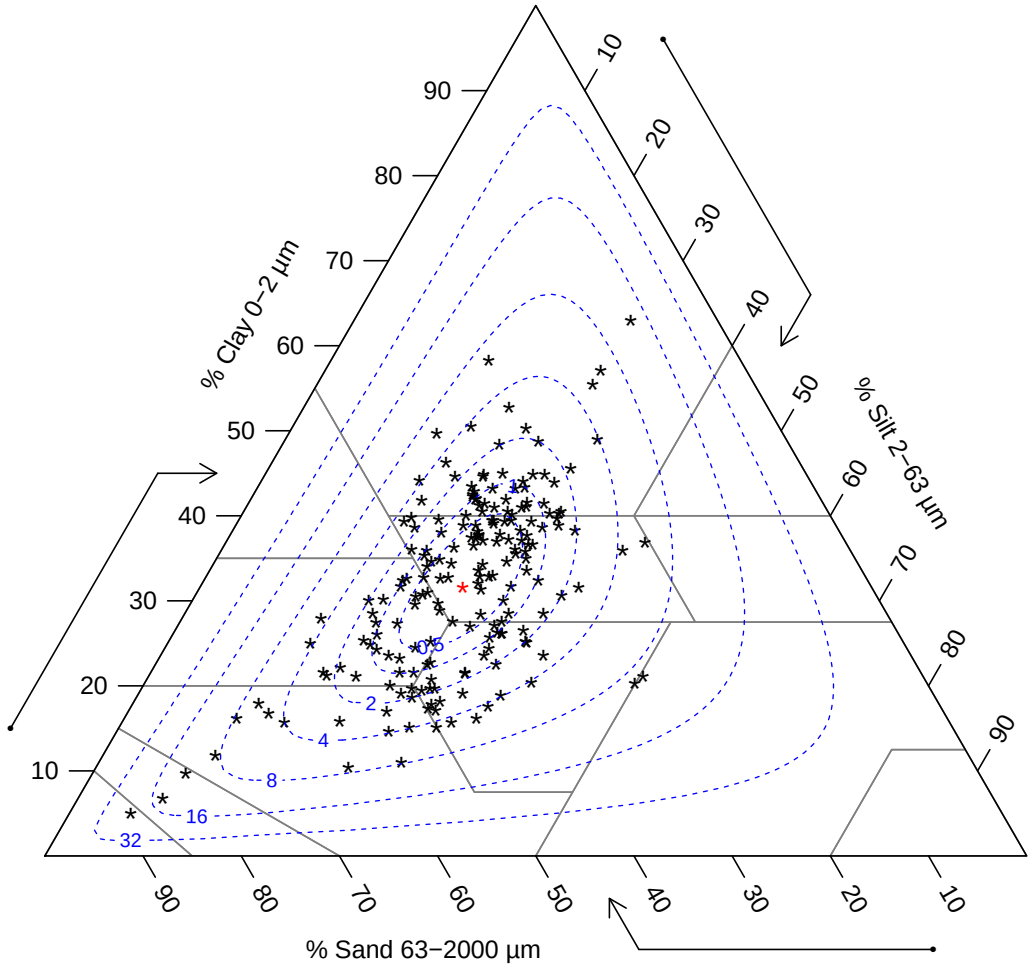


Figure 5.17: Ternary diagram of sampled soil texture classes at the landscape scale, Rio di Costara. The grey lines indicate the transition between different textural classes in accordance with the USDA soil taxonomy (Soil Survey Staff 1999) and the FAO guidelines for soil description (FAO 2006). The red-coloured point represents the centre of the given composition. Contours of equal compositional Mahalanobis distance (up to 32) from the mean are added as blue-coloured dashed lines.

The opposed signs for elevation and wetness index may be explained by erosional processes. Finer particles are moved downslope by surface runoff and accumulate at lower positions where soil moisture is usually higher. This observation corresponds to reverse conditions found for sand content being the textural fraction that remains at higher elevation after removal of finer particles. Silt content,

however, is markedly less related to terrain attributes in the present study. At least slight correlations are obtained between soil textural fractions and potential incoming solar radiation. It was, therefore, reasonable to choose solar radiation and wetness index as stratifying variables for soil sampling (see section 4.2 on

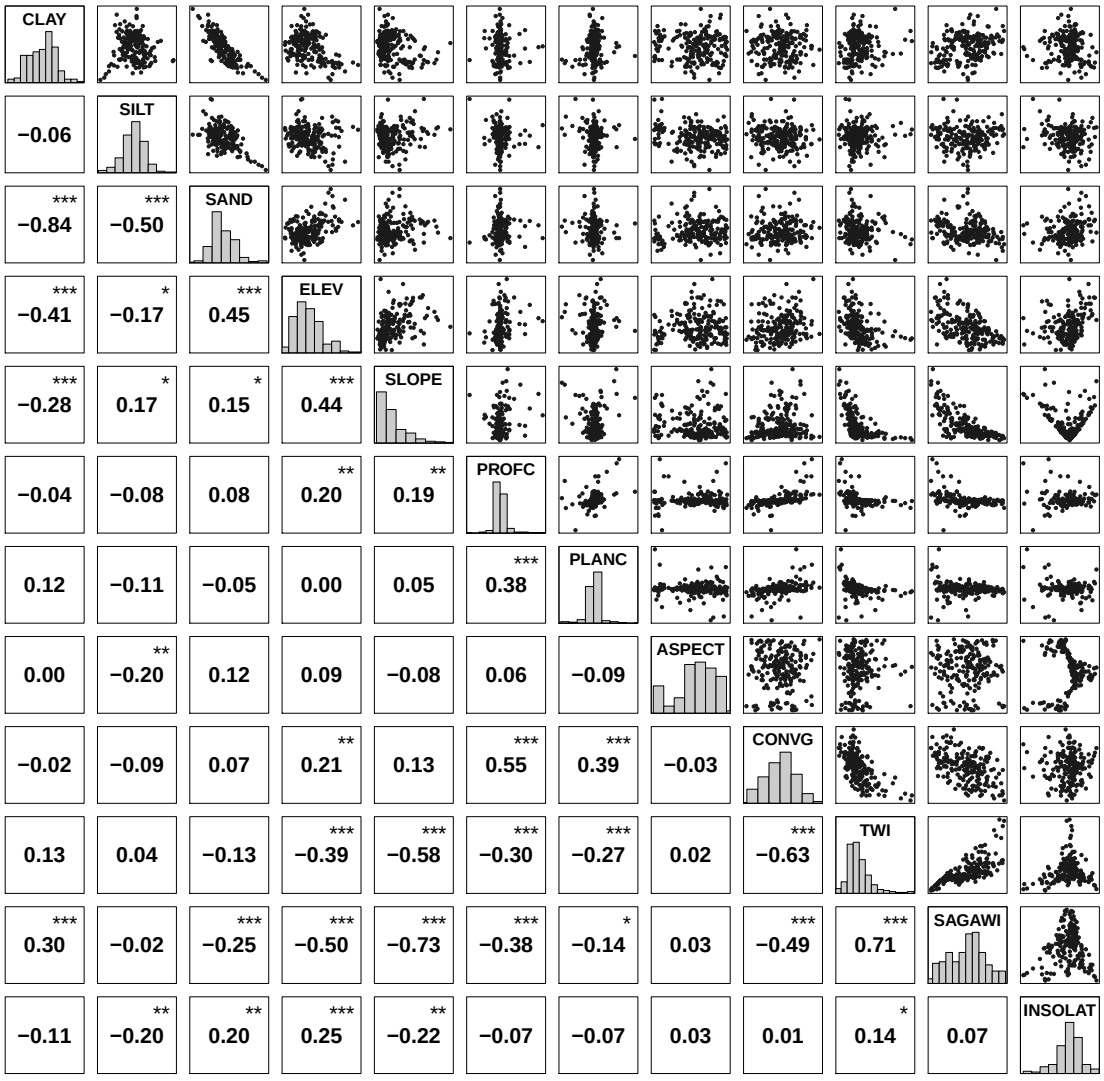


Figure 5.18: Scatterplot matrix, Rio di Costara. The lower triangle displays Pearson's correlation coefficients of each pairwise variable combination. Statistical significance is indicated as follows: * $P < 0.05$, ** $P < 0.01$, *** $P < 0.001$. Diagonal boxes show histograms of each parameter (for full names, see table 4.1 on page 39) and bivariate scatterplots are drawn in the upper panel.

page 30). However, from originally 13 land-surface parameters listed in table 4.1 on page 39, length-slope factor, direct and diffuse solar radiation are omitted from figure 5.18 because they almost perfectly correlated to slope, potential incoming solar radiation and elevation, respectively. Additionally, stream power index exhibits a statistically inconvenient distribution and is excluded from further analysis due to extremal values.

For variable selection with regard to spatial prediction of soil textural fractions at the landscape scale, correlation analysis was repeated using alr-transformed clay and silt content. Land-surface parameters are excluded if they exceed an arbitrarily chosen critical correlation level of 0.65 among each other. Thus, for instance, slope is selected for removal due to its proximity to wetness indices. Elevation, SAGA wetness index and potential incoming solar radiation are finally selected as continuous predictors.

Until now, data analysis focused on feature space and geographic location played no role in the tools that were used. Nevertheless, knowledge about the spatial structure of a dataset is of utmost importance for decision making in the context of (geo)statistical modelling. Proportional symbol postplots are particular useful to examine global trends in the variables of interest across an entire region. Figure 5.19 exhibits some clear tendency towards lower clay contents in the very North of the Rio di Costara catchment. These low values occur almost entirely inside an area dominated by Paleozoic rocks and, thus, correspond to implications drawn from the boxplots in figure 5.16 on page 119. Very high sand content is concentrated in the north-eastern boundary region which is associated with Oligo-Miocene sediments. The spatial distribution of silt content is more even, but shows some clustering of high values at the very south where quaternary deposits are located. It is obvious from figure 5.19 that the assumption of a constant (spatial) mean, inherent in most (geo)statistical techniques, is implausible in the given case study. However, the examined regional trend is strongly associated with geological distinctions. Next to global issues, proportional symbol postplots allow for the visual detection of local anomalies. The labelled points (IDs = 735 and 511) in the clay content plot of figure 5.19 are such outliers compared to their nearest neighbours.

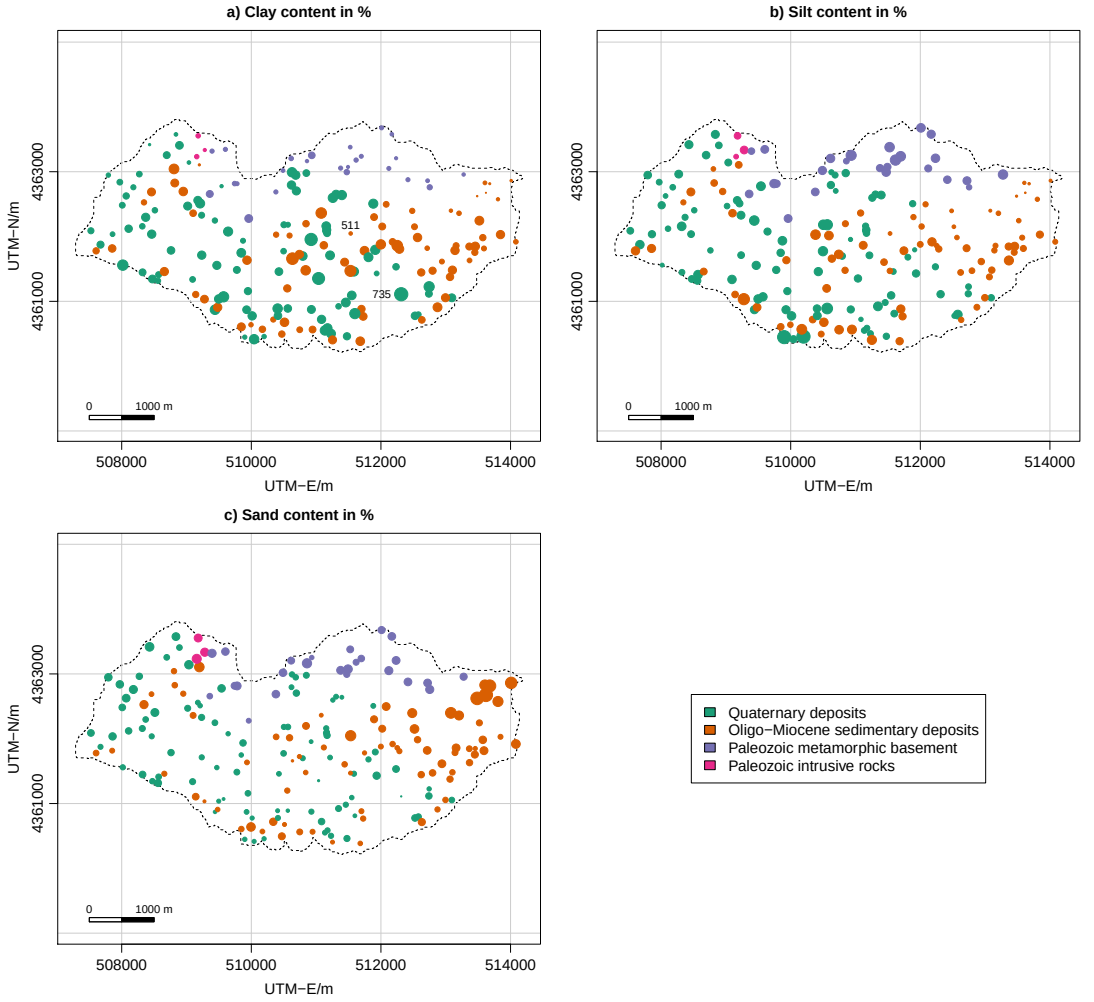


Figure 5.19: Postplots of measured clay, silt and sand content by geological unit in the Rio di Costara catchment. The size of the circles represents the proportion of the respective textural fractions.

Another important aspect of spatial structure is local spatial dependence which is independent from absolute locations and best investigated using variograms. Empirical variances plotted against distance, as done in figure 5.20, prove the existence of spatial auto-correlation for all three textural fractions in the given test site. Differences between point pairs close to each other are on average lower than those for higher separation distances. For clay content, two spherical models are fitted to lags 465 m and 4620 m, respectively. There are obviously two scales

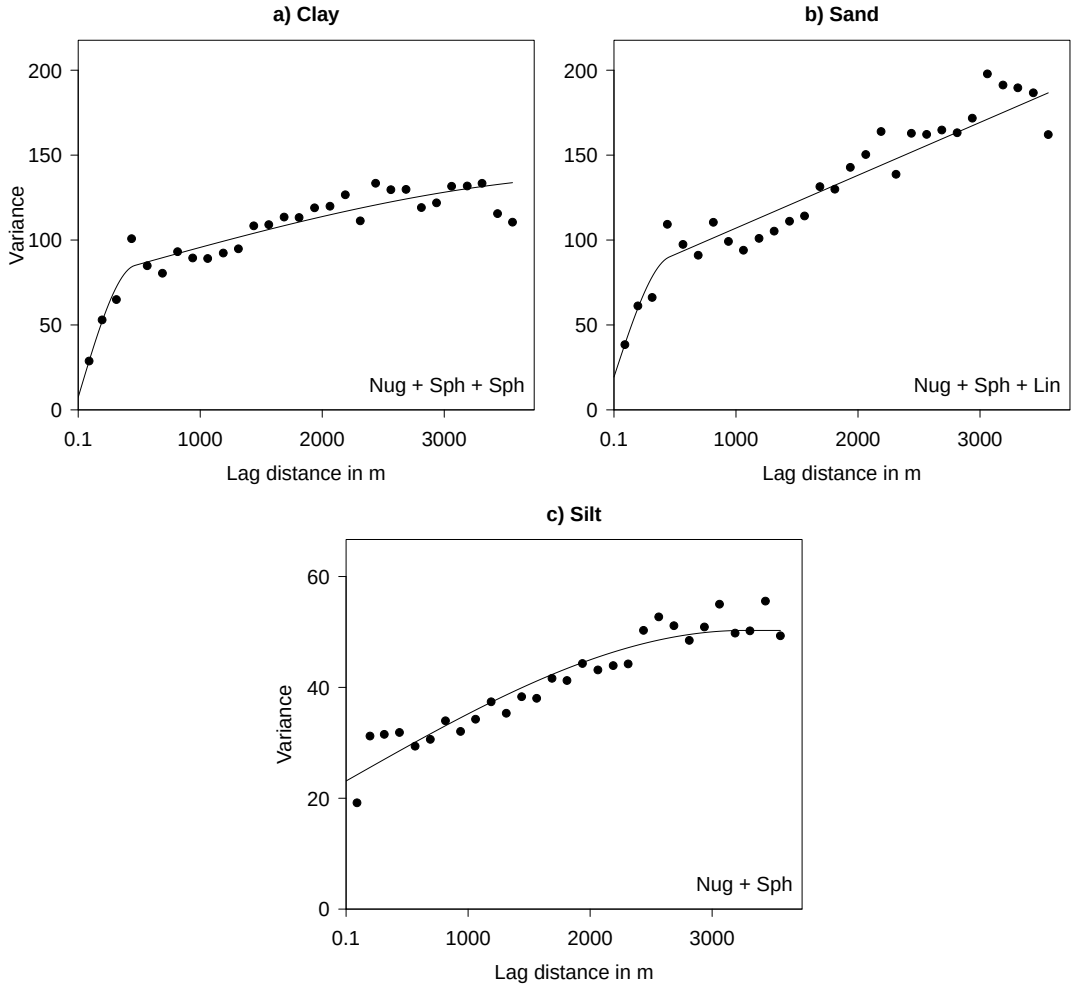


Figure 5.20: Variograms of soil textural fractions at the landscape scale. Empirical variances are estimated by the method-of-moments and fitted using weighted least squares (N/h^2). The sample size is 197, the number of point pairs inside lag 1 are 53 and the length of the diagonal of the box spanning the data is 7337 m.

of spatial variation, a short-range component with rapid increase of variance up to a (partial) sill of approximately 75 $\%^2$ and a gradual long-range trend converging to a total sill of about 140 $\%^2$. The nugget variance is 7 $\%^2$. Sand content exhibits a very similar nested structure as seen for clay, but with a more sharp increase after reaching a partial sill of 76 $\%^2$ at 450 m and a slightly higher nugget effect (20 $\%^2$). The third component is also clearly unbounded and better represented

by a linear model. Silt, however, is much more difficult to model especially at small distances. Variance moderately increases up to 50 %^{2} (of which the nugget value is 23 %^{2}) at a distance of 3260 m. Note that the overall sill variance derived from variogram modelling is almost three times higher for clay compared to silt, which also roughly corresponds to their observed variances.

Figure 5.21 displays variograms for alr-transformed clay and silt content of the calibration data from which spatial prediction models are developed. They appear clearly unbounded, indicating a non-stationary process. As a consequence, some de-trending is required before traditional (geo)statistical techniques can be applied in a reasonable manner.

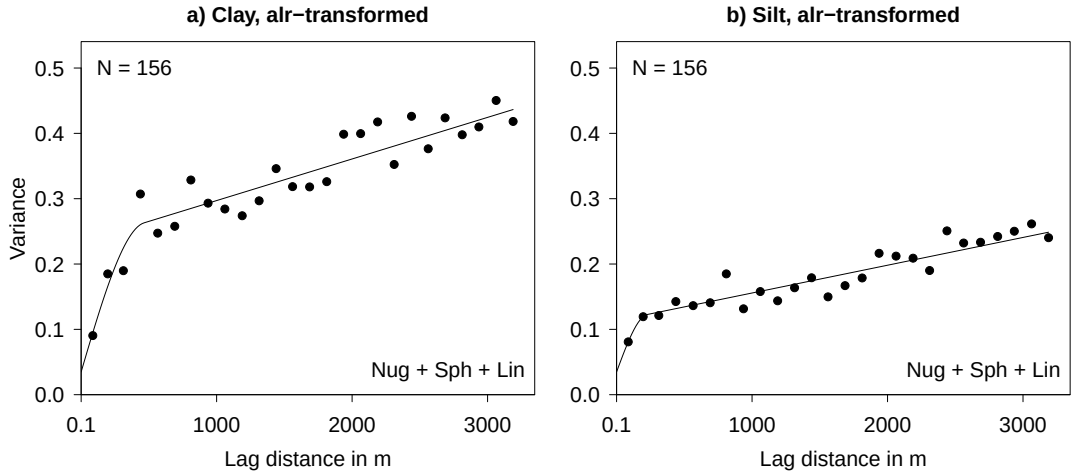


Figure 5.21: Variograms of alr-transformed clay and silt content at the landscape scale. Empirical variances are estimated by the method-of-moments. Nugget and spherical models are fitted using a) weighted least squares (N/h^2) or b) adjusted by eye. The number of point pairs inside lag 1 are 34.

5.2.2 Neural network training and estimation

The exploratory (spatial) data analysis section closed with the result that the process under study is clearly non-stationary. As a consequence, some de-trending component must be integrated into the spatial prediction procedure to produce

high-resolution maps of clay, silt and sand content in the Rio di Costara catchment. From correlation analysis, three land-surface parameter (elevation, SAGA wetness index and potential incoming solar radiation) were identified as useful explanatory factors for the given interpolation problem. Nevertheless, bivariate scatterplots revealed fairly non-linear relationships between target quantities and relevant environmental predictors. This non-linearity rather motivated the use of a neural network (NN), instead of common multiple linear regressions (MLRs) to address any large-scale variability among the dependent variables. A multi-layer perceptron (MLP) is modelled using the resilient back-propagation (Rprop) learning rule, applied to a sum-of-squares error function including some weight-decay term. The subsequent paragraphs present the results with regards to NN/MLP training and estimation.

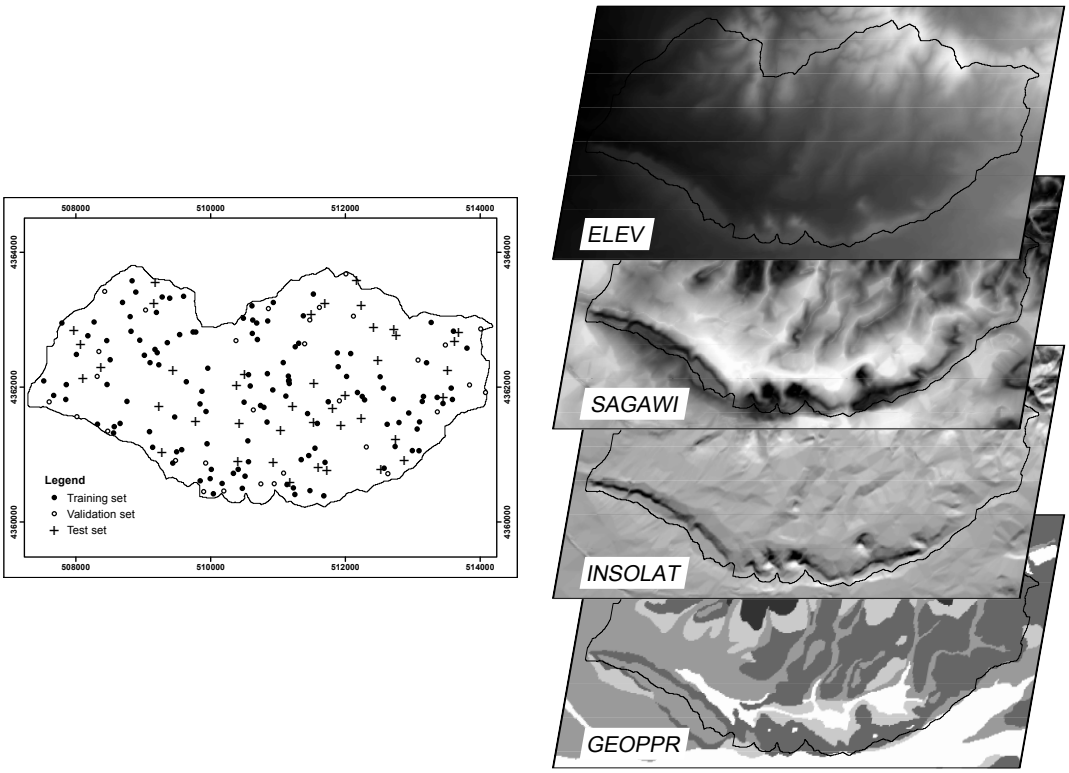


Figure 5.22: Soil spatial interpolation data and selected predictor maps at the Rio di Costara test site. Sample locations are subdivided into training, validation and test data (left) and maps of finally selected input variables for neural network modelling (right).

In addition to three selected terrain-related continuous predictors, encoded geological units were considered as neural network input. From originally six dummy variables, two (18 and 1233) were excluded as they consist of very few samples that actually fall into the corresponding geological classes leading to near-zero variance predictors. Together with measured alr-transformed clay and silt content, this set of (standardised) predictors sets the stage for MLP prediction in the present study. However, training the MLP model requires some more selections that need to be based on extra data. For this reason, three distinct data sets were produced in advance: a training set (N = 125) for finding the optimal network weights, a validation set (N = 31) to determine appropriate numbers of iteration and hidden units, and test data (N = 41) to later on evaluate the generalisation ability of the neural network residual cokriging model (see figure 5.22).

Table 5.12: Model performance of different neural network architectures

	Number of hidden units					
Number of training cycles	5	7	9	11	13	15
20	35.45	34.75	41.92	35.75	43.44	45.14
40	31.28	32.75	30.83	32.58	35.16	42.31
60	31.28	32.06	34.04	32.61	36.01	42.73

Average sum-of-squares errors (SSE) calculated from independent validation data (N = 31)

Table 5.12 summarises the average prediction errors of the validation set with regards to different combinations of iteration and hidden unit number. The latter is an important parameter to adjust, since it substantially defines the complexity of the resulting MLP model. Several networks are trained with varying numbers of hidden units selecting the best ones in terms of a prediction accuracy measure. The lowest rated, and thus, most promising models were obtained using 40 training cycles and nine hidden units. From ten models trained with this particular combination, the one with the minimum sum-of-squares error is considered for prediction purposes. Similarly, few hidden units were used, for instance, in climatic data applications by Demyanov et al. (1998), radioactive soil contamination studies by Kanevski et al. (2004) and wind speed estimation problems by Cellura et al. (2008) who created neural network models with 8, 5 and up to 15 hidden units, respectively.

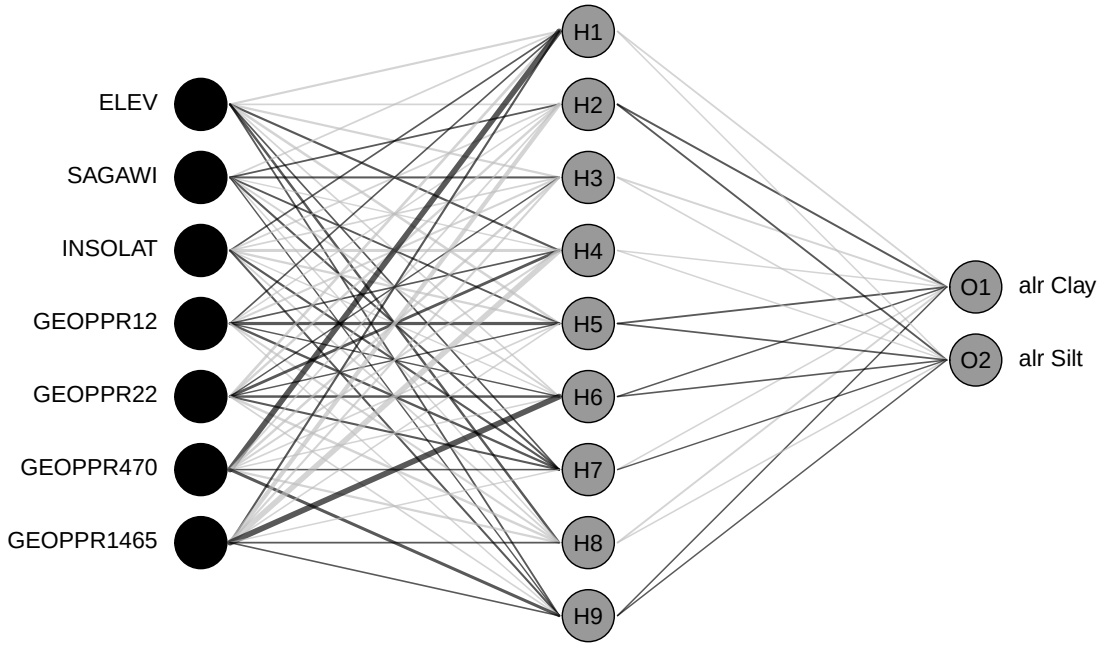


Figure 5.23: Neural interpretation diagram of data-driven trend estimation for soil spatial prediction at the landscape scale. Line width is proportional to the magnitude of each individual connection weight. Black-coloured connections have a positive sign and gray-coloured links are negative. Bias units are not shown.

The predominant objective of neural network applications to real-world problems is prediction. Due to their complex structure and demanding computation, NNs are often regarded as black box models (Rojas 1996). In other words, NNs are very flexible approximators that do not necessarily rely on restrictive assumptions, but provide little insight into the underlying system as, for instance, linear regression does in form of well-established summary statistics and diagnostics. Nevertheless, it is sometimes desirable to take a look behind the scenes of the trained NN/MLP network. This is done here by studying the MLP connection weights.

Figure 5.23 displays the final MLP architecture with seven input, nine hidden and two output layers fully connected from left to right. In addition, the shown neural interpretation diagram (NID) provides some qualitative information about the strength and direction of any connection weight. Focusing on the sign of the links roughly indicates, whether a certain input variable affects an output in a positive

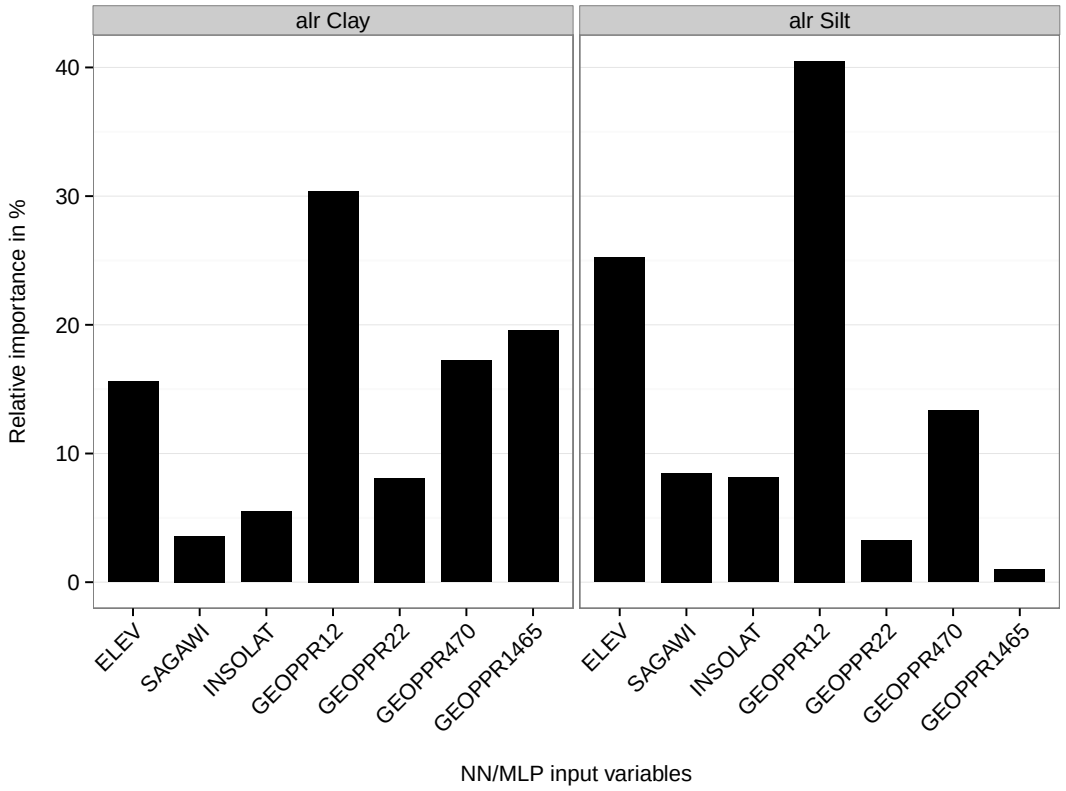


Figure 5.24: Bar plots of relative importance of input variables on neural network model output according to Garson's algorithm (Garson 1991)

or negative way. If both, input-hidden and hidden-output layer connections are of the same sign, the overall effect of a single input on the output is said to be positive. This is true, for instance, regarding the majority of hidden units related to alr Clay and SAGA wetness index. The interpreted positive relationship seems reasonable, since fine-earth particles more likely accumulate at lower positions where soil moisture is often higher. On the contrary, in cases with opposing signs between input-hidden and hidden-output layer connections, a negative influence of the studied input on the corresponding output is deduced. An example for such a relationship is given by elevation and alr Clay. Again, this observation from the trained network is in accordance with natural process understanding, since higher clay contents are usually expected at lower positions where fine soil material is accumulated by erosional processes. Additionally, interactions among

predictors can be detected from the NID, when two inputs are connected to the same hidden unit with opposing signs as observed here for elevation and SAGA wetness index.

The interpretation of the NID is particularly demanding in case studies with numerous connections. Besides network complexity, implications from figure 5.23 are hampered by the fact that input variables entered the network after standardisation and target quantities were considered as alr coordinates. Garson (1991) suggests an alternative method to determine a more convenient measure for demonstrating the relative importance of input variables on each individual network output. The degree of relative importance is determined in percent and its computation based on partitioning the connection weights (see Gevrey et al. 2003, Appx. 1). Figure 5.24 visualises the resulting numbers in terms of simple bar plots. Obviously, eluvial-colluvial deposits (GEOPPR12) contribute most intensively to alr-transformed clay and silt contents in the given MLP model with relative importance values of 30 % and 40 %, respectively. In addition, elevation is highly influential on both studied outputs. Geological units play an important role especially regarding alr-transformed clay content. However, they appear less relevant for alr Silt. This observation corresponds to results from exploratory data analysis (refer to the grouped boxplots in figure 5.16 on page 119). Note that Garson's algorithm operates on absolute weight values and, thus, provides no information on the direction of relationships. Nevertheless, in conjunction with NIDs, relative importance bar plots provide some basic impressions about the inherent mechanisms of neural networks used for prediction.

5.2.3 Geostatistical analysis of neural network residuals

The successfully trained neural network focuses on modelling long-range variability among the alr-transformed target quantities using secondary information (land-surface parameters, geological map). Following the MLP estimation, model residuals are analysed for remaining spatial auto-correlation. This is done by computing Moran's I test statistics from a spatial weights matrix based on eight nearest neighbours. Resulting p-values of 0.049 for alr-transformed clay content

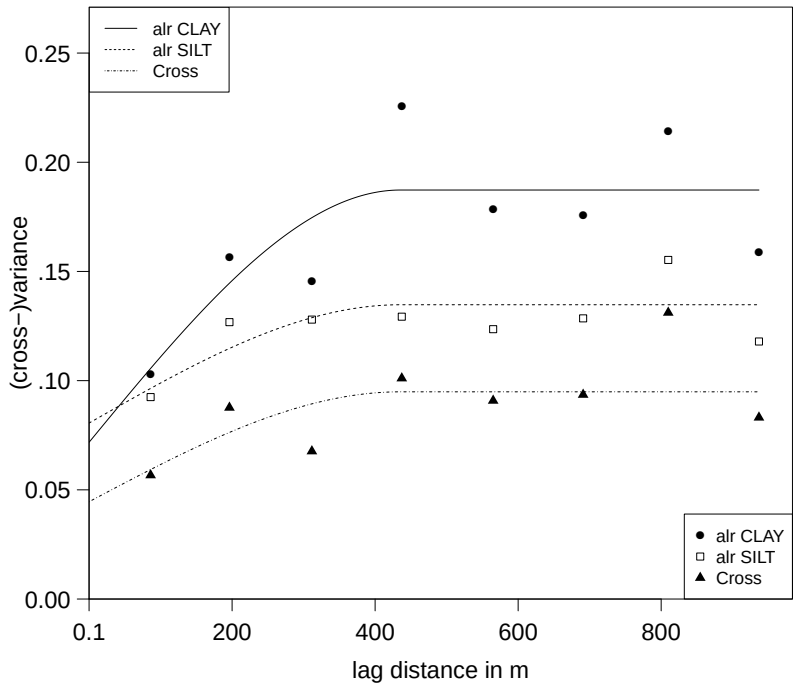


Figure 5.25: Residual (cross-)variograms of alr-transformed target quantities in the Rio di Costara catchment. The auto- and cross-(semi)variances come from method-of-moments estimation with a linear model of coregionalisation fitted by (weighted) least squares including a subsequent check for positive-definiteness.

and 0.032 regarding alr Silt are just below the 0.05 significance level indicating that there is still a certain amount of local spatial dependence among the MLP residuals. Thus, subsequent kriging-interpolation seems reasonable. A Pearson's correlation coefficient of 0.62 between the two fitted NN outputs motivate the use of cokriging rather than kriging each residual set individually. Consequently, this section presents the (cross-)variogram modelling results based on the linear model of coregionalisation (LMC). Note that the geostatistical analysis of NN/MLP residuals operates on the reunited calibration set (156 samples).

The empirical auto- and cross-variograms computed from MLP residuals shown in figure 5.25 strongly suggest to choose a spherical model with nugget effect for characterising any present local spatial dependence. The estimated variances

linearly increase up to 435 m and clearly converge to a sill. The fact, that the residual (cross-)variograms are bounded contradicts to observations taken from variography based on the original targets (see figure 5.21 on page 125). This is a strong indication that the trend-removal strategy using neural network models was successful in the given case study.

Table 5.13: Residual (cross-)variogram characteristics, Rio di Costara test site

Parameter	Model	Fit. method	Nugget	p. sill	Range	np[1]	NSR
alr Clay	Sph	LMC by WLS	0.0718	0.1155	435.2	34	38.3
alr Silt	Sph	LMC by WLS	0.0806	0.0542	435.2	34	59.8
Cross	Sph	LMC by WLS	0.0446	0.0503	435.2	68	-

Sph = Spherical model, LMC = Linear model of coregionalization, WLS = Weighted least squares

p. sill = partial sill, np[1] = number of point pairs within the first lag bin, NSR = Nugget-to-Sill-Ratio in %

Table 5.13 numerically underlines the graphical interpretations by listing important (cross-)variogram characteristics. Nugget-to-Sill-Ratios of 38 % and 60 % indicate some moderate auto-correlation behaviour for both NN output residuals. However, there is less spatial dependence observed in the silt-related variable. Thus, the LMC fits well the empirical variances and clearly reflects some short-range variability among the alr-transformed target residuals, while any large-scale variation is already addressed by the previously fitted MLP model.

5.2.4 Final prediction and model validation

After modelling the long-range component of spatial variation among the studied soil textural fractions using a neural network and geostatistically interpolating the residuals, final predictions at unvisited locations are obtained from simply adding up both parts. Back-transformed results of estimated clay, silt and sand content are presented in figure 5.26 and appear plausible with regards to both, general process knowledge and case-specific conclusions drawn from exploratory (spatial) data analysis. For instance, consistently less clay content is predicted within in the mountainous northern part of the Rio di Costara catchment. In addition to erosional processes, these Paleozoic hills are made of micaceous and quartz-rich metasandstones which likely weather into more sandy soil material.

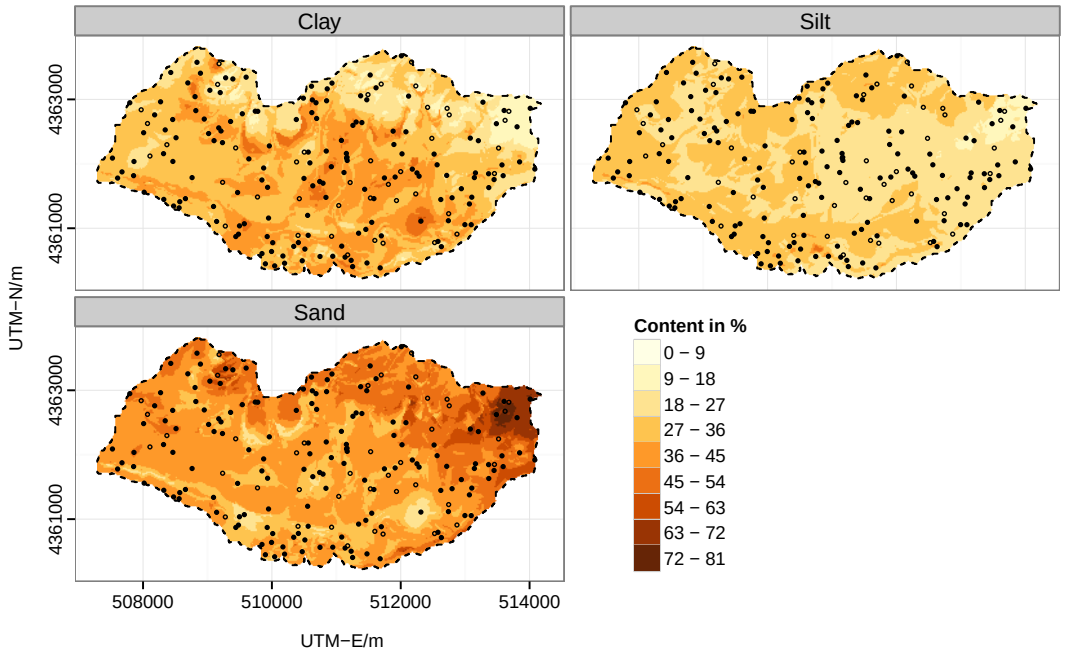


Figure 5.26: Neural network residual cokriging estimates of soil textural fractions at the landscape scale. The prediction results are shown after additive generalised logistic back-transform. Black-coloured dots represent sample locations used for calibration, while hollow data points refer to independent test data.

Even higher predicted sand contents occur in the very north-eastern corner of the study area, where sandy-conglomerate alternations from the Nurallao formation combined with higher elevations are predominant. Again, greater sand content is estimated for the relatively small intercalation of granitic material in the north-western edge of the catchment. Another remarkable feature is that more clay content, and contrarily, less sandy material occur in the eluvial-colluvial belt between the (northern) mountains and the adjacent inner-catchment part of the Campidano plain. Note that this geological unit has also been identified as the most contributing factor during neural network modelling. However, while clay and sand content are well-distinguished and often opposed in magnitude due to the compositional character of the three soil textural fractions, the silt prediction map appears far more homogeneous. This impression is not entirely owed to the equal colour scheme used for visualisation, but is consistent with the lower range of that particular fraction reported from initial summary statistics.

Table 5.14: Summary statistics, normality tests and checks for remaining spatial correlation of model residuals at the landscape scale

	Min	Max	Mean	Std.	Skew	W	p	I	z	p
Clay	−18.53	22.96	−1.59	8.28	0.38	0.98	0.566	0.10	1.27	0.203
Silt	−11.33	10.20	0.93	5.43	−0.30	0.97	0.400	0.24	2.69	0.007
Sand	−31.51	21.38	0.66	9.71	−0.52	0.96	0.119	0.00	0.24	0.812

W = Shapiro-Wilk W, I = Global Moran's I based on 4-nearest neighbour weights, Std. = standard deviation, z = z-score

Residual statistics, calculated from a hold-out sample of 41 points, reveals some initial impressions on the overall model quality. A mean error, also known as (statistical) bias, of -1.59 indicates some slight under-estimation of the clay content, provided that the measured value is subtracted from the prediction result during error calculation. On the contrary, silt and sand content are slightly over-estimated on average with mean errors of 0.93 and 0.66, respectively. In addition to rather small biases, the error distributions are mostly symmetric as desired, reported in terms of skewness and significant ($P > 0.05$) Shapiro-Wilk test results. Nevertheless, the largest absolute discrepancy between predicted and measured values is 32 % with regards to sand representing some rather severe mismatch in the context of soil textural fractions.

Moving from average error metrics to individual estimates, scatterplots of actual versus NNRCK-predicted values are shown in figure 5.27. All three slopes of the displayed regressions are greater than 1 indicating that high measured values are (slightly) under-predicted, while low actual clay, silt and sand contents are over-predicted by tendency. This consistent behaviour at the margins of the value ranges basically results from the smoothing property of the krige interpolator and appears to be comparatively weak in the given case study. In fact, the majority of test data points are well-packed around the desired 1:1 lines, suggesting that the NNRCK model effectively predicts the observed target quantities. This is numerically underlined by R-squared values of 0.57, 0.49 and 0.63 for clay, silt and sand content, respectively.

A few individual samples are very poorly matched, of which point 511 is most strikingly incorrect with regards to clay and sand estimations. This becomes particularly obvious within the context of bubble plots, where absolute residual

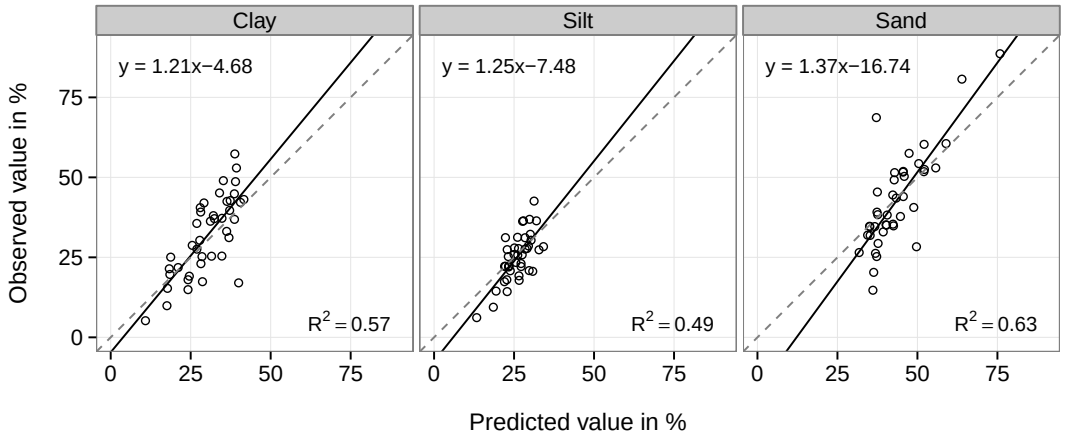


Figure 5.27: Scatterplots of actual versus predicted values in the Rio di Costara catchment. The solid lines represent linear regression models of actual measurements on back-transformed NNRCK estimates at test data locations. The dashed lines are the (desired) 1:1 lines.

contents are reflected by the size of the circles and the direction of mismatch is highlighted by colour. It can be seen in figure 5.28 that the ID-labelled sample 511 shows not only very large absolute errors, but is also contradictory to any neighbouring validation points regarding the sign of the misprediction. In other words, this particular sample is either corrupted or the only representative of a local soil textural anomaly. The latter interpretation is supported by the fact that exploratory data analysis (see postplots in figure 5.19 on page 123) already highlighted the serious combination of unusual observations with rather isolated spatial position for point 511. EDA also identified another unfortunate sample (735) with similar features assigned to the calibration data. Due to the exact nature of the krige interpolator, this calibration sample 735 strongly influences the mismatch-direction of closest validation data points. This can be easily understood, for instance, from the clay-related map of figure 5.28, focusing on two adjacent red-coloured test points located slightly south-east from point 511. The overestimation of these two clay-samples is directly related to calibration point 735 and unusual with respect to the mismatch-sign of their superior neighbourhood. Consequently, it is strongly advised to revisit both sites during any possible subsequent soil surveys in the same region.

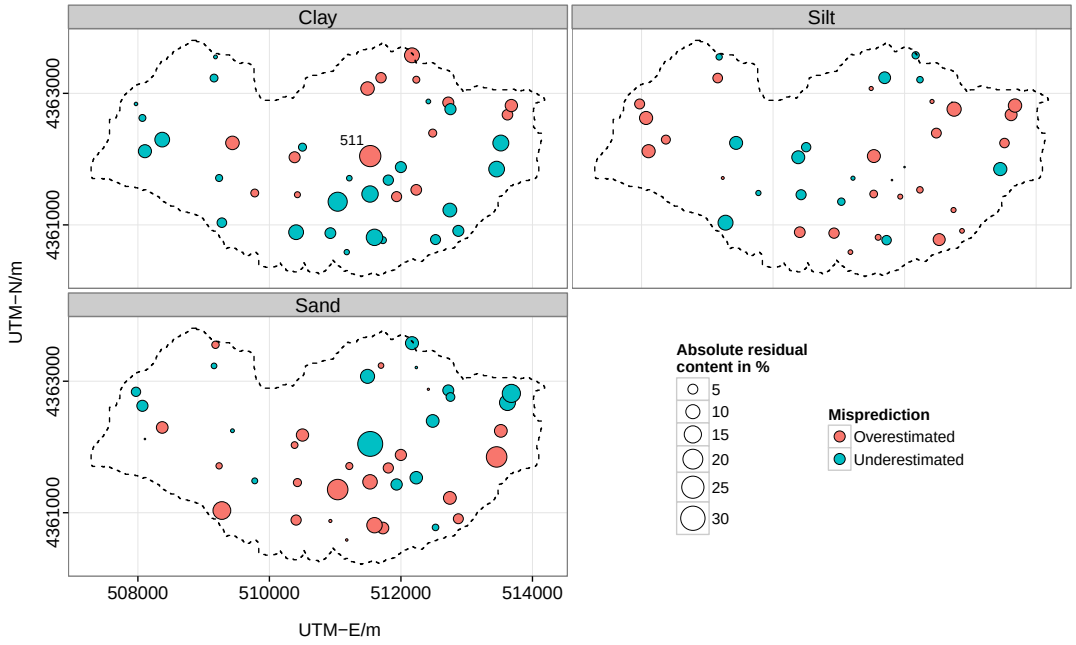


Figure 5.28: Bubble plot of NNRCK test residuals at the landscape scale

Besides analysing hotspots or residual outliers, bubble plots are useful graphical tools to detect systematic (spatial) misprediction. Equally coloured adjacent validation points occur, for instance, close to the southern boundary of the study area and at the vicinity of the Azienda San Michele in the western part of the catchment. In these regions, clay content seems systematically underestimated. A possible explanation might be the nearby Rio di Costara which only indirectly entered the NNRCK model through parent material encodings. In addition, only very little contributions were found for the corresponding input GEOPPR22 during NN modelling (see Garson's plots in figure 5.24 on page 129). Expected high clay contents at locations associated with alluvial deposits such as the one close to the river course are, therefore, likely under-predicted by the given model, because the influence of this particular geological unit on the resulting output is down-weighted through the trained NN model.

Thus, unfortunate spatial error clustering and one hotspot of distinct misprediction slightly weaken the positive overall impression of the ability of the NNRCK model to reliably map soil textural fractions at the Rio di Costara test site.

5.2.5 Model comparison

Neural network residual cokriging (NNRCK) combines the common krige interpolator with a universal approximator from the group of machine learning models. It implies a possible methodological improvement and needs to be assessed against standard spatial prediction techniques. This section compares the performance of NNRCK with inverse distance weighting (IDW), ordinary kriging (OK), regression kriging (RK), ordinary cokriging (COK) and regression cokriging (RCOK). The latter two models were built upon alr-transformed soil textural fractions, whereas IDW, OK and RK were applied to original target quantities. For details on each interpolation procedure refer to the corresponding R-script (.4.17).

Focusing on the presentation of clay estimates, IDW interpolation as shown in figure 5.29 appears fairly local with distinct hotspots at calibration points where high clay contents have been measured. These undesirable bull's eyes around known sample locations are a direct consequence of the exactness property of the IDW technique. Bull's eyes also emerge to a lesser extent from OK, which is, again, classified as an exact interpolator, if no nugget effect contributes to the utilised variogram model. In addition to local hotspots, all kriging outputs show a characteristic smoothing effect. Nevertheless, due to the application of a nested variogram model (refer to the original clay variogram in figure 5.20 on page 124), both local and regional variation is well-modelled by the given OK method.

Moving towards regression-based kriging variants, the RK clay prediction map reveals little difference by visual inspection compared to OK or IDW results. This observation is due to a weak regression performance with an associated coefficient of determination of 0.17 using elevation and the indicator-valued formation of San Vito sandstones as independent variables. Thus, only little variation is addressed by linear regression, so that trend residuals stay close to original clay content values. This eventually reduces the RK method to simple ordinary kriging.

Mapping products from bivariate COK and RCOK models are pretty similar to OK and RK results, respectively. This is in line with remarks given in the context of ternary diagram interpretation. It has been demonstrated that most sample

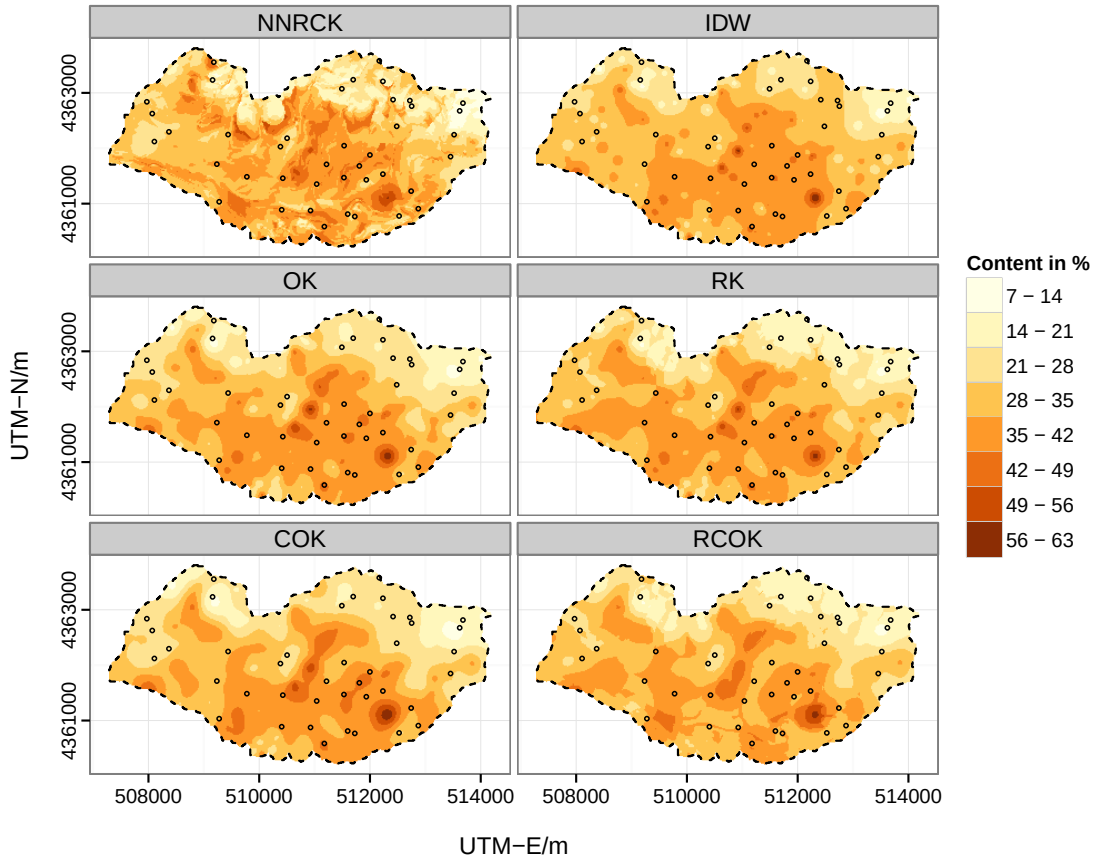


Figure 5.29: Estimates of clay content from different soil spatial interpolation methods in the Rio di Costara catchment. NNRCK, COK and RCOK are applied to alr-transformed soil textural fractions with prediction maps shown after agl-backtransform. The points refer to the 41 test samples.

points of the given test site are clustered in an area of the simplex where distortion of equal Mahalanobis distances, and thus, the influence of the compositional constraints is relatively small. However, the few data points near the vertices (particularly for large sand contents) benefit from using alr coordinates as interpolation input. For instance, COK sand content estimations of 72.2 % (ID = 513) and 66.9 % (ID = 816) are closer to the actual observations of 88.6 % and 80.7 %, respectively, than OK-based predictions accounting for 67.7 % and 63.2 % sand content. In addition, the differences between COK- and OK-estimates for these two samples (4.4 % and 3.6 %) are greater than the average absolute difference of 1.4 % computed from the remaining 39 validation points.

NNRCK-related spatial patterns were already thoroughly analysed in the previous section. The most obvious difference of the NNRCK clay prediction map with respect to all other applied models is its heterogeneous appearance, which clearly resembles the maps of key contributors identified from NN modelling.

Table 5.15: Cohen's κ statistic of agreement among soil texture maps

INTERPOLATION MODEL	NNRCK	IDW	OK	RK	COK	RCOK
NNRCK	1	0.468	0.482	0.565	0.502	0.582
IDW	-	1	0.700	0.671	0.635	0.618
OK	-	-	1	0.681	0.796	0.644
RK	-	-	-	1	0.629	0.824
COK	-	-	-	-	1	0.645
RCOK	-	-	-	-	-	1

Soil texture classes determined in accordance with the USDA Soil taxonomy

In addition to evaluating prediction maps rather subjectively by means of visual inspection, Cohen's κ statistic of agreement, which jointly considers all three compositions, is supplemented. Table 5.15 summarises the results based on soil texture maps computed from the corresponding estimates of particle-size fractions in accordance with the USDA Soil taxonomy. The highest match between soil texture maps occurs for RCOK- and RK-results with a κ value of 0.82, indicating almost perfect agreement following the interpretation proposal of Landis and Koch (1977), cited in Sterlacchini et al. (2011). Substantial agreement exists between outputs from OK and COK expressed by a κ value of 0.80. Similarity between NNRCK predictions and all other model outputs is less obvious, but still classified as moderate agreement. Therefore, quantitative assessment of conformity between categorical soil texture maps certainly highlights former findings from visual comparisons of predicted clay content maps. To summarise, all six considered interpolation techniques yield rather similar prediction results with NNRCK being the model that deviates most from the others, offering the highest degree of heterogeneity.

After comparing model output either by visual inspection or by some numerical measure of agreement, attention is shifted to prediction performance in terms of common error metrics. Table 5.16 provides validation measures computed from clay, silt and sand content of an independent hold-out sample. Both, root mean

Table 5.16: Validation measures of soil spatial interpolation at landscape scale

	Clay		Silt		Sand		
	RMSE	EF	RMSE	EF	RMSE	EF	STRESS
NN residual cokriging	8.329	0.53	5.441	0.45	9.617	0.59	0.737
Inverse distance weighting	8.922	0.47	5.350	0.47	10.049	0.55	0.827
Ordinary kriging	8.521	0.51	5.744	0.39	10.035	0.55	0.835
Regression kriging	8.290	0.54	5.773	0.39	10.062	0.55	0.863
Ordinary cokriging	8.564	0.51	5.645	0.41	9.566	0.59	0.797
Regression cokriging	8.227	0.55	5.943	0.35	9.980	0.56	0.837

RMSE = root mean squared error, EF = model efficiency, STRESS = standardized residual sum of squares

squared errors (RMSE) and model efficiencies (EF, Nash and Sutcliffe (1970)) emphasise an overall similarity regarding prediction performances of the used interpolation techniques. The most promising model in terms of RMSE and EF for clay content is RCOK with values of 8.2 and 0.55, respectively. IDW performs best with regards to silt content reaching a RMSE of 5.4 and an EF of 0.47. Sand content exhibits its lowest RMSE of 9.6 and its highest EF of 0.59 by applying the NNRCK and COK models. For completeness, error histograms of all involved interpolation techniques are shown in figure 5.30.

Table 5.17: Percentage number of best prediction counts

	NNRCK	IDW	OK	RK	COK	RCOK
Clay	29.2	14.6	14.6	7.3	19.5	14.6
Silt	36.6	17.1	17.1	9.8	14.6	4.9
Sand	24.4	17.1	24.4	9.8	14.6	9.8

Counts calculated from independent test data (N = 41)

However, when taking into account that soil textural fractions are compositional, model performance can be accordingly examined by more appropriate Aitchison distances. Doing so, significantly rearranges the rank of best prediction models. A STRESS value of 0.74 clearly favours the NNRCK model, which is also verified by counting the cases where each considered model produces the lowest absolute error. The percentage number of these best prediction counts are summarised in table 5.17. In 29 %, 37 % and 24 % of all independent test samples, NNRCK yields the lowest absolute error and, thus, performs best more often than any reference model for either soil textural fraction.

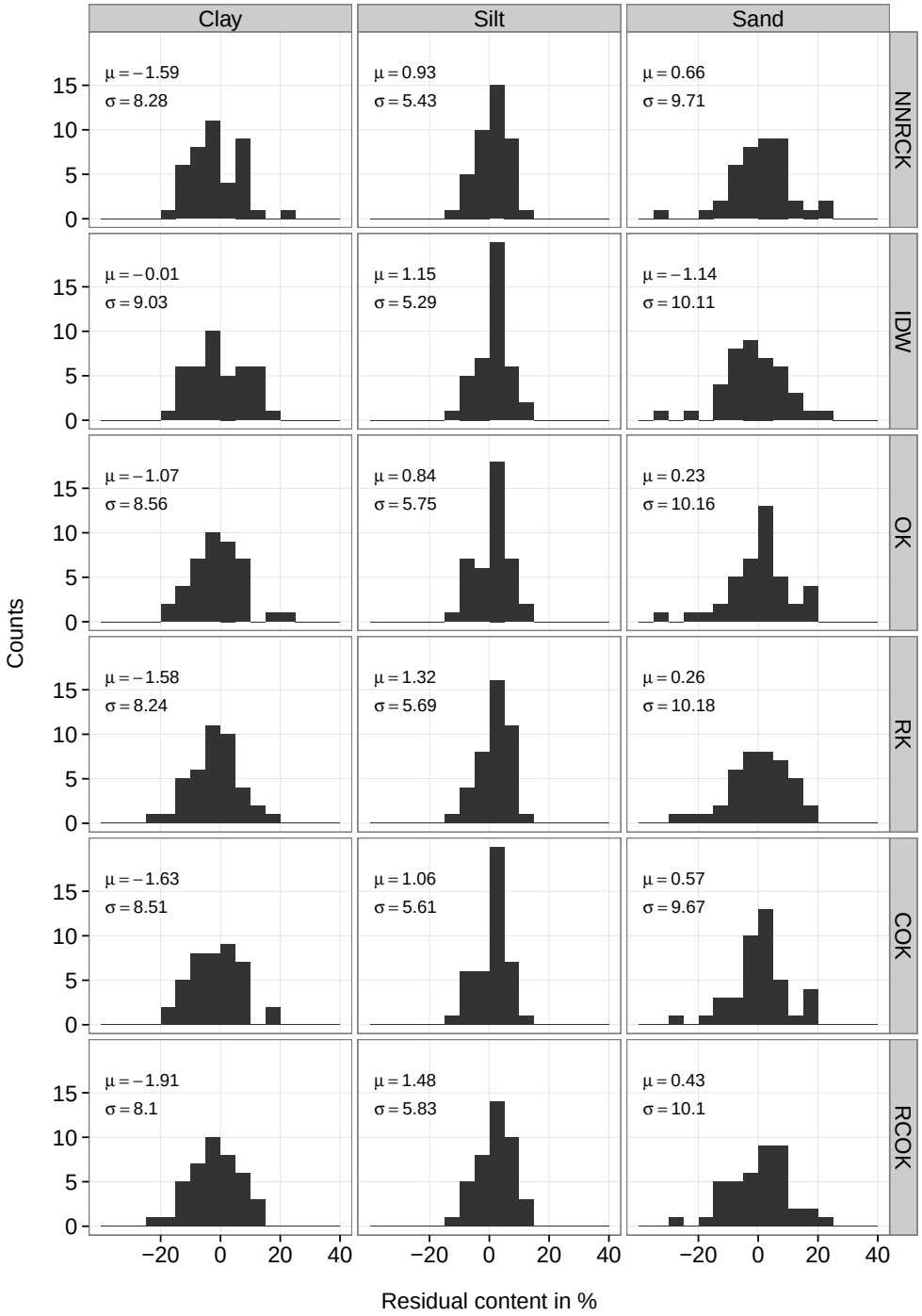


Figure 5.30: Histograms of validation residuals at the Rio di Costara test site, calculated from a hold-out sample of 41 points. The parameter μ and σ are the arithmetic mean values and standard deviations of the distributions.

5.2.6 Interim conclusion

From correlation analysis, three land-surface parameter (elevation, SAGA wetness index and potential incoming solar radiation) were identified as useful explanatory factors in the given interpolation problem. In addition, exploratory data analysis revealed an apparent influence of geological categories on variations of soil texture in the Rio di Costara catchment. In combination with 197 point measurements, these two groups of environmental covariates set the stage for training of a neural network to address large-scale variability among the dependent variables. Eventually, a multi-layer perceptron with nine hidden units was modelled using resilient back-propagation in 40 training cycles. After de-trending, remaining small-scale variation among the stationary neural network residuals were tackled by cokriging based on a fitted linear model of coregionalisation. Validation of final results from independent test data suggests that the neural network residual cokriging model predicts well on average. In a last step, the promoted NNRCK model is found to be superior to common interpolation techniques, particularly in the context of multivariate distance measures, best prediction counts and plausibility considerations from visual inspection.

Chapter 6

Discussion

The previous chapter provided a wide selection of interpreted figures and tables to demonstrate the interesting findings related to this thesis. The following chapter carefully discusses the various results by placing them in scientific context. In addition, it revisits unexpected observations and highlights limitations of the conducted studies. The discussion is, thereby, structured according to the specific objectives formulated in chapter 1.2 on page 3. It begins with the provision of geographically continuous information on soil textural fractions at two different spatial scales, followed by an evaluation of neural network performances in hybrid digital soil mapping frameworks. Subsequently, the applicability of geophysical measurements as covariates in small-scale digital soil mapping is discussed, exemplified by the two investigated Mediterranean fields. The final section highlights the merits of web-based distribution of soil mapping products.

6.1 Digital mapping of soil textural fractions

Referring to the first specific objective of this dissertation accurate maps of clay, silt and sand content were provided at different levels of spatial extent at an assigned Mediterranean basin. The creation of these geographically continuous soil property maps from limited sets of point data was conducted following the con-

cepts of hybrid spatial interpolation techniques. Consequently, the resulting maps strongly reflect the empirical correlations found between the target quantities and available scorpan factors. In addition to this external drift component, they show patterns modelled from residual auto- and cross-variograms representing random spatial variation.

The use of ancillary (co)variables

For the entire Rio di Costara catchment, continuous environmental covariates were derived from a digital elevation model. In accordance with the results of a correlation analysis, three land-surface parameter (elevation, SAGA wetness index and potential incoming solar radiation) were selected as input for soil textural trend estimation. Among these considered covariates, elevation showed the highest absolute association with sand content ($\rho = 0.45$). Additionally, elevation was recognised as the most influencing terrain attribute with regard to alr-transformed clay and silt contents in the final neural network architecture, showing relative importance values of 15 % and 25 %, respectively. This significant impact of altitude on soil spatial variability in undulating landscapes is well-known from early studies quantifying soil-environment relationships (Moore et al. 1993; McKenzie et al. 2000). More recently, Ließ et al. (2012) identified altitude as most relevant predictor for soil texture mapping in a machine learning application (random forests) at the landscape scale.

In addition to DEM derivatives, exploratory data analysis also revealed an apparent impact of geological categories on soil textural variability at the given test site. Focusing on log-ratio transformed target variables, eluvial-colluvial deposits played the most important role in the given neural network model with relative importance values of 30 % and 40 %, respectively. This generally corresponds to the results by Gobin et al. (2001) and Ließ et al. (2012) who successfully used elevation-based covariates, but also acknowledged the remarkable influence of parent material on the spatial variability of soil texture. Greve et al. (2012) even found that parent material was more influential than terrain attributes using regression-trees in a national-scale soil (texture) survey.

In contrast to soil texture interpolation at the landscape scale, dummy variables of geological categories were omitted from the set of possible predictors regarding the two agricultural fields, because their original reference scale of 1:25,000 was too coarse to support fine-scale mapping. Instead, geophysical measurements were used to represent parent material in the scorpan-based model framework at the San Michele farm. Ignoring minor differences between field 21 and 33, on the one hand, as well as clay, silt and sand content, on the other, the highest Pearson's correlation coefficients were generally related to gamma-ray nuclides. A maximum value of 0.58 was measured between topsoil sand content and uranium concentration. This absolute number lies in a similar range of what has been reported by Buchanan et al. (2012). However, the signs of coefficients and, thus, the direction of inter-relations observed in this thesis are unusual with respect to observations from other case studies (e.g. Megumi and Mamuro 1977; Taylor et al. 2002). A detailed discussion of these differences will follow in section 6.3.

Aside from gamma-ray nuclides, other covariates were considered at the field scale including x- and y-coordinates, apparent electrical conductivity (ECa) measures from electromagnetic induction and a set of land-surface parameters (elevation, slope and SAGA wetness index). Here, focusing on the latter attributes, up to moderate correlations with soil textural fractions were found (see figures 5.4 and 5.5 at pages 93 and 94, respectively). Yet more importantly, rather diverse relations to clay, silt and sand content occurred in terms of direction comparing the two studied fields. For example, this is evidenced by an unusual correlation between sand and elevation ($\rho = -0.38$) at field 21 as opposed to an expected positive relationship between the same variables at field 33 ($\rho = 0.33$). The latter is slightly lower, but otherwise in line with observations from landscape scale. The opposed correlation signs at field 21 are likely due to a former river course, which has led to higher sand contents in the lower south-eastern part of field 21. This local cause generally confirms Sumfleth and Duttmann (2008) and Lennartz et al. (2009) who found that the explanatory power of relief-based covariates for soil mapping decreases with finer spatial scales due to growing site-specific influences. In addition, lower measurements and partly unusual soil-landscape relationships at the field scale may arise from the fact that the used DEM with an original

resolution of 10 m may possibly be too coarse to properly reflect terrain details at that particular level of spatial extent. Furthermore, the conventional 3 x 3 moving window used to calculate many of the land-surface parameters might be undersized and better adjusted to match the true topographic features of the San Michele farm. Cavazzi et al. (2013) and Maynard and Johnson (2014) recently emphasised the importance of scale and neighbourhood size regarding the calculation of DEM derivatives used as environmental covariates in digital soil mapping. Apart from possible issues related to grid resolution or analysis scale, elevation errors and positional uncertainties in the DEM data source affect the quality of derived terrain attributes. Vacca et al. (2014) reported a vertical and horizontal accuracy of 2.5 m for the given DEM, which is negligible at the landscape scale, but likely influential regarding field level terrain analysis.

Shortcomings concerning observed relationships between single covariates and target quantities were particularly apparent in respect to field 33. As a consequence, this led to relatively poor regression performances with explained variance levels of 41 % for alr Clay and 28 % for alr Silt. On the contrary, regression modelling at field 21 was much more successful in terms of R-squared values of 0.67 (alr Clay) and 0.74 (alr Silt). This difference in performance is largely due to additionally considering x- and y-coordinates which were useful to address global spatial structure at field 21, but had no apparent influence with regard to field 33. Triantafilis and Lesch (2005) also combined EMI-based ECa measures with point coordinates to map topsoil clay content, but reached a slightly higher explained variance level of 79 %, because of more distinct relationships between clay and EMI-signals. On the other hand, van der Klooster et al. (2011) relied on gamma-ray nuclides (K and Th) for topsoil clay mapping at multiple fields and found coefficients of determination ranging from 0.78 to 0.95. However, both mentioned studies only investigated clay content and did not simultaneously model other members of the composition. Taking this into consideration, and with regard to back-transformed predictions, clay mapping at field 21 has led to a lower error (4.4 %) compared to the 4.9 % of Triantafilis and Lesch (2005) and was almost as accurate as the results by van der Klooster et al. (2011) with reported RMSE values of 3 % to 4 % depending on the different fields.

Spatial auto- and cross-correlations

Ancillary (co)variables were meant to support any global trend estimation among the alr-transformed target quantities using either neural networks or linear regression. Subsequently, model residuals were analysed for remaining spatial auto- and cross-correlations. At the landscape scale, a linear model of coregionalisation (LMC) was fitted, showing significant local spatial dependence up to 435 m. Very similar spherical curve shapes, but with slightly shorter ranges (approx. 300 m) corresponding to a smaller study area, have been found by Lark and Bishop (2007). At field 33, Nugget-to-Sill-Ratios of about 60 % clearly indicate the presence of only moderate spatial dependencies up until 240 m. By contrast, the LMC fit was less uniformly structured (different partial sills) at field 21 exhibiting linear increases in dissimilarity only up to 56 m. An identical range for clay content was reported, for instance, by De Benedetto et al. (2012).

While residual (cross-)variograms were the ones that entered the final prediction models, versions related to raw and alr-transformed response variables helped to detect their level of spatial complexity prior to mapping. With regard to landscape scale and at field 21, most raw and alr-transformed variograms revealed two different structures defining the spatial variation of soil textural fractions. Variograms related to field 21 were characterised by ranges of approximately 135 m until a second unbounded structure appeared with larger distances. The two ranges of spatial dependence for the entire Rio di Costara catchment were 465 m and 4620 m, respectively. This two-fold spatial structure is common and has been reported, for instance, by Piccini et al. (2014). At the landscape scale, the long-range part of soil texture variation is largely controlled by regional influences of geology and topography. Short-range variability is more difficult to explain, but might be attributed to more localised influences from (past) erosional processes, land use and soil management practices or depositional peculiarities due to former river channels. Such an ancient creek bed crosses the San Michele farm in a south-easterly direction strongly affecting parts of field 21 and is, therefore, likely responsible for the global structure observed from clay and sand content measurements at that particular field (see Cassiani et al. 2012).

To statistically summarise the results from raw, alr-transformed and residual variogram analysis, it can be stated that at field 21 and particularly at the landscape scale, obvious trends were successfully removed by the GLS regression and neural network models, respectively. This was not true for field 33, where long-range variability clearly persisted among the residuals.

Validation of the prediction methods

Both, soil-landscape relationships as well as spatial auto- and cross-correlations strongly define the plausibility of resulting soil maps. This also implies that related findings directly affect the overall accuracy of the hybrid prediction models.

With respect to the entire Rio di Costara catchment, the multivariate prediction of clay, silt and sand content reached an explained variance level of 57 %, 49 % and 63 %, respectively. The corresponding root mean squared errors (RMSE) from independent testing were 8.3 %, 5.4 % and 9.6 %. These errors appear relatively high compared to other large-scale mapping applications regarding soil textural fractions. For example, Piccini et al. (2014) reported RMSE values of 4.7 % and 8.2 % for mapped clay and sand content, respectively, using regression kriging in an Italian region (100 km^2) which is also dominated by clay loam texture. Neural networks trained by Zhao et al. (2009) also using the resilient back-propagation algorithm have led to very similar errors of 4.8 % (clay) and 7.6 % (sand). However, being a dimensioned measure, RMSEs are generally difficult to compare between regions or variables. This is particularly true, if no standard deviations of the target quantities are provided, as in the case of Zhao et al. (2009). Using the standard deviations reported by Piccini et al. (2014) for normalisation of their RMSE values and subsequent re-calculation as proposed by Hengl et al. (2004) leads to a rough guess on explained variability in the order of 50 %. This is less accurate compared to what has been obtained for the Rio di Costara catchment and may be explained by the more simplified regression kriging method applied by Piccini et al. (2014). Other machine learning models based on random forests were used by Ließ et al. (2012) or Akpa et al. (2014) resulting in slightly lower proportions of explained topsoil texture variability compared to this work.

At field 21, multivariate mapping of clay, silt and sand content led to explained variance levels of 41 %, 29 % and 78 %, respectively. The deviation-based error metrics were in the range of 3.3 % (silt) to 4.4 % (clay). Thus, RMSE values were remarkably lower compared to landscape scale which is mainly due to the smaller study area, but also likely influenced by the different validation basis. Unlike at landscape level where an extra test dataset has been used to validate the model, leave-one-out cross-validation was performed at the field scale due to significantly less available soil samples. Cross-validation techniques in general, and its leave-one-out version in particular, are known to underestimate the true error, as outlined, for instance, by Bengio and Grandvalet (2004).

Focusing on relative validation measures, several other studies which compare well with the results obtained at field 21 have been published. For instance, Priori et al. (2014) have determined equally high R-squared values of 0.65 and 0.74 regarding sand content using gamma radiometric data in either neural network or support vector machine approaches. Focusing on the clay fraction, Rodrigues et al. (2015) found a moderate coefficient of determination of 0.46 relying on multiple soil sensors in a principal component stepwise regression framework. The latter procedure, but without an explicit mapping objective, was also used by Mahmood et al. (2012) resulting in rather diverse R-squared values of 0.34 and 0.87 for clay and sand contents, respectively. This is not only in line with the magnitude of values obtained here, but also resembles the performance differences by fraction. Note, however, that both, Mahmood et al. (2012) and Rodrigues et al. (2015) studied contrasting soils. The results of this work are only compared to the outcome of those sites (or samples) that provide similar texture conditions as judged from reported descriptive statistics and soil type delineations.

As opposed to field 21, models at field 33 with corresponding R-squared values of 0.16 (clay), 0.11 (silt) and 0.42 (sand) must be rejected with regards to fine-soil fractions. Moreover, these very low prediction performances cannot compete with results from other field-level case studies. One major reason for comparatively inaccurate interpolation results at field 33 lies in the weak empirical correlations between covariates and target quantities. In addition, sample size ($N = 64$) and/or point distribution were obviously not sufficient to properly address spatial topsoil

texture variability through auto- and cross-correlation analysis. On the one hand, these unfortunate observations are caused by poor quality predictors. Discussing the (co)variables already revealed some shortcomings of the derived land-surface parameters (coarse DEM) and geophysical attributes (signal predominated by other soil properties). Thus, new information, particularly regarding the parent material, is required to significantly improve the scorpan-based part of the given prediction model. On the other hand, finding an appropriate model for field 33 is hampered by great spatial heterogeneities among the target variables. Large soil variability at the San Michele farm is generally known from previous surveys (Aru 1966) and particularly apparent at field 33 as further demonstrated by observations from own soil profiles (see pages 23 to 27). Moreover, additional evidence is provided by distinct soil colour differences visible from satellite images (see Cassiani et al. 2012).

Taking all the (estimation) results of this thesis into consideration, different processes of soil formation are responsible for the current spatial distribution of clay, silt and sand content at the field and catchment scale. It follows that individual soil spatial prediction models are generally required to properly address the scale-dependent peculiarities of soil variation. Moreover, unintended and unexpected differences observed between adjacent agricultural fields call for individual spatial interpolation strategies even at the same scale, if great soil heterogeneities characterise the study area. This includes a site-specific selection of environmental covariates and, thus, strongly emphasises the importance of an extensive exploratory data analysis prior to (geo)statistical data processing.

In each soil texture mapping application of this work – i.e. independent of spatial scale – silt content was the most difficult fraction to predict. Although equally oriented performance ratios have been reported elsewhere (Krüger 2007; Akpa et al. 2014; Wetterlind et al. 2015), their explanation remains difficult and site-specific. One possible reason might be that significant accumulations of silty material due to aeolian processes occur in the given test sites. Accordingly, Yaalon (1997) mentioned Sahara dust as an important source of fine soil material in Mediterranean soils. Such an apparent clustering of high silt content values was observed in the central part of field 21. However, particularly with regard to the

entire Rio di Costara catchment, such wind-induced tendencies towards higher silt percentages are not quantified and, consequently, not part of the prediction model. This is a distinct shortcoming, since accumulation of wind-borne material occurs rather spatially diffuse and can be quite significant. For example, Sevink and Kummer (1984) found silt-size deposits of up to 200 t/ha in the top 10 cm layer of depression soils on the near-by Giara plateau in central Sardinia. Another factor that might help to explain the difficulties observed from silt mapping is concerned with the laboratory methods. Incorrect measurements are more likely to occur between clay and silt fractions during sedimentary analysis than by sieving to determine sand percentages. Therefore, the interpolation basis for silt content mapping is possibly more erroneous than for any other of the modelled textural classes.

Meeting the compositional requirements of soil textural fractions

Among others, Odeh et al. (2003) repeatedly stated that untransformed components of compositional soil data should never be mapped one by one. In accordance with this recommendation and following the suggestions by Pawlowsky et al. (1995) and Lark and Bishop (2007), trend residuals of alr coordinates were cokriged, as illustrated in the methodological framework of this thesis.

Independent from spatial scale, ternary diagram interpretations have shown that most sample points of the considered test sites are barely affected by the additional model requirements of compositional data. Similar results were found by Lark and Bishop (2007) in regard to their East Creek case study. Additionally, Cohen's κ statistic revealed substantial agreement among soil texture maps from ordinary kriging (raw values) and cokriging (alr coordinates), on the one hand, and regression kriging (raw) and regression cokriging (alr), on the other. From this, it follows that by ignoring the compositional restrictions and using raw values for prediction is, on average, sufficiently accurate from a practical point of view. However, at the landscape scale a few test data estimations (for those points with large measured sand contents) were considerably more realistic using alr coordinates as model input.

The strong observed similarity between univariate and bivariate kriging models also raises the question, whether it is reasonable to use the more complex cokriging technique rather than estimating the alr-transformed target variables in isolation. Webster and Oliver (2007) discussed this issue in detail concluding that differences between kriging and cokriging results likely diminish, if equally sampled variables are hardly correlated. This is not true at the landscape scale, where the two fitted neural network outputs were significantly correlated ($\rho = 0.62$). However, considering cross-correlations at the field scale is less likely of any value as judged from non-significant correlation coefficients of 0.18 (field 33) and 0.28 (field 21).

Taking everything into consideration, using alr coordinates is advised when dealing with compositions to meet additional model requirements caused by the non-negativity and constant sum constraints (e.g. 100 %) of particle-size fractions. Moreover, cokriging should be used for geostatistical modelling if interdependencies between the (transformed) response variables are documented. In doing so, spatial interpolation is statistically sound and, thus, generally more likely to produce accurate soil property maps. However, less experienced practitioners might as well disregard this additional complex procedure. In any case, following Lark and Bishop (2007) a thorough investigation of the ternary diagram is mandatory during exploratory data analysis to check for data points that are likely affected by compositional constraints.

To further improve the accuracy of the proposed methodology particularly in scientific applications, the biased agl-(back)transformation of final estimates should be replaced by a more sophisticated procedure. Pawlowsky-Glahn and Egozcue (2006) proposed an unbiased back-transform using Gauss-Hermite quadrature applied, for instance, by Ward and Mueller (2012). Moreover, Menafoglio et al. (2014) recently introduced a new functional compositional kriging approach to spatially predict from particle-size curves rather than classes. This seems especially useful for better support of soil-related process modelling, but assumes more advanced lab-analyse techniques to provide an adequate interpolation set of high-dimensional particle-size data.

The practical usefulness of the results

First and foremost, geographically continuous maps of soil textural fractions at the catchment scale can serve as valuable base data for more complex soil parameters as required, e.g. in hydrological modelling. For example, the estimation of groundwater recharge strongly depends on soil characteristics such as moisture content at field capacity which can be deduced from texture using pedo-transfer functions (see Ehlers et al. 2015; Herrmann et al. 2015, for recently published applications in the Mediterranean region). In addition to the provision of input data for hydrological modelling purposes, the produced soil property maps might be considered as decision making support for sustainable agricultural irrigation management, which is an essential requirement for (future) farming activities in the Mediterranean area. Moreover, detailed maps of clay, silt and sand content can contribute to better-adapted prevention and mitigation of soil degradation, particularly useful regarding the island of Sardinia. The accuracy of the maps created in the frame of this dissertation are generally sufficient for these purposes, considering the great soil heterogeneities and relatively large extent of the catchment area.

As part of an agricultural research farm, soil maps presented at field 21 are likely to be used for studying the textural influences on soil fertility, water-holding capacity and other soil functions related to the optimisation of site-specific crop management. However, results obtained for fine-soil fractions at field 33 should be used with caution, if at all, due to poor prediction performances.

6.2 Performance of neural networks in soil spatial interpolation

In the frame of soil texture mapping at the landscape scale, another more specific objective of this dissertation was related to the usefulness of artificial neural networks as de-trending method in hybrid spatial interpolation. In response, the performance of neural network residual cokriging (NNRCK) was compared

to inverse distance weighting (IDW), ordinary kriging (OK), regression kriging (RK), ordinary cokriging (COK) and regression cokriging (RCOK). This part of the discussion focuses on NNRCK and its distinct evaluation in view of the other methods. For a detailed description and interpretation of differences among all involved models return to section 5.2.5 on page 137.

Prediction performances of the NNRCK method and all benchmark models were listed in table 5.16 on page 140. Comparing univariate validation measures (root mean squared errors and model efficiencies) showed that the promoted NNRCK approach is always among the top-three methods. It is also the model that predicts the three target quantities (clay, silt and sand content) in the most balanced way. However, the differences between the considered methods are relatively small in total. For instance, model efficiencies of sand content only differ by a magnitude of 0.04, ranging from 0.55 to 0.59. This is consistent with model comparison results described by Padarian et al. (2012) and Guo et al. (2013) and can partly be explained by comparatively low correlations between the input and output variables. Moreover, 125 calibration samples might not be enough to fully exhaust the potential of a data-driven, machine-based learning technique such as neural networks. For example, Minasny et al. (2008) related the usefulness of data-mining tools to training sets of at least 200 cases.

Nevertheless, particularly by including multivariate distance measures and best prediction counts, NNRCK was found markedly superior to common interpolation techniques. For example, the NNRCK model reached a standardised residual sum-of-squares value of 0.74 as calculated from the 41 independent test samples. This is clearly more favourable than the same measure of 0.8 found for the second best performing model (ordinary cokriging). Moreover, percentage numbers of test cases where NNRCK produced the lowest absolute error were constantly higher than for any reference model and independent from soil textural fraction (see table 5.17 on page 140). In addition, this overall impression of a preferable NNRCK model is supplemented with plausibility considerations from visual inspection. Contour maps from NNRCK offer a desirable, higher degree of soil spatial heterogeneity than standard kriging does. Similarly obvious differences occur, for example, in the prediction maps (soil organic matter content) provided

by Dai et al. (2014) comparing ANN-kriging results with estimates from universal kriging and IDW. Padarian et al. (2012) also reached more detailed spatial distributions of soil organic carbon content modelled by a coupled neural network and kriging technique as compared to other (co)kriging variants.

Focusing on the chosen network architecture, its complexity can be considered as adequate for capturing long-range variability among the target quantities. The best illustration of this is given by comparison of theoretical variograms fitted to alr-transformed variables (see figure 5.21 at page 125) with the linear model of coregionalisation derived from neural network residuals (see fig. 5.25 at p. 131). The formerly unbounded spatial structure reduced to a single component of local spatial dependence indicating that the trend has been successfully modelled by the three-layer perceptron. This de-trending ability of small-sized networks corresponds to findings from Kanevski et al. (1997) or Cellura et al. (2008).

While assessing the generalisation ability of the final models, it should be noted that neural network training relies on less points than the competitive methods. This possibly weakens the overall NNRCK performance to some extent, but is unavoidable, since finding the optimal net architecture required further splitting of the calibration sample. This dataset partitioning based on a modified Kennard-Stone algorithm and is in line with Galvão et al. (2005) and Saptoro et al. (2012) who emphasised the great importance of a fortunate data division in modelling a neural network. In combination with the efforts made during soil sampling to receive calibration and test datasets that proved to be representative for each other, the resulting three datasets can be considered as a rather comfortable set-up. However though, a few individual points turned out to be very badly predicted during validation (see section 5.2.4 on page 132). Thus, to further increase the NNRCK performance in the Rio di Costara catchment, the surrounding area of these conspicuous samples should be considered for supplementary survey.

To summarise, comparison of the proposed NNRCK method with traditional digital soil mapping techniques revealed a better ability of NNRCK to provide accurate soil spatial predictions. At the same time, it has been demonstrated that machine learning based on the multi-layer perceptron can efficiently model

long-range soil variation and, thus, is a promising tool to address the trend component in hybrid interpolation methods. In addition, the counter-argument of many other studies (e.g. Ließ et al. 2012; Greve et al. 2012; Brungard et al. 2015) that neural networks are black box models could be invalidated. Instead, several methods were presented to evaluate the covariate contributions to the final network. Altogether and with regard to the Rio di Costara catchment, data-driven learning improved soil texture mapping at the landscape scale, but was not an option at both, fields 21 and 33 due to significantly smaller sample sizes.

6.3 Integration of geophysical sensor data

Field-scale mapping of soil textural fractions was closely linked to the question whether minimal-invasive geophysical measurements can provide useful predictors for regression-based interpolation techniques. Results from (linear) model testing indicated that gamma-ray nuclides and ECa measured by EMI devices significantly contributed to the successful estimation of clay, silt and sand content at field 21. By contrast, only modest prediction performances were achieved with regard to field 33. Moreover, in view of a more process-based understanding issues remain to be resolved as will become apparent from a detailed discussion on bivariate empirical correlations.

As mentioned earlier in this chapter, the highest absolute inter-relations between soil textural fractions and geophysical sensor data were found for gamma-ray nuclides. However, the signs of Pearson's correlation coefficients, as experienced from the considered field sites of the San Michele farm, correspond to observations by Priori et al. (2014) found in soils developed from similarly heterogeneous parent material, but are contrary to what is known from most other case studies (e.g. Megumi and Mamuro 1977; Taylor et al. 2002; Buchanan et al. 2012). This finding indicates that the received gamma-signal is likely superimposed by some other soil property than texture. In fact, there is a very high negative correlation up to 0.9 between gamma-ray nuclides and calcium carbonate content at both studied fields. Thus, calcium carbonate strongly attenuates measured gamma-ray counts

which is more remarkable in areas with higher percentages of fine-soil fractions. In addition to obvious carbonate influences, observed gamma-ray measurements proved to be sensitive to gravel content at field 33, where dose rate, for instance, showed significant positive correlation of 0.37 with the amount of coarser soil particles. In contrast to this, gravel content had no apparent influence on the relationships between textural fractions and radio-nuclide concentrations at field 21, because its measured percentages were on average considerably lower and also spatially much less variable. Focusing on ratios of pairwise nuclides instead of pure gamma-ray counts led to lower, but commonly directed associations with the amount of clay, silt and sand particles (see Petersen et al. 2012).

As a second source of geophysical predictors, apparent electrical conductivity (ECa) measures from electromagnetic induction were considered, but found only significantly related to silt content ($\rho = 0.38$) at field 21. This obviously lower influence on topsoil properties compared to gamma-ray spectrometry is mainly due to the different depth-response curves of the two sensing methods. Whereas about 90 % of the gamma-ray signal originates from the top 30 cm of the surveyed soil volume (Cook et al. 1996), EMI devices cover the first 75 cm to 150 cm in depth depending on operation mode. In mapping practice, Taylor et al. (2010) and Piikki et al. (2013) both concluded from their multi-sensor studies that gamma radiometric data is more likely to be promising regarding topsoil properties, while EMI measurements better suit subsoil interpolation. Unlike with pure gamma-ray signals, relationships between ECa measures and soil textural fractions were not greatly disturbed by other soil attributes. For instance, well-known influences by soil moisture have been found negligible at the time of the measurements as indicated by observed volumetric water contents below 10 %. Moreover, ratios of (lab-analysed) ECa to clay content of 0.5, on average, clearly point to non-saline conditions in the given test sites (see McBratney et al. 2005).

Altogether, significant correlations were found for both, electrical conductivity and gamma radiometric data with at least one measured soil textural fraction. Thus, either sensing method provided potentially relevant secondary variables for soil mapping at the studied agricultural fields of the San Michele farm. However, severe empirical correlations between the sensor values themselves required

additional processing to avoid any negative effects of multicollinearity in the final regression model. Consequently, the multi-source (sensor) data were fused by principal component analysis to provide effective and efficient predictor sets. This procedure is in accordance with De Benedetto et al. (2012) and Castrignanò et al. (2012) who also concluded that principal components derived from multi-source data are more suitable for soil property interpolation than single sensor values. In this work, model validation at field 21 revealed comparatively high performances as evidenced by R-squared values of 0.41, 0.29 and 0.78 for clay, silt and sand content, respectively. This was partly due to integrated x- and y-coordinates, but analysing the loadings of the PCs selected from stepwise regression procedure showed that gamma-ray counts, nuclide ratios and ECa measures significantly contributed to the created prediction maps. In addition, comparing estimations of clay, silt and sand content with corresponding ordinary kriging results, leads to relative model improvements of 4 % (silt) to 11 % (sand) as calculated from root mean squared errors. However, poor model performance was achieved in regard to field 33 for reasons discussed in section 6.1.

To conclude, multi-source geophysical sensor data were found potentially useful to support and improve field-scale soil texture mapping. However, their integration as covariates has also revealed some severe limitations, particularly with regard to process-based interpretation and when applied to field sites characterised by great soil heterogeneity. Thus, further research is required to better resolve the signals from gamma-ray spectrometry and EMI devices, particularly in the presence of highly calcareous soils. Moreover, an advanced functional understanding of the relationships between certain soil properties and sensor-based (co)variables is essential to accept the additional costs that these on-the-go techniques entail, if applied, for instance, in precision agriculture. With regards to the San Michele farm, this requires additional knowledge to quantify the spatial distribution of soil parent material, and more importantly, to properly reflect mineralogy.

As a final remark on geophysical covariates, it should be noted that observations from proximal soil sensors themselves are point measurements. To serve as predictors in regression-based soil spatial interpolation models, these datasets need to be mapped as well. Due to the very high number of points – sample sizes rang-

ing from 1478 (gamma-ray spectrometry) to 40710 (electromagnetic induction) – available in the frame of this thesis, ordinary kriging was sufficiently adequate to create geographically continuous sensor data maps. However, in regard to future applications, other processing techniques might be used, for instance, based on multiple-point simulations as suggested by Meerschman et al. (2014) who used EMI sensing to detect cryogenic features in the subsoil.

6.4 Online dissemination of final soil maps

The final step of the methodological framework and the last specific objective of this dissertation was related to the delivery of achieved mapping results. This task is of great importance, particularly in times of rapidly growing amounts of geographically continuous soil information and ever increasing numbers of applications that rely on these datasets. Among others, Wilson et al. (2012) stressed the need for online, on-demand information systems to provide long-term visibility and access to mapped soil data.

The online platform, described in section 4.8 on page 83, was originally deployed in association with the international research project CLIMB and primarily focuses on hydrological modelling output. As another thematic priority, it successfully disseminates and documents the final soil maps of this thesis in terms of geospatial web services (CSW, WMS, WCS). In distributing the spatial resources using interoperable access formats and describing them in a standards-based manner, a sustainable storage of digital soil mapping products could be established. The implementation of this system totally relies on free and open-source software (FOSS) and made use of the geospatial content management system GeoNode. Focusing on FOSS in general, and the GeoNode suite in particular, was highly beneficial in the sense that software could be easily adapted to meet project-specific requirements, e.g. regarding an additional time series application (related to hydrological indicators). Moreover, maintaining the geoportal solution is much easier due to lower costs, since no license fees need to be paid. At the time of writing this dissertation, other geospatial web-platforms using GeoNode

include the WFPGeoNode – a data sharing application of the United Nations World Food Programme – as well as the information system (PaRIS) of the Pacific Catastrophe Risk Assessment and Financing Initiative. Regarding European soil data, a recent example of a standard-compliant soil information portal has been provided by the GS Soil project (Feiden et al. 2010), based on the software bundle InGrid (Kruse and Konstantinidis 2010).

Once operational, GeoNode turned out to be a good choice for deploying the desired web-platform. However, full INSPIRE-compliance could not be realised, for instance, regarding the use of harmonised layer names or additional supported languages in the context of view services. This is partly due to limitations of the current GeoServer instance and its specific behaviour of adding the workspace as a prefix to each layer name. Yet more importantly than slight violations of international standards, response times grew rapidly serving a huge amount of spatial layers. This issue might be solved by increasing the number of map servers administrating the GIS-related web services and feeding GeoNode's Django-based web interface. Lately, such an approach has been presented by Arias et al. (2015), who as well used GeoNode components to create their distributed water information systems. Following this option also offers the possibility to separate soil-related layers from hydrological resources and, thus, avoiding to provide them in one single service.

To conclude, the platform presented in the frame of this thesis stands as an example that digital soil mapping data can be effectively delivered to a broader community using free and open-source software. It allows access and use of valuable input variables for own specific environmental workflows. Moreover, its metadata documentation is not only of use to experts in soil science, but also to stakeholders, policy makers and other users interested in the available resources long after the project has ended. In addition to deployment and extension of an own portal solution, an increased exchange with other existing soil information infrastructures is intended. Depending on the kind of data eventually considered for disclosure, this includes large-scale international initiatives such as GlobalSoilMap (IUSS), or World Soil Profiles (ISRIC), but also applies to regional projects like the newly established Database Soil Sardinia (AGRIS Sardegna).

Chapter 7

Conclusions and future work

Despite the strong anthropogenic influences and great soil heterogeneities that characterise the Mediterranean region, appropriate procedures have been created to map soil textural fractions of the topsoil at the field and catchment scale in southern Sardinia (Italy).

With regards to the entire Rio di Costara catchment, 197 soil observations and two groups of environmental covariates (land-surface parameters and geological categories) were adequate to train a multi-layer perceptron and, thus, to address large-scale variability among the target quantities. Remaining short-range variation was successfully estimated by fitting a linear model of coregionalisation and subsequent cokriging of the residuals. Testing showed that the proposed neural network residual cokriging approach outperforms common digital soil mapping techniques. This suggests that machine learning algorithms such as artificial neural networks are an effective and efficient tool for multivariate, non-linear trend analysis in soil spatial prediction.

Focusing on two agricultural fields at the San Michele farm, spatial interpolation of clay, silt and sand content was performed with regression-based hybrid methods using explanatory variables derived from geophysical measurements and digital terrain analysis. In response to severe multicollinearity and to avoid possible issues due to spurious correlations, principal components of multi-source data were

used as predictors in the GLS regression part instead of raw covariates. Results from leave-one-out cross-validation at field 21 indicated that the regression cokriging model predicts soil textural fractions fairly well. Contrasting this, rather weak model performances were obtained with regards to clay and silt content at field 33. To conclude, the use of electro-magnetical and gamma radiometric data as ancillary information for digital mapping of soil textural fractions at the field scale were found difficult when dealing with heterogeneous and partly calcareous parent material.

To further increase the quality of soil spatial interpolation models, other data sources might be considered to cover additional factors of soil formation as well as to provide better predictor sets in terms of accuracy and resolution. Such extra information can be obtained, for instance, from remote sensing data recorded by spaceborne sensors or instruments mounted on airborne platforms like unmanned aerial vehicles (UAV). Among others, Sumfleth and Duttmann (2008) used vegetation indices from various satellite scenes to predict clay, silt and sand content at the landscape scale. In contrast to optical imagery, Singh and Kathpalia (2007) introduced an approach for the direct retrieval of soil texture along with soil moisture and surface roughness from radar data. Compared to satellite imagery, UAV systems are more operationally flexible, which makes them particularly relevant with regard to field-level surveys. Depending on the sensors used, common UAV data acquisition comprises multi- or hyper-spectral and thermal imaging, LiDAR scanning and the creation of photogrammetric digital terrain models. In addition, repeated overflights facilitate the set-up of soil monitoring systems and, thus, allow for the detection of changes in soil properties over time or season (see Archer et al. 2015, and references therein). Assuming another visit to the Sardinian test sites, a recently acquired UAV system could be used to derive new high-resolution elevation models for both investigated agricultural fields. In doing so, a better basis would be created for scorpan-based DSM as well as for general future research activities at the San Michele farm.

Focusing on data-driven learning algorithms such as neural networks, an increased number of input variables is likely to have positive implications for the overall model performance. However, the relationships between observations of a specific

soil property and its associated factors of soil formation often vary with spatial scale. Therefore, further research is required to properly dissect available covariates, for instance, using wavelet functions (Mendonça-Santos et al. 2006) and to create multi-scale modelling frameworks. The latter has been done by Behrens et al. (2014) by developing a hyper-scale terrain analysis approach (ConStat) to support DSM applications. Miller et al. (2015) recently extended this concept to a broader range of covariates including remote sensing predictor variables and geological categories. With regard to this work, both strategies appear attractive for soil mapping concerning the entire Rio di Costara catchment. The multi-layer perceptron used as data mining tool to exploit available covariates will likely be able to handle much larger predictor sets. Yet more importantly, it would be interesting to see whether modelled soil-landscape relationships become even more realistic by integrating terrain attributes from across multiple spatial scales rather than relying on one single subset of land-surface parameters.

In addition to ongoing developments to increase the number and quality of scorpan-based predictors, better analysis tools are required for the provision of reliable soil observations. This applies to the methods used to collect and analyse soil material, but also affects the prior definition of representative samples with regard to location and number. Moreover, for digital soil mapping in data-sparse regions and for general soil inventories at coarser spatial scales, the use of legacy soil data – historical maps and soil profile descriptions – is strongly advised. Sulaeman et al. (2013) described essential steps of legacy soil data harmonisation and database development for digital mapping purposes in Indonesia. As repeatedly indicated in this thesis, available soil information in the study area was generally scarce and of little value for spatial interpolation issues. By now, official authorities have started to process and store legacy soil data in an island-wide database (Database Soil Sardinia, DBSS, see Vacca et al. 2014). If this database eventually becomes accessible to the public, it might help to improve at least large-scale digital soil mapping activities on Sardinia.

Although it is commonly accepted that data quality has a higher impact on soil spatial prediction accuracy (Minasny et al. 2008), methodological progress remain a factor to advance the predictive power in digital soil mapping applications.

Regardless of specific improvements with respect to neural network modelling, recent trends in (soil) geostatistics indicate the use of multiple-point simulations (MPS). Relying on the determination of training images, MPS basically provides a more natural way to incorporate soil process knowledge into the description of spatial auto-correlation compared to common two-point statistics such as the variogram. Thus, in contrast to scorpan-based approaches, MPS tools attempt to improve the random model component rather than to look for more sophisticated functional representations of soil formation in the fixed effects. By replacing the variogram and its simplifying assumption of an underlying Gaussian process, this approach is especially promising for mapping spatially complex soil properties such as bulk density or soil moisture content. As a continuation of this work, MPS might be considered to better address the complex patterns of soil separates at field 33, where regression- and variogram-based mapping have led to relatively poor prediction results. One of the first multivariate examples of MPS in soil science is given by Meerschman et al. (2014), who simultaneously detected fossil ice-wedges and interpolated sensor data from EMI measurements. Lark (2012a) emphasized the difficulty of generating adequate training images and proposed stochastic geometric models for different modes of soil spatial variation.

Particularly at the catchment scale, this thesis has succeeded in creating digital soil maps of adequate quality and resolution. Subsequent research projects are highly appreciated to use these results, for instance, as input data in hydrological modelling or to assess soil functions and degradation risks. Carré et al. (2007) and Finke (2012) have raised the importance of converting the increasing amount of geographically continuous soil information into spatial soil functional understanding. To fulfil the needs of a digital soil assessment rather than mapping, information on the uncertainty of estimated soil property values is needed as much as three-dimensional representations of the investigated soil body. Goovaerts (2001) discussed some approaches to quantify uncertainty in geostatistical applications focusing on either prediction variances or simulation-based methods. Recently, Taalab et al. (2015) emphasized the ability of Bayesian models to formalise uncertainty. Case studies that investigate soil variation with depth are provided by Kempen et al. (2011), Lacoste et al. (2014) and Poggio and Gimona (2014).

As a final remark, it should be noted that the branch of digital soil mapping has developed in recent years from merely a research field to an operational stage. This dissertation contributes by presenting a comprehensive approach for the digital mapping and online dissemination of soil textural fractions. It, thereby, increases the data availability in the Sardinian test sites. Moreover, it provides a methodological option for a statistically sound determination of (top)soil information in cases where intensive conventional soil surveys are not feasible due to time and/or budget limitations. The proposed concept seems especially useful in times of climate-induced shortages of water resource and growing land degradation risks where modelling requires precise and spatially continuous soil knowledge. It is important to stress that all observations and results presented in this thesis primarily apply to Mediterranean landscapes. In a next step, its underlying methodology will be adapted to other regions like the loess covered hills bordering the North German Lowlands. Furthermore, it will be tested at coarser spatial scales and with respect to other environmental target quantities including meteorological variables in order to improve the modelling of soil and landscape processes.

References

Abu-Mostafa, Y. S., M. Magdon-Ismail, and H.-T. Lin (2012). *Learning From Data*. AMLBook (Cit. on pp. 10, 32).

Accornero, M. and L. Marini (2007). A water-rock interaction study in kinematic mode for the Arenarie di San Vito formation (Sardinia, Italy). In: *Proc. IMWA Symposium*, pp. 363–367 (Cit. on p. 18).

AG Boden (2005). *Bodenkundliche Kartieranleitung. 5. verbesserte und erweiterte Auflage*. Ed. by W. Eckelmann, H. Sponagel, W. Grottenthaler, K.-J. Hartmann, R. Hartwich, P. Janetzko, H. Joisten, D. Kühn, K.-J. Sabel, and R. Traidl. Stuttgart: Schweizerbart'sche Verlagsbuchhandlung (Cit. on p. 21).

Aitchison, J. (1982). The statistical analysis of compositional data. In: *Journal of the Royal Statistical Society. Series B (Methodological)* 44.2, pp. 139–177 (Cit. on pp. 9, 55).

Aitchison, J. (1986). *The statistical analysis of compositional data*. Springer (Cit. on pp. 55, 56).

Akaike, H. (1998). Information Theory and an Extension of the Maximum Likelihood Principle. In: *Selected Papers of Hirotugu Akaike*. Springer, pp. 199–213 (Cit. on p. 60).

Akpa, S. I. C., I. O. A. Odeh, T. F. A. Bishop, and A. E. Hartemink (2014). Digital Mapping of Soil Particle-Size Fractions for Nigeria. In: *Soil Science Society of America Journal* 78.6, pp. 1953–1966 (Cit. on pp. 2, 10, 148, 150).

Allen, D., M. Pringle, K. Page, and R. Dalal (2010). A review of sampling designs for the measurement of soil organic carbon in Australian grazing lands. In: *The Rangeland Journal* 32.2, pp. 227–246 (Cit. on p. 31).

Amundson, R., A. A. Berhe, J. W. Hopmans, C. Olson, A. E. Sztein, and D. L. Sparks (2015). Soil and human security in the 21st century. In: *Science* 348.6235, p. 1261071 (Cit. on p. 1).

Archer, N., B. Rawlins, S. Grebby, B. Marchant, and B. Emmett (2015). "Identify the opportunities provided by developments in earth observation and remote sensing for national scale monitoring of soil quality". This item has been internally reviewed but not externally peer-reviewed. Nottingham, UK, British Geological Survey. URL: http://nora.nerc.ac.uk/510783/ (Cit. on p. 162).

Arias, C., M. A. Brovelli, and R. Moreno (2015). Open Data, Open Specifications and Free and Open Source Software: A powerful mix to create distributed Web-based water information systems. In: *EGU General Assembly Conference Abstracts*. Vol. 17, p. 8631 (Cit. on p. 160).

Arrouays, D., N. McKenzie, J. Hempel, A. R. de Forges, and A. B. McBratney (2014). *GlobalSoilMap: Basis of the global spatial soil information system*. CRC Press (Cit. on p. 2).

Aru, A. (1966). *Rilevamento Pedologico dell'Azienda San Michele*. Centro Regionale Agrario Sperimentale, Cagliari (Cit. on pp. 21, 22, 34, 97, 98, 150).

Aru, A. (1985). The soils of Sardinia and their state of conservation. In: *Geoökodynamik* 6, pp. 71–84 (Cit. on p. 21).

Bachmaier, M. and M. Backes (2011). Variogram or Semivariogram? Variance or Semivariance? Allan Variance or Introducing a New Term? In: *Mathematical Geosciences* 43.6, pp. 735–740 (Cit. on p. 50).

Barthold, F. K., M. Wiesmeier, L. Breuer, H.-G. Frede, J. Wu, and F. B. Blank (2013). Land use and climate control the spatial distribution of soil types in the grasslands of Inner Mongolia. In: *Journal of Arid Environments* 88, pp. 194–205 (Cit. on p. 10).

Behrens, T., H. Förster, T. Scholten, U. Steinrücken, E.-D. Spies, and M. Goldschmitt (2005). Digital soil mapping using artificial neural networks. In: *Journal of Plant Nutrition and Soil Science* 168, pp. 21–33 (Cit. on pp. 10, 64).

Behrens, T., K. Schmidt, L. Ramirez-Lopez, J. Gallant, A.-X. Zhu, and T. Scholten (2014). Hyper-scale digital soil mapping and soil formation analysis. In: *Geoderma* 213, pp. 578–588 (Cit. on p. 163).

Bellocchi, G., M. Rivington, M. Donatelli, and K. Matthews (2010). Validation of biophysical models: issues and methodologies. A review. In: *Agronomy for Sustainable Development* 30.1, pp. 109–130 (Cit. on p. 79).

Belsley, D. A., E. Kuh, and R. E. Welsch (2005). *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*. Wiley (Cit. on p. 103).

Bengio, Y. and Y. Grandvalet (2004). No unbiased estimator of the variance of k-fold cross-validation. In: *The Journal of Machine Learning Research* 5, pp. 1089–1105 (Cit. on p. 149).

Benzi, R., R. Deidda, and M. Marrocu (1997). Characterization of temperature and precipitation fields over Sardinia with principal component analysis and singular spectrum analysis. In: *International Journal of Climatology* 17.11, pp. 1231–1262 (Cit. on p. 14).

Bergmeir, C. and J. M. Benítez (2012). Neural Networks in R Using the Stuttgart Neural Network Simulator: RSNNS. In: *Journal of Statistical Software* 46.7, pp. 1–26. URL: http://www.jstatsoft.org/v46/i07/ (Cit. on pp. 59, 65, 67).

Beven, K. J. and M. J. Kirkby (1979). A physically based, variable contributing area model of basin hydrology. In: *Hydrological Sciences Bulletin* 24.1, pp. 43–69 (Cit. on pp. 39, 40).

Beyer, H. L. (2011). *Geospatial Modelling Environment (version 0.5.3 Beta). (software)*. URL: http://www.spatialecology.com/gme (Cit. on pp. 33, 34).

Bishop, C. M. (1995). *Neural Networks for Pattern Recognition*. Clarendon press, Oxford (Cit. on pp. 62–67).

Bivand, R. (2014). *spdep: Spatial dependence: weighting schemes, statistics and models*. R package version 0.5-74. URL: http://CRAN.R-project.org/package=spdep (Cit. on p. 61).

Bivand, R., T. Keitt, and B. Rowlingson (2014). *rgdal: Bindings for the Geospatial Data Abstraction Library*. R package version 0.8-15. URL: http://CRAN.R-project.org/package=rgdal (Cit. on p. 54).

Bivand, R. S., E. Pebesma, and V. Gomez-Rubio (2013). *Applied spatial data analysis with R, Second edition*. Springer, New York. URL: http://www.asdar-book.org/ (Cit. on p. 54).

Boettinger, J. L., D. W. Howell, A. C. Moore, A. E. Hartemink, and S. Kienast-Brown, eds. (2010). *Digital Soil Mapping: Bridging Research, Environmental Application, and Operation*. Progress in Soil Science. Springer (Cit. on p. 7).

Boni, M., G. Balassone, L. Fedele, and N. Mondillo (2009). Post-Variscan hydrothermal activity and ore deposits in southern Sardinia (Italy): selected examples from Gerrei (Silius Vein System) and the Iglesiente district. In: *Periodico di Mineralogia* 78.3, pp. 19–35 (Cit. on p. 17).

Borovicka, T., M. Jirina Jr., P. Kordik, and M. Jirina (2012). Selecting representative data sets. In: *Advances in Data Mining Knowledge Discovery and Applications*. Ed. by A. Karahoca. InTech. Chap. 2, pp. 43–70 (Cit. on p. 48).

Brenning, A. (2008). Statistical geocomputing combining R and SAGA: The example of landslide susceptibility analysis with generalized additive models. In: *SAGA – Seconds Out*. Ed. by J. Böhner, T. Blaschke, and L. Montanarella. Vol. 19. Hamburger Beiträge zur Physischen Geographie und Landschaftsökologie. Institut für Geographie der Universität Hamburg, pp. 23–32 (Cit. on pp. 38, 40).

Breusch, T. S. and A. R. Pagan (1979). A simple test for heteroscedasticity and random coefficient variation. In: *Econometrica* 47.5, pp. 1287–1294 (Cit. on p. 61).

Brungard, C. W., J. L. Boettinger, M. C. Duniway, S. A. Wills, and T. C. Edwards (2015). Machine learning for predicting soil classes in three semi-arid landscapes. In: *Geoderma* 239, pp. 68–83 (Cit. on pp. 9, 156).

Brus, D. J. and J. J. de Gruijter (1997). Random sampling or geostatistical modelling? Choosing between design-based and model-based sampling strategies for soil (with Discussion). In: *Geoderma* 80.1–2, pp. 1–44 (Cit. on pp. 30, 31).

Brus, D. and G. Heuvelink (2007). Optimization of sample patterns for universal kriging of environmental variables. In: *Geoderma* 138, pp. 86–95 (Cit. on p. 31).

Buchanan, S., J. Triantafilis, I. O. A. Odeh, and R. Subansinghe (2012). Digital soil mapping of compositional particle-size fractions using proximal and remotely sensed ancillary data. In: *Geophysics* 77.4, pp. 201–211 (Cit. on pp. 9, 145, 156).

Burrough, P. A., J. Bouma, and S. R. Yates (1994). The state of the art in pedometrics. In: *Geoderma* 62.1, pp. 311–326 (Cit. on p. 8).

Burt, J. E., G. M. Barber, and D. L. Rigby (2009). *Elementary statistics for geographers*. Guilford Press (Cit. on pp. 47, 48).

Callegary, J. B., T. P. A. Ferré, and R. W. Groom (2007). Vertical spatial sensitivity and exploration depth of low-induction-number electromagnetic-induction instruments. In: *Vadose Zone Journal* 6.1, pp. 158–167 (Cit. on p. 44).

Carmignani, L., R. Carosi, A. Dipisa, M. Gattiglio, G. Musumeci, G. Oggiano, and P. C. Pertusati (1994). The hercynian chain in Sardinia (Italy). In: *Geodinamica Acta* 7.1, pp. 31–47 (Cit. on p. 17).

Carré, F., A. B. McBratney, T. Mayr, and L. Montanarella (2007). Digital soil assessments: Beyond DSM. In: *Geoderma* 142.1, pp. 69–79 (Cit. on p. 164).

Cassiani, G., N. Ursino, R. Deiana, G. Vignoli, J. Boaga, M. Rossi, M. T. Perri, M. Blaschek, R. Duttmann, S. Meyer, R. Ludwig, A. Soddu, P. Dietrich, and U. Werban (2012). Noninvasive Monitoring of Soil Static Characteristics and Dynamic States: A Case Study Highlighting Vegetation Effects. In: *Vadose Zone Journal* 11.3 (Cit. on pp. 42, 108, 147, 150).

Castrignanò, A., M. T. F. Wong, M. Stelluti, D. De Benedetto, and D. Sollitto (2012). Use of EMI, gamma-ray emission and GPS height as multi-sensor data for soil characterisation. In: *Geoderma* 175–176, pp. 78–89 (Cit. on pp. 12, 158).

Casula, G., A. Cherchi, L. Montadert, M. Murru, and E. Sarria (2001). The Cenozoic graben system of Sardinia (Italy): geodynamic evolution from new seismic and field data. In: *Marine and Petroleum Geology* 18, pp. 863–888 (Cit. on p. 17).

Cavazzi, S., R. Corstanje, T. Mayr, J. Hannam, and R. Fealy (2013). Are fine resolution digital elevation models always the best choice in digital soil mapping? In: *Geoderma* 195, pp. 111–121 (Cit. on p. 146).

Cellura, M., G. Cirrincione, A. Marvuglia, and A. Miraoui (2008). Wind speed spatial estimation for energy planning in Sicily: A neural kriging application. In: *Renewable Energy* 33.6, pp. 1251–1266 (Cit. on pp. 127, 155).

Chessa, P. A. and A. Delitala (1997). Objective analysis of daily extreme temperatures of Sardinia (Italy) using distance from the sea as independent variable. In: *International Journal of Climatology* 17.13, pp. 1467–1485 (Cit. on p. 14).

Chiles, J.-P. and P. Delfiner (2012). *Geostatistics: Modeling Spatial Uncertainty*. 2nd ed. Vol. 497. Wiley (Cit. on p. 71).

Christensen, R. (1991). *Linear Models for Multivariate, Time Series, and Spatial Data*. Springer (Cit. on pp. 53, 69–71, 77).

Cimmery, V. (2007). *User Guide for SAGA (version 2.0)* (Cit. on p. 40).

Clark, I. (1979). *Practical geostatistics*. Applied Science Publishers London (Cit. on p. 51).

Clay, D. E., J. Chang, D. D. Malo, C. G. Carlson, C. Reese, S. A. Clay, M. Ellsbury, and B. Berg (2001). Factors influencing spatial variability of soil apparent electrical conductivity. In: *Communications in Soil Science and Plant Analysis* 32.19–20, pp. 2993–3008 (Cit. on p. 43).

Cockx, L., M. Van Meirvenne, U. W. A. Vitharana, L. P. C. Verbeke, D. Simpson, T. Saey, and F. M. B. Van Coillie (2009). Extracting topsoil information from EM38DD sensor data using a neural network approach. In: *Soil Science Society of America Journal* 73.6, pp. 2051–2058 (Cit. on pp. 12, 44).

Conrad, O. (2006). SAGA: Entwurf, Funktionsumfang und Anwendung eines Systems für Automatisierte Geowissenschaftliche Analysen. PhD thesis. Göttingen University (Cit. on p. 40).

Conti, P., L. Carmignani, and A. Funedda (2001). Change of nappe transport direction during the Variscan collisional evolution of central-southern Sardinia (Italy). In: *Tectonophysics* 332.1–2, pp. 255–273 (Cit. on p. 17).

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. In: *Technometrics* 19.1, pp. 15–18 (Cit. on p. 61).

Cook, R. D. and S. Weisberg (1983). Diagnostics for heteroscedasticity in regression. In: *Biometrika* 70.1, pp. 1–10 (Cit. on p. 61).

Cook, S. E., R. J. Corner, P. R. Groves, and G. J. Grealish (1996). Use of airborne gamma radiometric data for soil mapping. In: *Australian Journal of Soil Research* 34.1, pp. 183–194 (Cit. on pp. 45, 157).

Corwin, D. L. and S. M. Lesch (2005). Apparent soil electrical conductivity measurements in agriculture. In: *Computers and Electronics in Agriculture* 46.1, pp. 11–43 (Cit. on pp. 42, 43).

Cressie, N. (1990). The Origins of Kriging. In: *Mathematical Geology* 22.3, pp. 239–252 (Cit. on pp. 8, 69).

Cressie, N. and C. K. Wikle (1998). The Variance-Based Cross-Variogram: You Can Add Apples and Oranges. In: *Mathematical Geology* 30.7, pp. 789–799 (Cit. on p. 74).

Cressie, N. A. (1993). *Statistics for spatial data*. Wiley, New York (Cit. on pp. 51–53, 69, 71, 77).

Critchfield, H. J. (1983). *General Climatology*. Prentice-Hall, Inc. Englewood Cliffs, NJ (Cit. on p. 14).

Dai, F., Q. Zhou, Z. Lv, X. Wang, and G. Liu (2014). Spatial prediction of soil organic matter content integrating artificial neural network and ordinary kriging in Tibetan Plateau. In: *Ecological Indicators* 45, pp. 184–194 (Cit. on pp. 11, 155).

Dale, M. R. T. and M.-J. Fortin (2014). *Spatial Analysis: A Guide For Ecologists*. 2nd ed. Cambridge University Press (Cit. on p. 8).

Dalgaard, P. (2008). *Introductory statistics with R*. Springer (Cit. on p. 47).

Dalu, G. A. and A. Cima (1983). Three-dimensional airflow over Sardinia. In: *Il Nuovo Cimento C* 6.5, pp. 453–472 (Cit. on pp. 15, 16).

Daszykowski, M., B. Walczak, and D. L. Massart (2002). Representative subset selection. In: *Analytica Chimica Acta* 468.1, pp. 91–103 (Cit. on p. 57).

Davis, B. M. (1987). Uses and Abuses of Cross-Validation in Geostatistics. In: *Mathematical Geology* 19.3, pp. 241–248 (Cit. on p. 82).

De Benedetto, D., A. Castrignano, D. Sollitto, F. Modugno, G. Buttafuoco, and G. lo Papa (2012). Integrating geophysical and geostatistical techniques to map the spatial variation of clay. In: *Geoderma* 171–172, pp. 53–63 (Cit. on pp. 43, 147, 158).

De Gruijter, J. J. and C. J. F. ter Braak (1990). Model-free estimation from spatial samples: a reappraisal of classical sampling theory. In: *Mathematical Geology* 22.4, pp. 407–415 (Cit. on p. 31).

De Gruijter, J. J., D. J. J. Walvoort, and P. F. M. Van Gams (1997). Continuous soil maps - a fuzzy set approach to bridge the gap between aggregation levels of process and distribution models. In: *Geoderma* 77.2, pp. 169–195 (Cit. on p. 9).

De Gruijter, J. J., D. J. Brus, M. F. P. Bierkens, and M. Knotters (2006). *Sampling for natural resource monitoring*. Springer (Cit. on pp. 31, 34).

Delitala, A., D. Cesari, P. A. Chessa, and M. N. Ward (2000). Precipitation over Sardinia (Italy) during the 1946–1993 rainy seasons and associated large-scale climate variations. In: *International Journal of Climatology* 20.5, pp. 519–541 (Cit. on p. 15).

Demyanov, V., M. Kanevski, S. Chernov, E. Savelieva, and V. Timonin (1998). Neural network residual kriging application for climatic data. In: *Journal of Geographic Information and Decision Analysis* 2.2, pp. 215–232 (Cit. on p. 127).

Dierke, C. and U. Werban (2013). Relationships between gamma-ray data and soil properties at an agricultural test site. In: *Geoderma* 199, pp. 90–98 (Cit. on pp. 12, 44, 45).

Diggle, P. J. and P. J. Ribeiro Jr. (2007). *Model-based Geostatistics*. Springer, New York (Cit. on pp. 8, 71).

Duttmann, R. and K. Sumfleth (2007). Predictive mapping of soil characteristics in paddy rice landscapes of the central eastern Jiangxi Province/China. In: *GEOÖKO* 28, pp. 72–103 (Cit. on p. 41).

EC (2007). DIRECTIVE 2007/2/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 14 March 2007 establishing an Infrastructure for Spatial Information in the European Community (INSPIRE). In: *Official Journal of the European Union* L 108. European Commission, pp. 1–14 (Cit. on p. 83).

EC (2008). COMMISSION REGULATION (EC) No 1205/2008 of 3 December 2008 implementing Directive 2007/2/EC of the European Parliament and of the Council as regards metadata (Text with EEA relevance). In: *Official Journal of the European Union* L.326, pp. 12–30 (Cit. on p. 83).

EC (2009). COMMISSION REGULATION (EC) No 976/2009 of 19 October 2009 implementing Directive 2007/2/EC of the European Parliament and of the

Council as regards the Network Services. In: *Official Journal of the European Union* L.274, pp. 9–18 (Cit. on p. 83).

EC (2011). *Technical Guidance for the implementation of INSPIRE View Services*. Tech. rep. 3.1. Joint Research Centre (Cit. on p. 83).

EC (2013). *INSPIRE Metadata Implementing Rules: Technical Guidelines based on EN ISO 19115 and EN ISO 19119*. Tech. rep. 1.3. Joint Research Centre (Cit. on p. 83).

Egozcue, J. J., V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal (2003). Isometric Logratio Transformations for Compositional Data Analysis. In: *Mathematical Geology* 35.3, pp. 279–300 (Cit. on p. 55).

Ehlers, L., F. Herrmann, M. Blaschek, R. Duttmann, and F. Wendland (2015). Sensitivity of mGROWA-simulated groundwater recharge to changes in soil and land use parameters in a Mediterranean environment and conclusions in view of ensemble-based climate impact simulations. In: *Science of The Total Environment*, pp. –. ISSN: 0048-9697. DOI: http://dx.doi.org/10.1016/j.scitotenv.2015.04.122. URL: http://www.sciencedirect.com/science/article/pii/S0048969715300565 (Cit. on p. 153).

Esbensen, K. H. and P. Geladi (2010). Principles of Proper Validation: use and abuse of re-sampling for validation. In: *Journal of Chemometrics* 24.3-4, pp. 168–187 (Cit. on p. 82).

Fanshawe, T. R. and P. J. Diggle (2012). Bivariate geostatistical modelling: a review and an application to spatial variation in radon concentrations. In: *Environmental and Ecological Statistics* 19.2, pp. 139–160 (Cit. on p. 76).

FAO (2006). *Guidelines for soil description*. Ed. by R. Jahn, H. P. Blume, V. B. Asio, O. Spaargaren, and P. Schad. Food and Agriculture Organization of the United Nations, Rome (Cit. on p. 120).

Faraway, J. J. (2004). *Linear Models with R*. CRC Press (Cit. on p. 61).

Feiden, K., F. Kruse, V. Epitropou, and K. Karatzas (2010). The GS SOIL portal prototype and its integrated network. In: *24th International Conference on Informatics for Environmental Protection in cooperation with Intergeo*. Ed. by K. Greve. Shaker, pp. 420–428 (Cit. on p. 160).

Finke, P. A. (2012). On digital soil assessment with models and the Pedometrics agenda. In: *Geoderma* 171, pp. 3–15 (Cit. on p. 164).

Florinsky, I. V., R. G. Eilers, G. R. Manning, and L. G. Fuller (2002). Prediction of soil properties by digital terrain modelling. In: *Environmental Modelling and Software* 17.3, pp. 295–311 (Cit. on p. 41).

Florinsky, I. V. (2011). *Digital terrain analysis in soil science and geology*. Academic Press (Cit. on pp. 37, 38).

Fox, J. and S. Weisberg (2011). *An R Companion to Applied Regression*. 2nd ed. Sage Publications (Cit. on p. 61).

Funedda, A., L. Carmignani, P. C. Pertusati, A. Forci, P. Calzia, F. Marongiu, G. Pisanu, and M. Serra (2013). *Note illustrative della Carta Geologica d'Italia alla scala 1:50.000. Foglio 548 Senorbi*. Servizio Geologico d'Italia, Regione Autonoma della Sardegna, Organo Cartografico dello Stato. URL: http://www.isprambiente.gov.it/Media/carg/note_illustrative/548_Senorbi.pdf (visited on 05/13/2013) (Cit. on pp. 17–19).

Gabriel, K. R. (1971). The biplot graphic display of matrices with application to principal component analysis. In: *Biometrika* 58.3, pp. 453–467 (Cit. on p. 49).

Galvão, R. K. H., M. C. U. Araujo, G. E. José, M. J. C. Pontes, E. C. Silva, and T. C. B. Saldanha (2005). A method for calibration and validation subset partitioning. In: *Talanta* 67.4, pp. 736–740 (Cit. on pp. 57, 155).

Garson, D. G. (1991). Interpreting neural-network connection weights. In: *AI Expert* 6.4, pp. 47–51 (Cit. on pp. 67, 129, 130).

Gauch, H. G., J. T. Hwang, and G. W. Fick (2003). Model Evaluation by Comparison of Model-Based Predictions and Measured Values. In: *Agronomy Journal* 95.6, pp. 1442–1446 (Cit. on p. 81).

GDI-DE (2008). *Deutsche Übersetzung der Metadatenfelder des ISO 19115 Geographic information Metadata*. Tech. rep. AK Geodienste, GDI-DE Koordinierungsstelle (Cit. on p. 83).

GDI-DE (2011). *Handlungsempfehlungen für die Bereitstellung von INSPIRE konformen Darstellungsdiensten (INSPIRE View Services)*. Tech. rep. 1.0. AK Geodienste, GDI-DE Koordinierungsstelle (Cit. on p. 83).

Gee, G. W. and D. Or (2002). Particle-size analysis. In: *Methods of Soil Analysis. Part 4. Physical Methods*. Ed. by J. H. Dane and C. Topp. SSSA Book Series: 5. Soil Science Society of America. Chap. 2.4, pp. 255–293 (Cit. on p. 36).

Gessler, P. E., I. D. Moore, N. J. McKenzie, and P. J. Ryan (1995). Soil-landscape modelling and spatial prediction of soil attributes. In: *International Journal of Geographical Information Systems* 9, pp. 421–432 (Cit. on pp. 8, 9, 31).

Gevrey, M., I. Dimopoulos, and S. Lek (2003). Review and comparison of methods to study the contribution of variables in artificial neural network models. In: *Ecological Modelling* 160.3, pp. 249–264 (Cit. on pp. 67, 130).

Gilmore, G. R. (2008). *Practical Gamma-ray Spectrometry*. 2nd ed. Wiley (Cit. on p. 46).

Giorgi, F. and P. Lionello (2008). Climate change projections for the Mediterranean region. In: *Global and Planetary Change* 63.2, pp. 90–104 (Cit. on p. 1).

Gobin, A., P. Campling, and J. Feyen (2001). Soil-Landscape Modelling to Quantify Spatial Variability of Soil Texture. In: *Physics and Chemistry of the Earth, Part B: Hydrology, Oceans and Atmosphere* 26.1, pp. 41–45 (Cit. on pp. 11, 144).

Goldberger, A. S. (1962). Best Linear Unbiased Prediction in the Generalized Linear Regression Model. In: *Journal of the American Statistical Association* 57.298, pp. 369–375 (Cit. on p. 70).

Goovaerts, P. (1997). *Geostatistics for Natural Resources Evaluation*. Oxford University Press (Cit. on pp. 11, 47, 48, 72, 74, 75).

Goovaerts, P. (2001). Geostatistical modelling of uncertainty in soil science. In: *Geoderma* 103.1, pp. 3–26 (Cit. on p. 164).

Goulard, M. and M. Voltz (1992). Linear Coregionalization Model: Tools for Estimation and Choice of Cross-Variogram Matrix. In: *Mathematical Geology* 24.3, pp. 269–286 (Cit. on p. 75).

Gray, J. M., T. F. A. Bishop, and J. R. Wilford (2014). Lithology as a powerful covariate in digital soil mapping. In: *GlobalSoilMap: Basis of the global spatial soil information system*. Ed. by D. Arrouays, N. McKenzie, J. Hempel, A. R. de Forges, and A. B. McBratney. CRC Press, pp. 433–440 (Cit. on p. 41).

Greenacre, M. (2010). *Biplots in Practice*. Fundación BBVA (Cit. on pp. 49, 95).

Greve, M. H., R. B. Kheir, M. B. Greve, and P. K. Bøcher (2012). Quantifying the ability of environmental parameters to predict soil texture fractions using

regression-tree model with GIS and LIDAR data: The case study of Denmark. In: *Ecological Indicators* 18, pp. 1–10 (Cit. on pp. 2, 10, 144, 156).

Grimm, R., M. Behrens T.and Märker, and H. Elsenbeer (2008). Soil organic carbon concentrations and stocks on Barro Colorado Islanddigital soil mapping using Random Forests analysis. In: *Geoderma* 146.1, pp. 102–113 (Cit. on p. 10).

Grunwald, S. (2006). *Environmental Soil-Landscape Modeling: Geographic Information Technologies and Pedometrics*. CRC Press (Cit. on p. 8).

Grunwald, S. (2009). Multi-criteria characterization of recent digital soil mapping and modeling approaches. In: *Geoderma* 152.3, pp. 195–207 (Cit. on p. 8).

GSDI (2004). *Developing Spatial Data Infrastructures: The SDI Cookbook*. Ed. by D. D. Nebert. 2.0. Global Spatial Data Infrastructure (Cit. on p. 84).

Guo, P.-T., W. Wu, Q.-K. Sheng, M.-F. Li, H.-B. Liu, and Z.-Y. Wang (2013). Prediction of soil organic matter using artificial neural network and topographic indicators in hilly areas. In: *Nutrient Cycling in Agroecosystems* 95.3, pp. 333–344 (Cit. on pp. 10, 154).

Hartemink, A. E., A. B. McBratney, and M. de Lourdes Mendonça-Santos, eds. (2008). *Digital Soil Mapping with Limited Data*. Springer (Cit. on p. 7).

Hartge, K. H. and R. Horn (2009). *Die physikalische Untersuchung von Böden: Praxis, Messmethoden, Auswertung*. 4th ed. Schweizerbart'sche Verlagsbuchhandlung (Cit. on p. 36).

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The Elements of Statistical Learning*. 2nd ed. Springer (Cit. on pp. 10, 60).

Hengl, T., D. Rossiter, and A. Stein (2003). Soil sampling strategies for spatial prediction by correlation with auxiliary maps. In: *Australian Journal of Soil Research* 41, pp. 1403–1422 (Cit. on pp. 31, 32).

Hengl, T. (2003). Pedometric Mapping: bridging the gaps between conventional and pedometric approaches. PhD thesis. Wageningen University (Cit. on p. 7).

Hengl, T. (2006). Finding the right pixel size. In: *Computers & Geosciences* 32, pp. 1283–1298 (Cit. on pp. 57, 58).

Hengl, T. (2009). *A Practical Guide to Geostatistical Mapping*. Amsterdam (Cit. on pp. 8, 31, 52, 68, 77, 82).

Hengl, T. (2014). *plotKML: Visualization of spatial and spatio-temporal objects in Google Earth*. R package version 0.4-3/r485. URL: http://R-Forge.R-project.org/projects/plotkml/ (Cit. on p. 85).

Hengl, T., G. B. M. Heuvelink, and A. Stein (2004). A generic framework for spatial prediction of soil variables based on regression-kriging. In: *Geoderma* 120.1–2, pp. 75–93 (Cit. on pp. 11, 49, 76, 148).

Hengl, T., G. B. M. Heuvelink, and D. G. Rossiter (2007). About regression-kriging: From equations to case studies. In: *Comput. Geosci.* 33, pp. 1301–1315 (Cit. on pp. 11, 77).

Hengl, T., J. M. de Jesus, R. A. MacMillan, N. H. Batjes, G. B. M. Heuvelink, E. Ribeiro, A. Samuel-Rosa, B. Kempen, J. G. B. Leenaars, M. G. Walsh, and M. R. Gonzalez (2014). SoilGrids1km – Global Soil Information Based on Automated Mapping. In: *PLOS ONE* 9.8 (Cit. on p. 2).

Hengl, T., G. B. M. Heuvelink, B. Kempen, J. G. B. Leenaars, M. G. Walsh, K. D. Shepherd, A. Sila, R. A. MacMillan, J. M. de Jesus, L. Tamene, and J. E. Tondoh (2015). Mapping Soil Properties of Africa at 250 m Resolution: Random Forests Significantly Improve Current Predictions. In: *PLOS ONE* 10.6, e0125814 (Cit. on p. 2).

Herrmann, F., N. Baghdadi, M. Blaschek, R. Deidda, R. Duttmann, I. L. Jeunesse, H. Sellami, H. Vereecken, and F. Wendland (2015). Simulation of future groundwater recharge using a climate model ensemble and SAR-image based soil parameter distributions A case study in an intensively-used Mediterranean catchment. In: *Science of The Total Environment*, pp. –. ISSN: 0048-9697. DOI: http://dx.doi.org/10.1016/j.scitotenv.2015.07.036. URL: http://www.sciencedirect.com/science/article/pii/S004896971530379X (Cit. on p. 153).

Heuvelink, G. B. M. (2003). The Definition of Pedometrics. In: *Pedometron, IUSS* 15, pp. 11–12. URL: http://pedometrics.org/ (Cit. on p. 7).

Hijmans, R., N. Garcia, A. Rala, A. Maunahan, J. Weiczorek, and J. Kapoor (2012). *GADM database of Global Administrative Areas*. URL: www.gadm.org (Cit. on p. 13).

IAEA (2003). *Guidelines for radioelement mapping using gamma ray spectrometry data*. Tech. rep. 1363. International Atomic Energy Agency, Vienna (Cit. on pp. 44, 46).

Igel, C. and M. Hüsken (2000). Improving the Rprop Learning Algorithm. In: *Proceedings of the Second ICSC International Symposium on Neural Computation (NC 2000)*, pp. 115–121 (Cit. on p. 64).

IPCC (2014). *Climate Change 2014: Impacts, Adaptation, and Vulnerability. Part B: Regional Aspects. Contribution of Working Group II to the Fifth Assessment Report of the Intergovernmental Panel on Climate Change*. Ed. by V. R. Barros, C. B. Field, D. J. Dokken, M. D. Mastrandrea, K. J. Mach, T. E. Bilir, M. Chatterjee, K. L. Ebi, Y. O. Estrada, R. C. Genova, B. Girma, E. S. Kissel, A. N. Levy, S. MacCracken, P. R. Mastrandrea, and L. L. White. Cambridge University Press (Cit. on p. 2).

IUSS Working Group WRB (2006). *World reference base for soil resources 2006*. World Soil Resources Reports 103. FAO, Rome (Cit. on p. 22).

Jenny, H. (1941). *Factors of Soil Formation*. McGraw Hill (Cit. on pp. xiv, 9, 37, 41).

Jolliffe, I. (2002). *Principal Component Analysis*. 2nd ed. Springer (Cit. on p. 49).

Jouan-Rimbaud, D., D. Massart, C. Saby, and C. Puel (1997). Characterisation of the representativity of selected sets of samples in multivariate calibration and pattern recognition. In: *Analytica Chimica Acta* 350.1, pp. 149–161 (Cit. on p. 48).

Journel, A. G. and C. J. Huijbregts (1978). *Mining Geostatistics*. Academic Press, New York (Cit. on pp. 8, 51).

Kanevski, M., R. Parkin, A. Pozdnukhov, V. Timonin, M. Maignan, V Demyanov, and S. Canu (2004). Environmental data mining and modeling based on machine learning algorithms and geostatistics. In: *Environmental Modelling & Software* 19.9, pp. 845–855 (Cit. on pp. 11, 127).

Kanevski, M. and M. Maignan (2004). *Analysis and modelling of spatial environmental data*. EPFL Press (Cit. on pp. 47, 68).

Kanevski, M., V. Demyanov, and M. Maignan (1997). Mapping of Soil Contamination by Using Artificial Neural Networks and Multivariate Geostatistics. In: *Artificial Neural Networks – ICANN '97*. Ed. by W. Gerstner, A. Germond,

M. Hasler, and J.-D. Nicoud. Vol. 1327. Lecture Notes in Computer Science. Springer, pp. 1125–1130 (Cit. on pp. 11, 155).

Karlis, D., G. Saporta, and A. Spinakis (2003). A Simple Rule for the Selection of Principal Components. In: *Communications in Statistics-Theory and Methods* 32.3, pp. 643–666 (Cit. on pp. 49, 96, 100).

Kempen, B., D. J. Brus, and J. J. Stoorvogel (2011). Three-dimensional mapping of soil organic matter content using soil type–specific depth functions. In: *Geoderma* 162.1, pp. 107–123 (Cit. on p. 164).

Kennard, R. W. and L. A. Stone (1969). Computer Aided Design of Experiments. In: *Technometrics* 11.1, pp. 137–148 (Cit. on p. 57).

Kerry, R. and M. A. Oliver (2008). Determining nugget: sill ratios of standardized variograms from aerial photographs to krige sparse soil data. In: *Precision Agriculture* 9.1-2, pp. 33–56 (Cit. on pp. 32, 53).

Kim, J., S. Grunwald, R. G. Rivero, and R. Robbins (2012). Multi-scale Modeling of Soil Series Using Remote Sensing in a Wetland Ecosystem. In: *Soil Science Society of America Journal* 76.6, pp. 2327–2341 (Cit. on p. 10).

Kitanidis, P. K. (1993). Generalized Covariance Functions in Estimation. In: *Mathematical Geology* 25.5, pp. 525–540 (Cit. on p. 77).

Kitanidis, P. K. (1997). *Introduction to Geostatistics: Applications to Hydrogeology*. Cambridge University Press (Cit. on p. 8).

Kobayashi, K. and M. U. Salam (2000). Comparing Simulated and Measured Values Using Mean Squared Deviation and its Components. In: *Agronomy Journal* 92.2, pp. 345–352 (Cit. on p. 81).

Kovačević, M., B. Bajat, and B. Gajić (2010). Soil type classification and estimation of soil properties using support vector machines. In: *Geoderma* 154.3, pp. 340–347 (Cit. on p. 10).

Krüger, K. (2007). Regionalisierung von Bodeneigenschaften unter Verwendung digitaler Geländemodelle sowie multispektraler Fernerkundungsdaten am Beispiel einer Agrarlandschaft im Jungmoränengebiet Schleswig-Holsteins. PhD thesis. Kiel University (Cit. on p. 150).

Kruse, F. and S. Konstantinidis (2010). InGrid®– eine Software zum Aufbau von Umweltinformationssystemen. In: *Angewandte Geoinformatik 2010 – 22.*

AGIT-Symposium. Ed. by J. Strobl, T. Blaschke, and G. Griesebner. Wichmann, pp. 119–124 (Cit. on p. 160).

Kuang, B., H. S. Mahmood, M. Z. Quraishi, W. B. Hoogmoed, A. M. Mouazen, and E. J. van Henten (2012). Sensing Soil Properties in the Laboratory, In Situ, and On-Line: A Review. In: *Advances in Agronomy* 114, pp. 155–223 (Cit. on p. 43).

Kutner, M. H., C. J. Nachtsheim, and J. Neter (2004). *Applied Linear Regression Models*. McGraw–Hill (Cit. on p. 61).

La Jeunesse, I., C. Cirelli, D. Aubin, C. Larrue, H. Sellami, S. Afifi, A. Bellin, S. Benabdallah, D. N. Bird, R. Deidda, M. Dettori, G. Engin, F. Herrmann, R. Ludwig, B. Mabrouk, B. Majone, C. Paniconi, and A. Soddu (2015). Is climate change a threat for water uses in the Mediterranean region? Results from a survey at local scale. In: *Science of The Total Environment* (Cit. on p. 1).

Lacoste, M., B. Minasny, A. B. McBratney, D. Michot, V. Viaud, and C. Walter (2014). High resolution 3D mapping of soil organic carbon in a heterogeneous agricultural landscape. In: *Geoderma* 213, pp. 296–311 (Cit. on pp. 10, 164).

Lagacherie, P. and A. B. McBratney (2006). Spatial soil information systems and spatial soil inference systems: perspectives for digital soil mapping. In: *Digital Soil Mapping: An Introductory Perspective*. Ed. by P. Lagacherie, A. B. McBratney, and M. Voltz. Vol. 31. Developments in Soil Science. Elsevier, pp. 3–24 (Cit. on p. 7).

Lagacherie, P., A. B. McBratney, and M. Voltz, eds. (2006). *Digital Soil Mapping: An Introductory Perspective*. Vol. 31. Developments in Soil Science. Elsevier (Cit. on p. 7).

Landis, J. R. and G. G. Koch (1977). The Measurement of Observer Agreement for Categorical Data. In: *Biometrics* 33, pp. 159–174 (Cit. on p. 139).

Lark, R. M. (2000). Estimating variograms of soil properties by the method-of-moments and maximum likelihood. In: *European Journal of Soil Science* 51.4, pp. 717–728 (Cit. on p. 52).

Lark, R. M. and T. F. A. Bishop (2007). Cokriging particle size fractions of the soil. In: *European Journal of Soil Science* 58.3, pp. 763–774 (Cit. on pp. 9, 55, 56, 81, 147, 151, 152).

Lark, R. M., B. R. Cullis, and S. J. Welham (2006). On spatial prediction of soil properties in the presence of a spatial trend: the empirical best linear unbiased predictor (E-BLUP) with REML. In: *European Journal of Soil Science* 57.6, pp. 787–799 (Cit. on p. 77).

Lark, R. M. (2005). Soil Properties and Pedometrics. In: *Land Use, Land Cover and Soil Sciences*. Ed. by W. H. Verheye. Vol. 6. UNESCO-EOLSS Publishers, pp. 86–109 (Cit. on p. 8).

Lark, R. M. (2012a). A stochastic geometric model for continuous local trends in soil variation. In: *Geoderma* 189, pp. 661–670 (Cit. on p. 164).

Lark, R. M. (2012b). Towards soil geostatistics. In: *Spatial Statistics* 1, pp. 92–99 (Cit. on p. 8).

Lausch, A., S. Zacharias, C. Dierke, M. Pause, I. Kühn, D. Doktor, P. Dietrich, and U. Werban (2013). Analysis of Vegetation and Soil Patterns using Hyperspectral Remote Sensing, EMI, and Gamma-Ray Measurements. In: *Vadose Zone Journal* 12.4 (Cit. on p. 44).

Lawson, A. B. (2013). *Statistical Methods in Spatial Epidemiology*. Wiley (Cit. on p. 8).

Lennartz, B., R. Horn, R. Duttmann, H. H. Gerke, R. Tippkötter, T. Eickhorst, I. Janssen, M. Janssen, B. Rüth, T. Sander, X. Shi, K. Sumfleth, H. Taubner, and B. Zhang (2009). Ecological safe management of terraced rice paddy landscapes. In: *Soil and Tillage Research* 102.2, pp. 179–192 (Cit. on p. 145).

Ließ, M., B. Glaser, and B. Huwe (2012). Uncertainty in the spatial prediction of soil texture: Comparison of regression tree and Random Forest models. In: *Geoderma* 170, pp. 70–79 (Cit. on pp. 2, 10, 144, 148, 156).

Ludwig, R., A. Soddu, R. Duttmann, N. Baghdadi, S. Benabdallah, R. Deidda, M. Marrocu, G. Strunz, F. Wendland, G. Engin, C. Paniconi, F. Prettenthaler, I. Lajeunesse, S. Afifi, G. Cassiani, A. Bellin, B. Mabrouk, H. Bach, and T. Ammerl (2010). Climate-induced changes on the hydrology of Mediterranean basins – A research concept to reduce uncertainty and quantify risk. In: *Fresenius Environmental Bulletin* 19.10 (Cit. on pp. 3, 13, 42).

Mahmood, H. S., W. B. Hoogmoed, and E. J. van Henten (2012). Sensor data fusion to predict multiple soil properties. In: *Precision Agriculture* 13.6, pp. 628–645 (Cit. on p. 149).

Malone, B. P., A. B. McBratney, B. Minasny, and G. M. Laslett (2009). Mapping continuous depth functions of soil carbon storage and available water capacity. In: *Geoderma* 154.1, pp. 138–152 (Cit. on p. 10).

Mancosu, A. (2012). Taphonomy and Palaeoecology of multi-element skeleton invertebrates: a genetic model for exceptional preservation. PhD thesis. University of Cagliari (Cit. on pp. 18, 19).

Marchant, B. P. and R. M. Lark (2007). Estimation of Linear Models of Coregionalization by Residual Maximum Likelihood. In: *European Journal of Soil Science* 58.6, pp. 1506–1513 (Cit. on p. 75).

Martín-Fernández, J. A., R. A. Olea-Meneses, and V. Pawlowsky-Glahn (2001). Criteria to Compare Estimation Methods of Regionalized Compositions. In: *Mathematical Geology* 33.8, pp. 889–909 (Cit. on p. 81).

Mascaro, G, M Piras, R Deidda, and E. Vivoni (2013). Distributed hydrologic modeling of a sparsely monitored basin in Sardinia, Italy, through hydrometeorological downscaling. In: *Hydrology and Earth System Sciences* 17.10, pp. 4143–4158 (Cit. on pp. 14, 15).

Matheron, G. (1969). *Le krigeage universel*. Ed. by E. des Mines de Paris. Vol. 1. Les Cahiers du Centre de Morphologie Mathematique de Fontainebleau (Cit. on p. 11).

Matheron, G. (1971). *The theory of regionalized variables and its applications*. Vol. 5. École national supérieure des mines (Cit. on pp. 8, 50, 69).

Matheron, G. (1973). The intrinsic random functions and their applications. In: *Advances in Applied Probability* 5.3, pp. 439–468 (Cit. on p. 69).

Mayer, D. G., M. A. Stuart, and A. J. Swain (1994). Regression of Real-World Data on Model Output: An Appropriate Overall Test of Validity. In: *Agricultural Systems* 45.1, pp. 93–104 (Cit. on p. 81).

Maynard, J. J. and M. G. Johnson (2014). Scale-dependency of LiDAR derived terrain attributes in quantitative soil-landscape modeling: Effects of grid resolution vs. neighborhood extent. In: *Geoderma* 230, pp. 29–40 (Cit. on p. 146).

McBratney, A. B., M. L. Mendonca Santos, and B. Minasny (2003). On digital soil mapping. In: *Geoderma* 117, pp. 3–52 (Cit. on pp. xvi, 8, 9, 37, 41, 82).

McBratney, A. B., I. O. A. Odeh, T. F. A. Bishop, M. S. Dunbar, and T. M. Shatar (2000). An overview of pedometric techniques for use in soil survey. In: *Geoderma* 97.3–4, pp. 293–327 (Cit. on pp. 8, 11).

McBratney, A. B., B. Minasny, and B. M. Whelan (2005). Obtaining 'useful' high-resolution soil data from proximally-sensed electrical conductivity/resistivity (PSEC/R) surveys. In: *Precision Agriculture*. Ed. by J. V. Stafford. Wageningen Academic Publishers, pp. 503–510 (Cit. on p. 157).

McCulloch, W. S. and W. Pitts (1943). A logical calculus of the ideas immanent in nervous activity. In: *Bulletin of Mathematical Biophysics* 5, pp. 115–133 (Cit. on p. 61).

McKenzie, N. J. and P. J. Ryan (1999). Spatial prediction of soil properties using environmental correlation. In: *Geoderma* 89, pp. 67–94 (Cit. on pp. 9, 31, 32).

McKenzie, N. J., P. E. Gessler, P. J. Ryan, and D. A. OConnell (2000). The Role of Terrain Analysis in Soil Mapping. In: *Terrain Analysis: Principles and Applications*. Ed. by J. P. Wilson and J. C. Gallant. Wiley, pp. 245–265 (Cit. on p. 144).

McNeill, J. D. (1980). *Electromagnetic terrain conductivity measurement at low induction numbers*. Tech. rep. 6. Geonics Limited, Mississauga, ON, Canada (Cit. on p. 43).

Meerschman, E., M. Van Meirvenne, G. Mariethoz, M. M. Islam, P. De Smedt, E. Van De Vijver, and T. Saey (2014). Using bivariate multiple-point statistics and proximal soil sensor data to map fossil ice-wedge polygons. In: *Geoderma* 213, pp. 571–577 (Cit. on pp. 159, 164).

Megumi, K. and T. Mamuro (1977). Concentration of Uranium Series Nuclides in Soil Particles in Relation to Their Size. In: *Journal of Geophysical Research* 82.2, pp. 353–356 (Cit. on pp. 45, 145, 156).

Mehra, O. P. and M. L. Jackson (1960). Iron oxide removal from soils and clays by a dithionite-citrate system buffered with sodium bicarbonate. In: *Proceedings of the 7th National Conference on Clays and Clay Minerals*. Vol. 5, pp. 317–327 (Cit. on p. 36).

Menafoglio, A., A. Guadagnini, and P. Secchi (2014). A kriging approach based on Aitchison geometry for the characterization of particle-size curves in hetero-

geneous aquifers. In: *Stochastic Environmental Research and Risk Assessment* 28.7, pp. 1835–1851 (Cit. on p. 152).

Mendonça-Santos, M. L., A. B. McBratney, and B. Minasny (2006). Soil prediction with spatially decomposed environmental factors. In: *Digital Soil Mapping: An Introductory Perspective*. Ed. by P. Lagacherie, A. B. McBratney, and M. Voltz. Vol. 31. Developments in Soil Science. Elsevier, pp. 269–278 (Cit. on p. 163).

Miller, B. A., S. Koszinski, M. Wehrhan, and M. Sommer (2015). Impact of multi-scale predictor selection for modeling soil properties. In: *Geoderma* 239, pp. 97–106 (Cit. on p. 163).

Minasny, B. and A. McBratney (2006). A conditioned Latin hypercube method for sampling in the presence of ancillary information. In: *Computers & Geosciences* 32.9, pp. 1378–1388 (Cit. on pp. 31, 32).

Minasny, B. and A. B. McBratney (2007). Spatial prediction of soil properties using EBLUP with the Matrn covariance function. In: *Geoderma* 140, pp. 324–336 (Cit. on p. 70).

Minasny, B., A. B. McBratney, and R. M. Lark (2008). Digital Soil Mapping Technologies for Countries with Sparse Data Infrastructures. In: *Digital soil mapping with limited data*. Ed. by A. E. Hartemink, A. B. McBratney, and M. de Lourdes Mendonça-Santos. Springer. Chap. 2, pp. 15–30 (Cit. on pp. 8, 11, 32, 154, 163).

Minasny, B., B. P. Malone, and A. B. McBratney, eds. (2012). *Digital Soil Assessments and Beyond*. Proceedings of the 5th Global Workshop on Digital Soil Mapping 2012, Sydney, Australia. CRC Press (Cit. on p. 7).

Minasny, B., A. B. McBratney, B. P. Malone, and I. Wheeler (2013). Digital Mapping of Soil Carbon. In: *Advances in Agronomy* 118, pp. 1–47 (Cit. on p. 9).

Mitchell, T. (2005). *Web Mapping Illustrated: using open source GIS toolkits*. O'Reilly (Cit. on p. 83).

Moeys, J. (2014). *soiltexture: Functions for soil texture plot, classification and transformation*. R package version 1.2.13. URL: http://CRAN.R-project.org/package=soiltexture (Cit. on p. 56).

Moore, I. D., P. E. Gessler, G. A. Nielsen, and G. A. Peterson (1993). Soil attribute prediction using terrain analysis. In: *Soil Science Society of America Journal* 57, pp. 443–452 (Cit. on pp. 9, 32, 40, 41, 144).

Moran, P. A. P. (1950). Notes on Continuous Stochastic Phenomena. In: *Biometrika* 37.1/2, pp. 17–23 (Cit. on p. 61).

Mulder, V. L., S. De Bruin, M. E. Schaepman, and T. R. Mayr (2011). The use of remote sensing in soil and terrain mapping – A review. In: *Geoderma* 162.1, pp. 1–19 (Cit. on p. 12).

Myers, D. E. (1982). Matrix Formulation of Co-Kriging. In: *Mathematical Geology* 14.3, pp. 249–257 (Cit. on p. 75).

Nash, J. E. and J. V. Sutcliffe (1970). River flow forecasting through conceptual models, Part I – A discussion of principles. In: *Journal of Hydrology* 10.3, pp. 282–290 (Cit. on pp. xiv, 140).

Navarra, A. and L. Tubiana (2013). *Regional Assessment of Climate Change in the Mediterranean. Volume 2: Agriculture, Forests and Ecosystem Services and People*. Vol. 51. Advances in Global Change Research. Springer (Cit. on p. 1).

Neter, J., M. H. Kutner, C. J. Nachtsheim, and W. Wasserman (1996). *Applied Linear Statistical Models*. McGraw–Hill (Cit. on p. 60).

Nolde, M., R. Duttmann, M. Blaschek, and U. Klein (2010). Geodateninfrastrukturen und ihre Anwendungen in der Praxis. In: *PIK-Praxis der Informationsverarbeitung und Kommunikation* 33.4, pp. 245–252 (Cit. on p. 83).

Odeh, I. O. A., A. B. McBratney, and D. J. Chittleborough (1995). Further results on prediction of soil properties from terrain attributes: heterotopic cokriging and regression-kriging. In: *Geoderma* 67.3–4, pp. 215–226 (Cit. on pp. 11, 76).

Odeh, I. O. A., A. J. Todd, and J. Triantafilis (2003). Spatial prediction of soil particle-size fractions as compositional data. In: *Soil Science* 168.7, pp. 501–515 (Cit. on pp. 9, 151).

Olaya, V. (2004). *A gentle introduction to SAGA GIS*. 1.1 (Cit. on p. 40).

Olden, J. D. and D. A. Jackson (2002). Illuminating the black box: a randomization approach for understanding variable contributions in artificial neural networks. In: *Ecological modelling* 154.1, pp. 135–150 (Cit. on p. 67).

Oppo, I. (2010). Caratterizzazione Geologica ed Idraulica del Bacino del Rio Mannu. Master thesis. University of Cagliari (Cit. on p. 21).

Özesmi, S. L. and U. Özesmi (1999). An artificial neural network approach to spatial habitat modelling with interspecific interaction. In: *Ecological modelling* 116.1, pp. 15–31 (Cit. on p. 67).

Ozturk, M., O. Salman, and M. Koc (2011). Artificial neural network model for estimating the soil temperature. In: *Canadian Journal of Soil Science* 91.4, pp. 551–562 (Cit. on p. 10).

Padarian, J., J. Pérez-Quezada, and O. Seguel (2012). Modelling the distribution of organic carbon in the soils of Chile. In: *Digital Soil Assessments and Beyond*. Ed. by B. Minasny, B. P. Malone, and A. B. McBratney. Proceedings of the 5th Global Workshop on Digital Soil Mapping 2012, Sydney, Australia. CRC Press, pp. 329–333 (Cit. on pp. 154, 155).

Pagès, J. (2014). *Multiple Factor Analysis by Example Using R*. The R Series. CRC Press (Cit. on p. 49).

Papritz, A, H. R Künsch, and R Webster (1993). On the Pseudo Cross-Variogram. In: *Mathematical Geology* 25.8, pp. 1015–1026 (Cit. on p. 74).

Pawlowsky, V., R. A. Olea, and J. C. Davis (1995). Estimation of Regionalized Compositions: A Comparison of Three Methods. In: *Mathematical Geology* 27.1, pp. 105–127 (Cit. on pp. 9, 54, 55, 151).

Pawlowsky-Glahn, V. and J. J. Egozcue (2006). Compositional data and their analysis: an introduction. In: *Compositional Data Analysis in the Geosciences: From Theory to Practice*. Ed. by A. Buccianti, G. Mateu-Figueras, and V. Pawlowsky-Glahn. Vol. 264. 1. Geological Society of London, pp. 1–10 (Cit. on pp. 54, 152).

Pawlowsky-Glahn, V. and R. A. Olea (2004). *Geostatistical Analysis of Compositional Data*. Oxford University Press, New York (Cit. on pp. 55, 56).

Pebesma, E. J. (2004). Multivariate geostatistics in S: the gstat package. In: *Computers & Geosciences* 30, pp. 683–691 (Cit. on p. 59).

Pebesma, E. J. and R. S. Bivand (2005). Classes and methods for spatial data in R. In: *R News* 5.2, pp. 9–13 (Cit. on p. 54).

Pelletier, B., P. Dutilleul, G. Larocque, and J. W. Fyles (2004). Fitting the Linear Model of Coregionalization by Generalized Least Squares. In: *Mathematical Geology* 36.3, pp. 323–343 (Cit. on p. 75).

Petersen, H., T. Wunderlich, S. Attia al Hagrey, and W. Rabbel (2012). Characterization of some Middle European soil textures by gamma-spectrometry. In: *Journal of Plant Nutrition and Soil Science* 175.5, pp. 651–660. ISSN: 1522-2624 (Cit. on p. 157).

Piccini, C., A. Marchetti, and R. Francaviglia (2014). Estimation of soil organic matter by geostatistical methods: Use of auxiliary information in agricultural and environmental assessment. In: *Ecological Indicators* 36, pp. 301–314 (Cit. on pp. 11, 147, 148).

Piikki, K., M. Söderström, and B. Stenberg (2013). Sensor data fusion for topsoil clay mapping. In: *Geoderma* 199, pp. 106–116 (Cit. on pp. 12, 157).

Pike, R. J., I. S. Evans, and T. Hengl (2009). Geomorphometry: A Brief Guide. In: *Geomorphometry: Concepts, Software, Applications*. Ed. by T. Hengl and H. I. Reuter. Vol. 33. Developments in Soil Science. Elsevier, Amsterdam, pp. 3–30 (Cit. on pp. 38, 39).

Plant, R. E. (2012). *Spatial data analysis in ecology and agriculture using R*. CRC Press (Cit. on p. 47).

Poggio, L. and A. Gimona (2014). National scale 3D modelling of soil organic carbon stocks with uncertainty propagation – An example from Scotland. In: *Geoderma* 232, pp. 284–299 (Cit. on p. 164).

Poggio, L., A. Gimona, and M. J. Brewer (2013). Regional scale mapping of soil properties and their uncertainty with a large number of satellite-derived covariates. In: *Geoderma* 209, pp. 1–14 (Cit. on p. 10).

Priori, S., N. Bianconi, and E. A. C. Costantini (2014). Can γ-radiometrics predict soil textural data and stoniness in different parent materials? A comparison of two machine-learning methods. In: *Geoderma* 226, pp. 354–364 (Cit. on pp. 10, 12, 149, 156).

Rao, C. R. (2002). *Linear Statistical Inference and its Applications*. 2nd ed. Wiley (Cit. on p. 61).

Riedmiller, M. (1994). Advanced Supervised Learning in Multi-Layer Perceptrons – From Backpropagation to Adaptive Learning Algorithms. In: *Computer Standards & Interfaces* 16.3, pp. 265–278 (Cit. on pp. 64, 65).

Riedmiller, M. and H. Braun (1993). A Direct Adaptive Method for Faster Backpropagation Learning: The RPROP Algorithm. In: *Proceedings of the IEEE International Conference on Neural Networks (ICNN)*. Ed. by H. Ruspini, pp. 586–591 (Cit. on pp. 64, 65).

Robinson, G. K. (1991). That BLUP is a Good Thing: The Estimation of Random Effects. In: *Statistical Science* 6.1, pp. 15–32 (Cit. on p. 70).

Rodrigues, F. A., R. G. V. Bramley, and D. L. Gobbett (2015). Proximal soil sensing for Precision Agriculture: Simultaneous use of electromagnetic induction and gamma radiometrics in contrasting soils. In: *Geoderma* 243, pp. 183–195 (Cit. on pp. 12, 149).

Rojas, R. (1996). *Neural networks: A Systematic Introduction*. Springer (Cit. on pp. 64, 67, 128).

Royle, J. and D. Nychka (1998). An algorithm for the construction of spatial coverage designs with implementation in SPLUS. In: *Computers & Geosciences* 24.5, pp. 479–488 (Cit. on p. 31).

Rumelhart, D. E., G. E. Hinton, and R. J. Williams (1986). Learning Internal Representations by Error Propagation. In: *Parallel Distributed Processing: Explorations in the Micro-Structure of Cognition*. Ed. by D. E. Rumelhart and J. L. McClelland. The MIT Press. Chap. 8, pp. 318–362 (Cit. on pp. 64, 65).

Sachs, L. and J. Hedderich (2006). *Angewandte Statistik: Methodensammlung mit R*. Springer (Cit. on pp. 48, 61).

SAGA User Group Association (2011). *System for Automated Geoscientific Analysis (version 2.0.8). (software)*. URL: http://www.saga-gis.org/ (visited on 05/27/2014) (Cit. on p. 38).

Saptoro, A., M. O. Tadé, and H. Vuthaluru (2012). A Modified Kennard-Stone Algorithm for Optimal Division of Data for Developing Artificial Neural Network Models. In: *Chemical Product and Process Modeling* 7.1 (Cit. on pp. 57, 155).

Sarle, W. S. (1994). Neural Networks and Statistical Models. In: *Proceedings of the Nineteenth Annual SAS Users Group International Conference* (Cit. on pp. 61, 68).

Sarle, W. S. (2002). *The IEEE Transactions on Neural Networks (Neural Network FAQ)*. URL: ftp://ftp.sas.com/pub/neural/FAQ.html (Cit. on p. 67).

Schabenberger, O. and C. A. Gotway (2004). *Statistical Methods for Spatial Data Analysis*. CRC Press (Cit. on p. 77).

Schwertmann, U. (1964). Differenzierung der Eisenoxide des Bodens durch Extraktion mit Ammoniumoxalat-Lösung. In: *Zeitschrift für Pflanzenernährung, Düngung, Bodenkunde* 105.3, pp. 194–202 (Cit. on p. 37).

Scott, D. W. (1979). On optimal and data-based histograms. In: *Biometrika* 66.3, pp. 605–610 (Cit. on p. 47).

Scull, P., J. Franklin, O. A. Chadwick, and D. McArthur (2003). Predictive soil mapping: a review. In: *Progress in Physical Geography* 27.2, pp. 171–197 (Cit. on p. 8).

Scull, P., J. Franklin, and O. A. Chadwick (2005). The application of classification tree analysis to soil type prediction in a desert landscape. In: *Ecological Modelling* 181.1, pp. 1–15 (Cit. on p. 10).

Sevink, J. and E. A. Kummer (1984). Eolian dust deposition on the Giara di Gesturi basalt plateau, Sardinia. In: *Earth Surface Processes and Landforms* 9.4, pp. 357–364 (Cit. on p. 151).

Shapiro, S. S. and M. B. Wilk (1965). An analysis of variance test for normality (complete samples). In: *Biometrika* 52.3/4, pp. 591–611 (Cit. on pp. 47, 61).

Shepard, D. (1968). A two-dimensional interpolation function for irregularly-spaced data. In: *Proceedings of the 1968 23rd ACM national conference*. ACM, pp. 517–524 (Cit. on p. 68).

Singh, D. and A. Kathpalia (2007). An efficient modeling with GA approach to retrieve soil texture, moisture and roughness from ERS-2 SAR data. In: *Progress In Electromagnetics Research* 77, pp. 121–136 (Cit. on p. 162).

Snee, R. D. (1977). Validation of Regression Models: Methods and Examples. In: *Technometrics* 19.4, pp. 415–428 (Cit. on p. 82).

Soil Survey Staff (1999). *Soil taxonomy: A basic system of soil classification for making and interpreting soil surveys*. 2nd edition. Natural Resources Conser-

vation Service. U.S. Department of Agriculture Handbook 436 (Cit. on pp. 19, 92, 120).

Soil Survey Staff (2010). *Keys to Soil Taxonomy*. 11th ed. USDA-Natural Resources Conservation Service, Washington, DC (Cit. on p. 19).

Stein, M. L. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York (Cit. on pp. 8, 69, 71, 77).

Sterlacchini, S., C. Ballabio, J. Blahut, M. Masetti, and A. Sorichetta (2011). Spatial agreement of predicted patterns in landslide susceptibility maps. In: *Geomorphology* 125.1, pp. 51–61 (Cit. on pp. 82, 139).

Sudduth, K. A., N. R. Kitchen, W. J. Wiebold, W. D. Batchelor, G. A. Bollero, D. G. Bullock, D. E. Clay, H. L. Palm, F. J. Pierce, R. T. Schuler, and K. D. Thelen (2005). Relating apparent electrical conductivity to soil properties across the north-central USA. In: *Computers and Electronics in Agriculture* 46.1, pp. 263–283 (Cit. on p. 43).

Sulaeman, Y., B. Minasny, A. B. McBratney, M. Sarwani, and A. Sutandi (2013). Harmonizing legacy soil data for digital soil mapping in Indonesia. In: *Geoderma* 192, pp. 77–85 (Cit. on p. 163).

Sumfleth, K. and R. Duttmann (2008). Prediction of soil property distribution in paddy soil landscapes using terrain data and satellite information as indicators. In: *Ecological Indicators* 8, pp. 485–501 (Cit. on pp. 11, 145, 162).

Taalab, K., R. Corstanje, J. Zawadzka, T. Mayr, M. J. Whelan, J. A. Hannam, and R. Creamer (2015). On the application of Bayesian Networks in Digital Soil Mapping. In: *Geoderma* 259, pp. 134–148 (Cit. on pp. 10, 164).

Tamm, O. (1932). Über die Oxalatmethode in der chemischen Bodenanalyse. In: *Medföljer Skogsvardsföreningens Tidskrift* 1–2 (Cit. on p. 37).

Taylor, J. A., M. Short, A. B. McBratney, and J. Wilson (2010). Comparing the Ability of Multiple Soil Sensors to Predict Soil Properties in a Scottish Potato Production System. In: *Proximal Soil Sensing*. Ed. by R. A. Viscarra Rossel, A. B. McBratney, and B. Minasny. Springer, pp. 387–396 (Cit. on p. 157).

Taylor, M. J., K. Smettem, G. Pracilio, and W. Verboom (2002). Relationships between soil properties and high-resolution radiometrics, central eastern Wheatbelt, Western Australia. In: *Exploration Geophysics* 33.2, pp. 95–102 (Cit. on pp. 45, 145, 156).

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. In: *Economic Geography* 46, pp. 234–240 (Cit. on p. 68).

Triantafilis, J. and S. M. Lesch (2005). Mapping clay content variation using electromagnetic induction techniques. In: *Comput. Electron. Agr.* 46.1, pp. 203–237. ISSN: 0168-1699 (Cit. on pp. 12, 146).

Tukey, J. W. (1977). *Exploratory data analysis*. Reading, Mass. (Cit. on pp. 47, 48).

Tutiempo Network, S.L. (Apr. 2014). *Climate Cagliari / Elmas: Data reported by the weather station 165600 (LIEE)*. URL: http://www.tutiempo.net/en/Climate/Cagliari_Elmas/165600.htm (visited on 04/08/2014) (Cit. on p. 15).

Vacca, A, S Loddo, G Serra, and A Aru (2002). Soil degradation in Sardinia (Italy): main factors and processes. In: *7th International meeting on Soils with Mediterranean Type of Climate (selected papers). Bari: CIHEAM*, pp. 413–423 (Cit. on pp. 14, 15).

Vacca, A., ed. (2012). *Field trip guide - Land degradation in Mediterranean environments: Causes, processes and management*. COMLAND meeting and field trip in Sardinia (Cit. on pp. 1, 17, 19).

Vacca, A., S. Loddo, M. T. Melis, A. Funedda, R. Puddu, M. Verona, S. Fanni, F. Fantola, S. Madrau, V. A. Marrone, G. Serra, C. Tore, D. Manca, S. Pasci, M. R. Puddu, and P. Schirru (2014). A GIS based method for soil mapping in Sardinia, Italy: A geomatic approach. In: *Journal of Environmental Management* 138, pp. 87–96 (Cit. on pp. 3, 38, 146, 163).

Van den Boogaart, K. G. and R. Tolosana-Delgado (2013). *Analyzing Compositional Data with R*. Springer (Cit. on p. 56).

Van den Boogaart, K. G., R. Tolosana, and M. Bren (2014). *compositions: Compositional Data Analysis*. R package version 1.40-1. URL: http://CRAN.R-project.org/package=compositions (Cit. on p. 56).

Van der Klooster, E., F. M. van Egmond, and M. P. W. Sonneveld (2011). Mapping soil clay contents in Dutch marine districts using gamma-ray spectrometry. In: *European Journal of Soil Science* 62.5, pp. 743–753. ISSN: 1365-2389 (Cit. on p. 146).

Van Groenigen, J., W. Siderius, and A. Stein (1999). Constrained optimisation of soil sampling for minimisation of the kriging variance. In: *Geoderma* 87.3-4, pp. 239–259 (Cit. on p. 31).

Vašát, R., G. Heuvelink, and L. Boruvka (2010). Sampling design optimization for multivariate soil mapping. In: *Geoderma* 155, pp. 147–153 (Cit. on p. 31).

Vašát, R., L. Pavlu, L. Boruvka, O. Drábek, and A. Nikodem (2013). Mapping the Topsoil pH and Humus Quality of Forest Soils in the North Bohemian Jizerské hory Mts. Region with Ordinary, Universal, and Regression Kriging: Cross-Validation Comparison. In: *Soil and Water Research* 8.3, pp. 97–104 (Cit. on p. 82).

Vaysse, K. and P. Lagacherie (2015). Evaluating Digital Soil Mapping approaches for mapping GlobalSoilMap soil properties from legacy data in Languedoc-Roussillon (France). In: *Geoderma Regional* 4, pp. 20–30 (Cit. on p. 2).

Ver Hoef, J. M. and N. Cressie (1993). Multivariable Spatial Prediction. In: *Mathematical Geology* 25.2, pp. 219–240 (Cit. on pp. 73, 75).

Verheye, W. and D. de la Rosa (2006). Mediterranean soils. In: *Land Use, Land Cover and Soil Sciences*. Ed. by W. H. Verheye. UNESCO-EOLSS Publishers. URL: http://www.eolss.net/ (Cit. on pp. 20, 28).

Walvoort, D. J. J. and J. J. de Gruijter (2001). Compositional Kriging: A Spatial Interpolation Method for Compositional Data. In: *Mathematical Geology* 33.8, pp. 951–966 (Cit. on p. 9).

Ward, C. and U. Mueller (2012). Multivariate Estimation Using Log Ratios: A Worked Alternative. In: *Geostatistics Oslo 2012*. Ed. by P. Abrahamsen, O. Kolbjørnsen, and R. Hauge. Springer, pp. 333–343 (Cit. on pp. 81, 152).

Warner, B. and M. Misra (1996). Understanding Neural Networks as Statistical Tools. In: *The American Statistician* 50.4, pp. 284–293 (Cit. on p. 62).

Webster, R. (2000). Is soil variation random? In: *Geoderma* 97.3, pp. 149–163 (Cit. on p. 8).

Webster, R. (1994). The development of pedometrics. In: *Geoderma* 62.1, pp. 1–15 (Cit. on p. 8).

Webster, R. and M. A. Oliver (1992). Sample adequately to estimate variograms of soil properties. In: *Journal of Soil Science* 43.1, pp. 177–192 (Cit. on pp. 32, 53).

Webster, R. and M. A. Oliver (2007). *Geostatistics for Environmental Scientists*. Wiley (Cit. on pp. 9, 50, 52, 53, 68, 69, 71, 72, 75, 152).

Weisberg, S. (2014). *Applied Linear Regression*. 4th ed. Wiley (Cit. on p. 61).

Werban, U., T. Behrens, G. Cassiani, and P. Dietrich (2010). iSOIL: An EU Project to Integrate Geophysics, Digital Soil Mapping, and Soil Science. In: *Proximal Soil Sensing*. Ed. by R. A. Viscarra Rossel, A. B. McBratney, and B. Minasny. Springer. Chap. 8, pp. 103–110 (Cit. on pp. 3, 13, 42).

Wetterlind, J., K. Piikki, B. Stenberg, and M. Söderström (2015). Exploring the predictability of soil texture and organic matter content with a commercial integrated soil profiling tool. In: *European Journal of Soil Science* 66.4, pp. 631–638 (Cit. on p. 150).

Wiesmeier, M., F. Barthold, B. Blank, and I. Kögel-Knabner (2011). Digital mapping of soil organic matter stocks using Random Forest modeling in a semi-arid steppe ecosystem. In: *Plant Soil* 340.1-2, pp. 7–24 (Cit. on pp. 10, 32).

Wilson, J. P. and J. C. Gallant (2000). Digital Terrain Analysis. In: *Terrain analysis: Principles and applications*. Ed. by J. P. Wilson and J. C. Gallant. Wiley, New York. Chap. 1, pp. 1–27 (Cit. on pp. 37, 38).

Wilson, P. L., D. Jacquier, and B. A. Simons (2012). Digital Soil Map data in an on-line, on-demand world. In: *Digital Soil Assessments and Beyond*. Ed. by B. Minasny, B. P. Malone, and A. B. McBratney. Proceedings of the 5th Global Workshop on Digital Soil Mapping 2012, Sydney, Australia. CRC Press, pp. 293–297 (Cit. on p. 159).

Witten, I. H. and E. Frank (2005). *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann (Cit. on p. 9).

Yaalon, D. H. (1997). Soils in the Mediterranean region: what makes them different? In: *CATENA* 28.3–4, pp. 157–169 (Cit. on pp. 20, 28, 150).

Zell, A., G. Mamier, M. Vogt, N. Mache, R. Hübner, S. Döring, K.-U. Herrman, T. Soyez, M. Schmalzl, T. Sommer, A. Hatzigeorgiou, D. Posselt, T. Schreiner, B. Kett, G. Clemente, J. Wieland, and J. Gatter (1998). *SNNS: Stuttgart Neural Network Simulator*. User Manual, version 4.2. IPVR, University of Stuttgart and WSI, University of Tübingen. URL: http://www.ra.cs.uni-tuebingen.de/SNNS/ (Cit. on pp. 66, 67).

Zevenbergen, L. W. and C. R. Thorne (1987). Quantitative Analysis of Land Surface Topography. In: *Earth Surface Processes and Landforms* 12, pp. 47–56 (Cit. on pp. 39, 40).

Zhang, S., Y. Huang, C. Shen, H. Ye, and Y. Du (2012). Spatial prediction of soil organic matter using terrain indices and categorical variables as auxiliary information. In: *Geoderma* 171-172, pp. 35–43 (Cit. on p. 11).

Zhao, Z., T. L. Chow, H. W. Rees, Q. Yang, Z. Xing, and F.-R. Meng (2009). Predict soil texture distributions using an artificial neural network model. In: *Computers and Electronics in Agriculture* 65, pp. 36–48 (Cit. on pp. 10, 64, 148).

Acknowledgements

First and foremost, I would like to express my deep appreciation to my supervisor Prof. Dr. Rainer Duttmann for his scientific advice and helpful guidance during the past six years. I also wish to thank Prof. Dr. Ralf Ludwig from the University of Munich (LMU) for being the co-reviewer of this thesis and for constantly supporting me in his role as coordinator of the EU-FP7 project CLIMB.

I sincerely thank Dr. Ulrike Werban and her team from the MET Department of the UFZ Helmholtz Centre for Environmental Research for providing the geophysical sensor datasets used in this work. In addition, I owe special thanks to Dr. Antonino Soddu Pirellas (AGRIS Sardegna) and everyone at the San Michele research farm for their friendly help and support during field campaigns.

This thesis would not have been possible without the ongoing support of our laboratory assistants, technical employees and students. In particular, I am indebted to Julia Becker, Antje Berger, Ursula Bock, Daniel Gerken, Robert Minkler and Nicole Wilder for assistance with soil sampling and lab analyses. Furthermore, a cordial thank-you goes to Dr. Michael Nolde for numerous practical tips regarding web-mapping, and Kilian Etter for proofreading this thesis.

Among my department colleagues, Michael Kuhwald deserves special mention for his tremendous efforts in the laboratory, passionate support during three intensive field campaigns, and many illuminating conversations.

I also wish to thank Swen Meyer and his student assistants from the LMU Munich for an excellent cooperation. It was my pleasure to jointly carry out much of the fieldwork and to spend such an extremely productive and great time on Sardinia.

Moreover, the research leading to this dissertation has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no 244151.

Finally, I would like to show my gratitude to my family and friends for their motivational support. In particular, I sincerely thank my partner Laura for her unceasing encouragement and patience throughout this project.

Declaration of Authorship / Eidesstattliche Erklärung

Hiermit erkläre ich, dass die vorliegende Dissertationsschrift – abgesehen von der Beratung durch den Betreuer – nach Inhalt und Form meine eigene Arbeit ist. Weiterhin versichere ich, dass diese Abhandlung weder ganz noch in Teilen schon an anderer Stelle Gegenstand eines Prüfungsverfahrens war, veröffentlicht worden ist oder zur Veröffentlichung eingereicht wurde und ich sie unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft verfasst habe.

Kiel, 26. Oktober 2015

Michael Blaschek

europass Curriculum Vitae

PERSONAL INFORMATION

Michael Blaschek

Kämpenstr. 6, D-24106 Kiel

+49 431 53671974

blaschek@geographie.uni-kiel.de

Sex Male | Date of birth 31/12/1982 in Hanover | Nationality German

SCIENTIFIC CAREER

1st December 2009 – Present

Research Associate

Department of Geography, Kiel University (Germany)

- EU-FP7-project: Climate Induced Changes on the Hydrology of Mediterranean Basins (CLIMB)
  - Main responsibility for the geospatial data management of the project
  - Set-up of a geoportal for the dissemination of project results: lgi-climbsrv.geographie.uni-kiel.de
  - Creation of OGC-compliant web services and ISO-compliant metadata records
  - Coordination of work package activities including reporting and delegation
- Doctoral thesis: *Digital soil mapping at different spatial scales using machine learning algorithms and multivariate geostatistics in a Mediterranean basin*
  - Soil sampling, laboratory analysis, (geo)statistical modelling, web-based dissemination
- Teaching: Geostatistics, GIS, Field Studies – Sardinia, Physical Geography

EDUCATION AND TRAINING

October 2003 – November 2009

Diploma in Geography

grade 1.2
(best score 1.0 of 5.0)

Universities of Kiel (Germany) and Aberdeen (United Kingdom, 1 semester)

- Minor subjects: Meteorology and Soil Science
- Diploma thesis: *Regionalization of soil chemical properties at different spatial scales using regression kriging and artificial neural networks in paddy soil landscapes, China (in German)*
- Student assistant at the Chair of Physical Geography – Landscape Ecology and Geoinformation Science in Kiel (Germany) from 1st August 2005 to 30th September 2009
- Internship at the State Agency for Agriculture, Environment and Rural Areas Schleswig-Holstein in Flintbek (Germany) from 1st August to 26th September 2008
- Internship at the environmental planning office GEUM.tec GmbH in Hanover (Germany) from 19th February to 30th March 2007
- Scholarship to attend the Summer School of Polish Language and Culture in Poznań (Poland), August 2004

August 1996 – June 2002

Abitur (equivalent to A-levels)

grade 1.7
(best score 1.0 of 6.0)

St. Ursula-Schule, Hanover (Germany)

ADDITIONAL INFORMATION

Memberships
Awards

Memberships

- Member of the German Soil Science Society (DBG), FOSSGIS e.v.

Awards

- Outstanding Student Poster (OSP) Award 2015 at the EGU – GA, Vienna, April 2015: *A stratified two-stage sampling design for digital soil mapping in a Mediterranean basin*
- Best student poster (4th rank) at the 10th Conference on Geostatistics for Environmental Applications, Paris, July 2014: *A neural network residual cokriging approach to predict soil separates in a Mediterranean basin*

Appendix A

.1 Software and applications

Table 1 specifies the software components that were used for both, the creation of this document and the implementation of the web-based dissemination platform. The latter was briefly introduced in chapter 4.8 at page 83. Its technical details are thoroughly documented within the corresponding deliverable reports related to work package 2 of the EU-FP7 project CLIMB (Climate Induced Changes on the Hydrology of Mediterranean Basins). These reports are partly public and can be downloaded from the project's website at http://www.climb-fp7.eu. GeoNode is running on a virtual machine with 2.8 GHz processor, 6 GB of memory (RAM) and an Ubuntu 12.04 LTS 64-bit operating system.

Table 1: Main software and applications used in the frame of this thesis

Name	Title / Purpose	Version
ArcPad	Esri's mobile field mapping and data collection software	8
ArcMap	Main component of Esri's ArcGIS used here for map creation	10.1
CUED PhD	PhD thesis LaTeX template (Harish Bhanderi)	1.1
Dia	GTK+ based diagram creation program	0.97.2
GeoExplorer	WebGIS-client based on GeoExt	3.0.5
GeoNode	Open source geospatial content management system	2.0c5
GeoServer	Open source server for geospatial data sharing	2.4
GME	Geospatial Modelling Environment used for soil sampling tasks	0.5.3
PostgreSQL	Database back-end, enhanced by PostGIS 2.0.1	9.2.1
pycsw	CSW server implementation, written in Python	1.6.0
TeX Live	Comprehensive TeX system used for setting this thesis document	20130722-1
Texmaker	Cross-platform LaTeX editor used for thesis writing	4.0.3

.2 Functionality of the geoportal solution

This section illustrates the functionality of the CLIMB Geoportal with respect to digital soil mapping products in form of selected screenshots. The portal's design was customised in accordance with CLIMB-specific needs by editing the CSS and HTML files of the used GeoNode-based content management system.

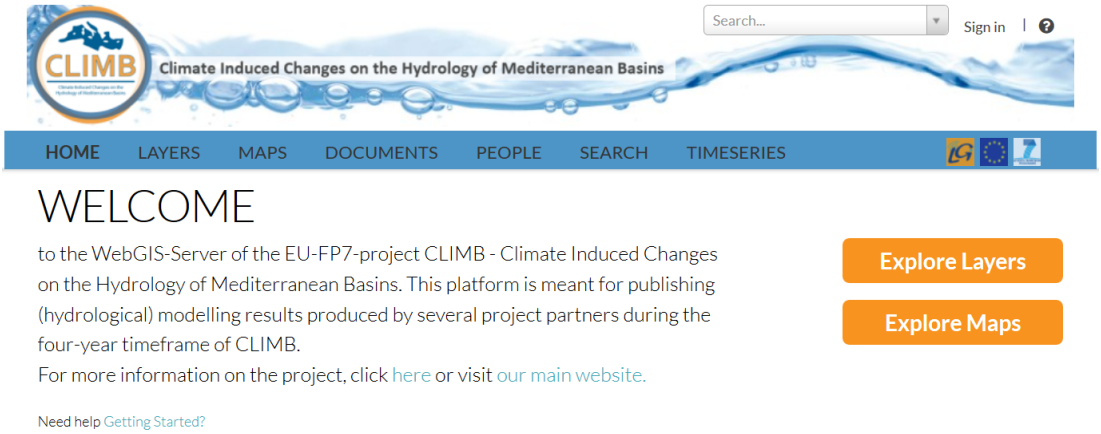


Figure 1: Home page of the dissemination web-platform (geoportal)
Source: http://lgi-climbsrv.geographie.uni-kiel.de

Inside the header section, the navigation tab represents the key element to switch from one topic covered by the CLIMB Geoportal to another highlighting the current subject by bold letters. Furthermore, it connects to the sign-in window in the right upper corner and provides direct access to the help page, meant to inform the user how to operate the portal in an efficient manner.

Following the Explore Layers button of the portal's home page, provides a list of all available spatial resources of the CLIMB Geoportal as shown in figure 2. Due to the fact that more than 3000 layers are available at the given web-platform, it is particularly useful to reduce the amount of resources listed at the Explore Layers page, for instance, by selecting the study site of interest. Refer to the Rio di Costara sub-category for digital soil mapping outputs of the Sardinian test site. The Upload Layers page is reserved for registered users only. Registration is granted upon request and requires the approval of the scientific project coordinator.

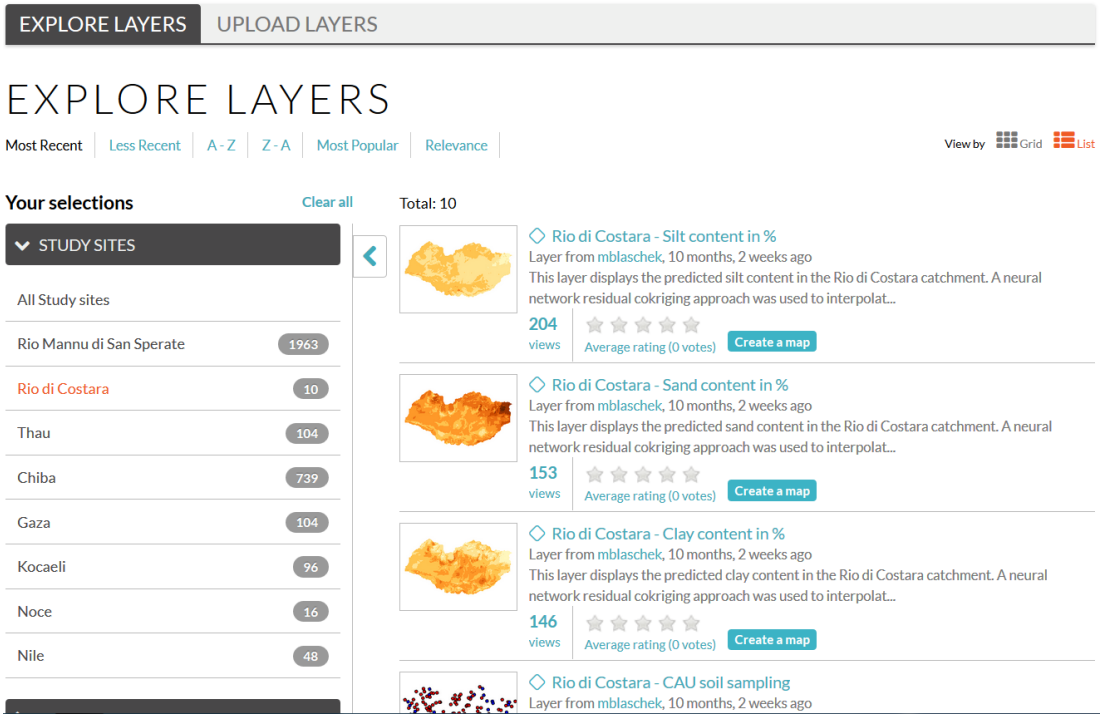


Figure 2: Layers section of the customised GeoNode interface

To get more information about a particular spatial resource, a click on its name in the layers list opens the layer details page showing a map preview and listing key metadata elements (see figures 3 and 4).

The upper part of the Layer details page shows a preview of the chosen resource in a map window that provides simple (GIS) functionality to interact with the given image. The button in the upper left hand corner, allows to change the underlying base map, for instance, from the default OpenStreetMap background to some satellite image. The following two buttons are meant to print the map or to move around inside the map area. To receive information on certain grid values, the Identify-button is used.

Below the map window, key metadata elements of the selected layer are presented including title, abstract, owner of the resource and supplemental information on the computation of the given parameter. The active layer can be utilised for map creation using GeoNode's built-in WebGIS-client (GeoExplorer).

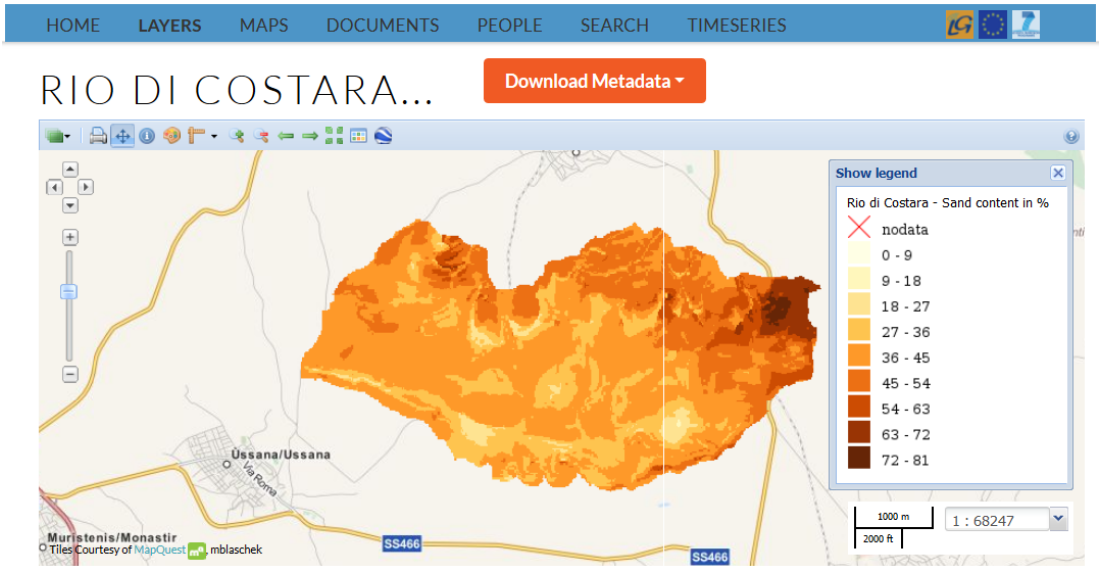


Figure 3: Upper part of the layer details page – Map window

The third element of the navigation tab connects to the Maps section of the given web-portal. This part of the CLIMB Geoportal is particularly meant for combining different layers from either the current platform itself or from remote web services (e.g. WMS). It provides a tool to disseminate certain information as pre-defined maps to soil experts, stakeholders, policy makers or other interested non-professionals. The functionality of the given GeoExplorer instance as presented in figure 5 almost equals the map preview of the Layers section.

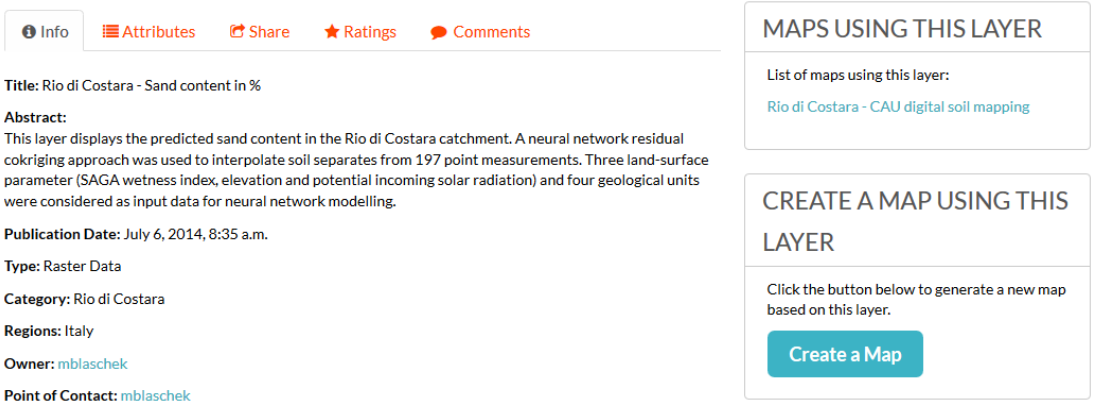


Figure 4: Lower part of the layer details page – Selected metadata

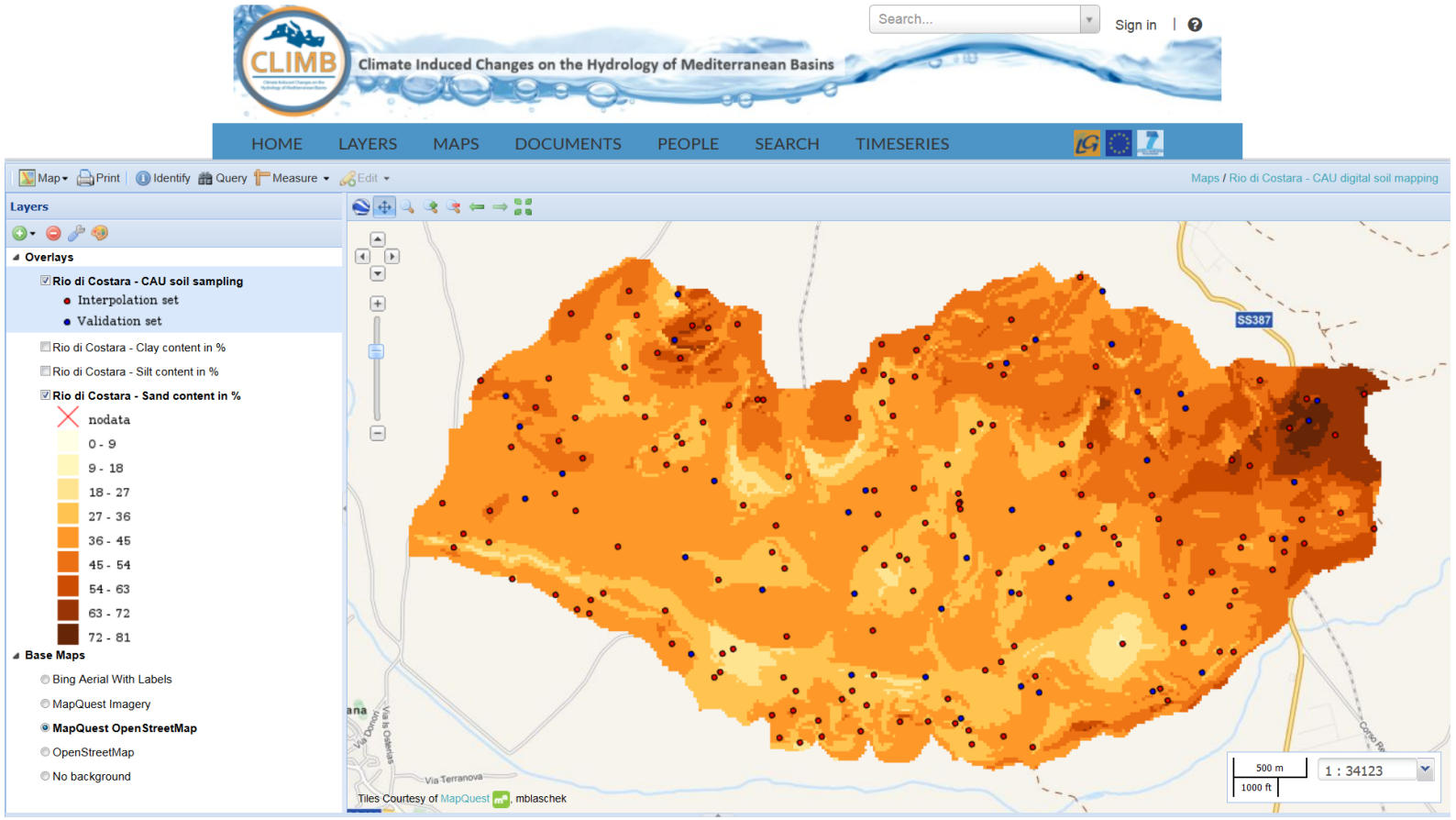


Figure 5: Built-in GIS-client based on the GeoExplorer web application

Appendix B

.3 R packages

This section lists essential R packages applied in the frame of this thesis. For details refer to the respective package-sites at http://cran.r-project.org. R itself was run in version 3.0.2 on a pc with an Ubuntu 13.10 64-bit operating system.

Table 2: Applied R packages

Name	Title / Purpose	Version
car	Companion to applied regression	2.0-19
caret	Training and plotting classification and regression models	6.0-22
climatol	Drawing wind rose and Walter and Lieth diagrams	2.2
compositions	Functions for the consistent analysis of compositional data	2013.6.15
FactoMineR	Multivariate exploratory data analysis and data mining	1.26
ggplot2	An implementation of the grammar of graphics	0.9.3.1
gstat	Geostatistical modelling, prediction and simulation	1.0-16
gridExtra	High-level functions for grid graphics	0.9.1
irr	Various coefficients of inter-rater reliability and agreement	0.84
lattice	High-level data visualization system	0.20-24
NeuralNetTools	Visualization and analysis tools for neural networks	1.0.1
plotrix	Various plots, labelling, axis and color scaling functions	3.5-3
plyr	Tools for splitting, applying and combining data	1.8
RColorBrewer	Provides color schemes for maps and other graphics	1.0-5
rgdal	Bindings for the Geospatial Data Abstraction Library	0.8-15
RSAGA	SAGA geoprocessing and terrain analysis in R	0.93-6
RSNNS	NNs in R using the Stuttgart Neural Network Simulator	0.4-6
soiltexture	Functions for soil texture plot, classification and transformation	1.2.13
sp	Classes and methods for spatial data	1.0-14
spdep	Spatial dependence: weighting schemes, statistics and models	0.5-74
xtable	Export tables to LaTeX or HTML	1.7-3
zoo	Infrastructure for regular and irregular time series	1.7-10

.4 R scripts

.4.1 cau_climatedata_calc_vDiss.R

```
#
################################################################################
#### Analysis of climate data from own weather station, Azienda San Michele ####
################################################################################
#
## last update: 08.04.2014

# climate tower from CAU and LMU, supported by AGRIS Sardegna
# installed in October 2010, repaired in March 2011, deconstructed in March 2013
# EAST: 508589, NORTH: 4362827, EPSG: 32632

# author: Michael Blaschek (blaschek@geographie.uni-kiel.de)
# contributions from: Jan Merkens (merkens@geographie.uni-kiel.de)

rm(list = ls())

# package "climatol" has been removed from CRAN --> installed from source..
install.packages("climatol_2.2.tar.gz", repos = NULL, type = "source")

library(zoo); library(ggplot2); library(climatol); library(sp)
# R v3.0.2, zoo_1.7-10, ggplot2_0.9.3.1, climatol_2.2, sp_1.0-14

load(paste(getwd(), "phd_calc_input", "metstation_CAU_LMU_processed_v1.RData"
  , sep = "/"))
path.fig <- paste(getwd(), "phd_calc_figures", sep = "/")

#-------------------------------------------------------------------------------
## Climate graph
#-------------------------------------------------------------------------------

d.cl$date <- as.Date(paste(d.cl$year, d.cl$month, d.cl$day, sep = "-"))

# longer time series available from Cagliari/Elmas weather station:
load(paste(getwd(), "phd_calc_input",
  "cagliari_elmas_metstation_1981_2010.RData", sep = "/"))

# precipitation:
p.agd <- aggregate(zoo(d.cl$PREC), d.cl$date, sum, na.rm = T)
p.agm <- aggregate(p.agd, as.yearmon(format(index(p.agd), "%Y-%m")), sum)
p.agmy <- aggregate(p.agm, as.numeric(format(index(p.agm), "%m")), mean)
prec2.agmy <- aggregate(zoo(d.cl2$PREC2),
  as.numeric(format(as.yearmon(date2), "%m")), mean, na.rm = T)

# temperature:
```

```
tmax.agd <- aggregate(zoo(d.cl$AIRTEMP200), d.cl$date, max)
tmax.agm <- aggregate(tmax.agd, as.yearmon(format(index(tmax.agd),"%Y-%m")),
  mean, na.rm = T)
tmax.agmy <- aggregate(tmax.agm, as.numeric(format(index(tmax.agm),"%m")), mean)
tmin.agd <- aggregate(zoo(d.cl$AIRTEMP200), d.cl$date, min)
tmin.agm <- aggregate(tmin.agd, as.yearmon(format(index(tmin.agd),"%Y-%m")),
  mean, na.rm = T)
tmin.agmy <- aggregate(tmin.agm, as.numeric(format(index(tmin.agm),"%m")), mean)
tmin.min <- aggregate(tmin.agd, as.yearmon(format(index(tmin.agd),"%Y-%m")),
  min, na.rm = T)
tmin.miny <- aggregate(tmin.min, as.numeric(format(index(tmin.min),"%m")), min)

tmax2.agmy <- aggregate(zoo(d.cl2$TMAX2),
  as.numeric(format(as.yearmon(date2), "%m")), mean, na.rm = T)
tmin2.agmy <- aggregate(zoo(d.cl2$TMIN2),
  as.numeric(format(as.yearmon(date2), "%m")), mean, na.rm = T)
tmin2.miny <- aggregate(zoo(d.cl2$TMINABS2),
  as.numeric(format(as.yearmon(date2), "%m")), min, na.rm = T)

# prepare plot, using diagwl function from climatol-package:
climate <- cbind(Prec = as.vector(p.agmy), Tmax = as.vector(tmax.agmy),
  Tmin = as.vector(tmin.agmy), Min = as.vector(tmin.miny))
climate <- t(climate)
colnames(climate) <- c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug",
  "Sep", "Okt", "Nov", "Dec")

climate2 <- cbind(Prec = as.vector(prec2.agmy), Tmax = as.vector(tmax2.agmy),
  Tmin = as.vector(tmin2.agmy), Min = as.vector(tmin2.miny))
climate2 <- t(climate2); climate2
colnames(climate2) <- c("Jan", "Feb", "Mar", "Apr", "May", "Jun", "Jul", "Aug",
  "Sep", "Okt", "Nov", "Dec")

pdf(paste(path.fig, "wldiagr_v2.pdf", sep = "/"))
diagwl(climate, alt = 150, per = "Oct 2010 to Jan 2013", mlab = "en",
  est = "CAU/LMU weather station, Azienda San Michele, Sardinia, Italy")
dev.off()

pdf(paste(path.fig, "wldiagr_165600_v1.pdf", sep = "/"))
diagwl(climate2, alt = 4, per = "Jan 1981 to Dec 2010", mlab = "en",
  est = "Cagliari/Elmas weather station, Lat 39.25, Lon 9.05, Sardinia, Italy")
dev.off()

rm(climate, p.agd, p.agm, p.agmy, tmax.agd, tmax.agm,
  tmax.agmy, tmin.agd, tmin.agm, tmin.agmy, tmin.min, tmin.miny,
  climate2, prec2.agmy, tmax2.agmy, tmin2.agmy, tmin2.miny)


#-------------------------------------------------------------------------
## windrose
#-------------------------------------------------------------------------
```

```
95
96 # 500cm
97 testc <- aggregate(d.cl$WIND500, list(dir = d.cl$WINDDIR2,
98   frq = ceiling(d.cl$WIND500)), FUN = function(x) sum(!is.na(x)))
99 testtab <- xtabs(x ~ frq + dir, testc)
100 testmatrix <- as.data.frame.matrix(testtab)
101
102 # calc. (and remove) % of circulating winds:
103 sum_rr0 <- sum(testmatrix[1,]); sum_rr <- sum(testmatrix)
104 per_rr0 <- round((sum_rr0/sum_rr)*100, 1); testmatrix <- testmatrix[-1,]
105
106 # sort columns (otherwise wrong illustr.) --> clockwise from N to NNW:
107 testmatrix <- testmatrix[c("N", "NNE", "NE", "ENE", "E", "ESE", "SE", "SSE",
108   "S", "SSW", "SW", "WSW", "W", "WNW", "NW", "NNW")]
109
110 testmatrix1 <- testmatrix[1:2,]
111 testmatrix1[1,] <- colSums(testmatrix[1:2,])
112 testmatrix1[2,] <- colSums(testmatrix[3:4,])
113 testmatrix1[3,] <- colSums(testmatrix[5:6,])
114 testmatrix1[4,] <- colSums(testmatrix[7:8,])
115 testmatrix1[5,] <- colSums(testmatrix[9:10,])
116 testmatrix1[6,] <- colSums(testmatrix[11:12,])
117 testmatrix1[7,] <- colSums(testmatrix[13:14,])
118 testmatrix1[8,] <- colSums(testmatrix[15:16,])
119 testmatrix1[9,] <- colSums(testmatrix[17:29,])
120 rownames(testmatrix1) <- c("0.1 to 2", "2.1 to 4", "4.1 to 6", "6.1 to 8",
121   "8.1 to 10", "10.1 to 12", "12.1 to 14", "14.1 to 16", "> 16.1")
122
123 # bpy.colors is supposed to be black-white printer friendly; requires sp-package
124 pdf(paste(path.fig, "windrose_v3.pdf", sep = "/"))
125 rosavent(testmatrix1, fnum = 4, fint = 5, uni = "m/s", key = T,
126   margen = c(0, 0, 4, 0), col = rev(bpy.colors(n = 9)),
127   main = paste("Annual wind rose in a height of 5m above ground\n
128     circulating winds: ", per_rr0, "%", sep = ""))
129 dev.off()
130
131 # end of script: cau_climatedata_calc_vDiss.R, 13.03.2014
```

.4.2 azienda_dta2k5_vDiss.R

```
1 #
2 ##############################################################################
3 ####  Extraction of LSP using SAGA at field 21 and 33, Azienda San Michele  ####
4 ##############################################################################
5 #
6 ## last update: 23.09.2014
7
8 # DTA based on "phd_calc_input/50000_548wgs84_azienda2k5tinlin.tif"
9 # Derived from http://www.sardegnageoportale.it/ --> Catalogo Dati -->
```

```
# Download --> Raccolte cartografiche --> Modello Digitale del Terreno SAR,
# passo 10m --> Scarica il DTM --> ascii_50000_548.asc (last access: 11.01.2011)

# Preprocessing (partly done in ArcGIS 10):
# 1. assign EPSG:3003 (info from website) and convert into GeoTIFF format
# 2. transform into UTM-WGS84 (EPSG:32632), +towgs84=-225,-65,9
# 3. clip to fit the Azienda San Michele
# 4. create TIN, then convert into raster, method = LINEAR, cellsize = 2.5m
# --> 50000_548wgs84p_azienda2k5tinlin.tif --> dem1
# 5. clip to fit the two fields only
# --> 50000_548wgs84_fields2k5tinlin.tif --> demt

# author: Michael Blaschek (blaschek@geographie.uni-kiel.de)
# reference: Tomislav Hengl - Analysis of DEMs in R+ILWIS/SAGA
#   (from http://spatial-analyst.net/wiki/)

rm(list = ls())

library(rgdal); library(RSAGA); library(compositions)
# R v3.0.2, SAGA v2.0.8, rgdal_0.8-15, RSAGA v0.93-6, compositions_1.40-1

myenv <- rsaga.env(path = "/usr/bin") # --> define, where SAGA is located..
#myenv <- rsaga.env(path = "C:/Program Files (x86)/SAGA/saga_2.0.8_bin_msw_x64")

dem1 <- readGDAL(paste(
  getwd(), "phd_calc_input", "50000_548wgs84_azienda2k5tinlin.tif", sep = "/"))
names(dem1)[1] <- "ELEV"

# fields 21 and 33 only, target grid:
demt <- readGDAL(paste(
  getwd(), "phd_calc_input", "50000_548wgs84_fields2k5tinlin.tif", sep = "/"))
names(demt)[1] <- "ELEV"


#-----------------------------------------------------------------------------
## Extraction of land-surface parameter (LSP)
#-----------------------------------------------------------------------------

# no long file-names are allowed (in SAGA)
# --> convert 50000_548wgs84_azienda2k5tinlin.tif into dem1.asc prior to loading
rsaga.esri.to.sgrd(env = myenv, in.grids = "phd_calc_input/dem1.asc",
  out.sgrds = "phd_calc_out/demanalysis_r2/dem_az2k5.sgrd", in.path = getwd())


# SAGA Wetness Index:
rsaga.geoprocessor(lib = "ta_hydrology", module = 15, env = myenv,
  param = list(DEM = "phd_calc_out/demanalysis_r2/dem_az2k5.sgrd",
    C = "phd_calc_out/demanalysis_r2/dtaout_az2k5/catcharea1.sgrd",
    GN = "phd_calc_out/demanalysis_r2/dtaout_az2k5/catchslope1.sgrd",
    CS = "phd_calc_out/demanalysis_r2/dtaout_az2k5/modcatcharea1.sgrd",
```

```
60     SB = "phd_calc_out/demanalysis_r2/dtaout_az2k5/sagawi1.sgrd", T = 10))
61
62 # primary attributes (local morphometry)
63 # after Zevenbergen & Thorne 1987 = METHOD 5:
64 # slope in rad (m/m):
65 rsaga.geoprocessor(lib = "ta_morphometry", module = 0, env = myenv,
66   param = list(ELEVATION = "phd_calc_out/demanalysis_r2/dem_az2k5.sgrd",
67     SLOPE = "phd_calc_out/demanalysis_r2/dtaout_az2k5/sloperad1.sgrd",
68     ASPECT = "phd_calc_out/demanalysis_r2/dtaout_az2k5/aspect1.sgrd",
69     HCURV = "phd_calc_out/demanalysis_r2/dtaout_az2k5/plancurv1.sgrd",
70     VCURV = "phd_calc_out/demanalysis_r2/dtaout_az2k5/profcurv1.sgrd",
71     METHOD = 5))
72
73 # convert curvatures from 1/m to 1/100m:
74 rsaga.geoprocessor(lib = "grid_calculus", module = 1, env = myenv,
75  param = list(GRIDS = "phd_calc_out/demanalysis_r2/dtaout_az2k5/profcurv1.sgrd",
76    RESULT = "phd_calc_out/demanalysis_r2/dtaout_az2k5/profcurv1001.sgrd",
77    FORMULA = "g1*100", FNAME = T))
78 rsaga.geoprocessor(lib = "grid_calculus", module = 1, env = myenv,
79  param = list(GRIDS = "phd_calc_out/demanalysis_r2/dtaout_az2k5/plancurv1.sgrd",
80    RESULT = "phd_calc_out/demanalysis_r2/dtaout_az2k5/plancurv1001.sgrd",
81    FORMULA = "g1*100", FNAME = T))
82
83 # convert aspect from rad to degree:
84 rsaga.geoprocessor(lib = "grid_calculus", module = 1, env = myenv,
85   param = list(GRIDS = "phd_calc_out/demanalysis_r2/dtaout_az2k5/aspect1.sgrd",
86     RESULT = "phd_calc_out/demanalysis_r2/dtaout_az2k5/aspectdeg1.sgrd",
87     FORMULA = "g1*180/pi()", FNAME = T))
88
89 # convergence/divergence index:
90 rsaga.geoprocessor(lib = "ta_morphometry", module = 2, env = myenv,
91   param = list(ELEVATION = "phd_calc_out/demanalysis_r2/dem_az2k5.sgrd",
92     CONVERGENCE = "phd_calc_out/demanalysis_r2/dtaout_az2k5/convg1.sgrd",
93     RADIUS = 3, DISTANCE_WEIGHTING_WEIGHTING = 0, SLOPE = T))
94
95
96 # secondary attributes (topographic indices):
97 # topographic wetness index:
98 rsaga.geoprocessor(lib = "ta_hydrology", module = 20, env = myenv,
99  param = list(SLOPE = "phd_calc_out/demanalysis_r2/dtaout_az2k5/sloperad1.sgrd",
100   AREA = "phd_calc_out/demanalysis_r2/dtaout_az2k5/catcharea1.sgrd",
101   TWI = "phd_calc_out/demanalysis_r2/dtaout_az2k5/twi1.sgrd"))
102
103 # stream power index:
104 rsaga.geoprocessor(lib = "ta_hydrology", module = 21, env = myenv,
105  param = list(SLOPE = "phd_calc_out/demanalysis_r2/dtaout_az2k5/sloperad1.sgrd",
106   AREA = "phd_calc_out/demanalysis_r2/dtaout_az2k5/catcharea1.sgrd",
107   SPI = "phd_calc_out/demanalysis_r2/dtaout_az2k5/streampow1.sgrd"))
108
109 # LS-Factor (Moore et al. 1991, Erosivity = 1):
```

```
rsaga.geoprocessor(lib = "ta_hydrology", module = 22, env = myenv,
 param = list(SLOPE = "phd_calc_out/demanalysis_r2/dtaout_az2k5/sloperad1.sgrd",
   AREA = "phd_calc_out/demanalysis_r2/dtaout_az2k5/catcharea1.sgrd",
   LS = "phd_calc_out/demanalysis_r2/dtaout_az2k5/ls1.sgrd",
   CONV = 0, METHOD = 0, EROSIVITY = 1, STABILITY = 0))

# converting the resulting grids to ESRI-ASCII Grid:
rsaga.sgrd.to.esri(
  in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_az2k5/sloperad1.sgrd",
  out.grids = "phd_calc_out/demanalysis_r2/dtaout_az2k5/sloperad1.asc",
  prec = 6, out.path = getwd(), env = myenv)
rsaga.sgrd.to.esri(
  in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_az2k5/sagawi1.sgrd",
  out.grids = "phd_calc_out/demanalysis_r2/dtaout_az2k5/sagawi1.asc",
  prec = 2, out.path = getwd(), env = myenv)
rsaga.sgrd.to.esri(
  in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_az2k5/plancurv1001.sgrd",
  out.grids = "phd_calc_out/demanalysis_r2/dtaout_az2k5/plancurv1001.asc",
  prec = 6, out.path = getwd(), env = myenv)
rsaga.sgrd.to.esri(
  in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_az2k5/profcurv1001.sgrd",
  out.grids = "phd_calc_out/demanalysis_r2/dtaout_az2k5/profcurv1001.asc",
  prec = 6, out.path = getwd(), env = myenv)
rsaga.sgrd.to.esri(
  in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_az2k5/aspectdeg1.sgrd",
  out.grids = "phd_calc_out/demanalysis_r2/dtaout_az2k5/aspectdeg1.asc",
  prec = 2, out.path = getwd(), env = myenv)
rsaga.sgrd.to.esri(
  in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_az2k5/convg1.sgrd",
  out.grids = "phd_calc_out/demanalysis_r2/dtaout_az2k5/convg1.asc",
  prec = 4, out.path = getwd(), env = myenv)
rsaga.sgrd.to.esri(
  in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_az2k5/twi1.sgrd",
  out.grids = "phd_calc_out/demanalysis_r2/dtaout_az2k5/twi1.asc",
  prec = 2, out.path = getwd(), env = myenv)
rsaga.sgrd.to.esri(
  in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_az2k5/streampow1.sgrd",
  out.grids = "phd_calc_out/demanalysis_r2/dtaout_az2k5/streampow1.asc",
  prec = 2, out.path = getwd(), env = myenv)
rsaga.sgrd.to.esri(
  in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_az2k5/ls1.sgrd",
  out.grids = "phd_calc_out/demanalysis_r2/dtaout_az2k5/ls1.asc",
  prec = 2, out.path = getwd(), env = myenv)


#------------------------------------------------------------------------------
## Creating dataset + target grid for digital soil mapping
#------------------------------------------------------------------------------

# all dem1-related relief parameter were clipped
```

```
#  in ArcGIS 10 to the boundaries of the fields 21 and 33, reload them:
j <- c("SAGAWI","SLOPE","ASPECT","PLANC","PROFC","CONVG","TWI","STREAMP","LS")
k <- c("sagawi1cl", "sloperad1cl", "aspectdeg1cl", "plancurv1001cl",
  "profcurv1001cl", "convg1cl", "twi1cl", "streampow1cl", "ls1cl")
m = 1

for (i in k) {
  demt@data[i] <- readGDAL(paste0("phd_calc_out/demanalysis_r2/dtaout_az2k5/",
    i, ".tif"))$band1
  names(demt)[m+1] <- j[m]; m <- m + 1
}
rm(i,j,k,m)

# load the laboratory data:
d21.all <- read.csv2("phd_calc_input/field21_v6.csv", na.strings = "-999")
d.33 <- read.csv2("phd_calc_input/field33_v2.csv")

d21.A <- subset(d21.all, d21.all$HORIZON == "A")
d21.B <- subset(d21.all, d21.all$HORIZON == "B")

# sort IDs 7-13 from d21.A to d.33 as they are actually part of f33,
#  but were sampled in 2010 along with f21-sampling:
d.33 <- rbind(d.33, d21.A[which(d21.A$ID %in% 7:13),])
d.33$ID[which(d.33$FIELD == 21)] <- 37:43 # new IDs

# remove the moved:
d21.A <- d21.A[-which(d21.A$ID %in% 7:13),]
d21.B <- d21.B[-which(d21.B$ID %in% 7:13),]

d.21 <- d21.A; rm(d21.all,d21.A)
# --> d.21: n = 43; d.33: n = 64

coordinates(d.21) <- ~ EAST + NORTH; coordinates(d.33) <- ~ EAST + NORTH
coordinates(d21.B) <- ~ EAST + NORTH

# ensure identical CRS (EPSG:32632):
proj4string(demt) <- CRS(NA)
proj4string(demt) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
proj4string(d.33) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
proj4string(d.21) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
proj4string(d21.B) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")

# logratio transform to account for the compositional character of the targets:
d.comp <- acomp(d.21@data, parts = c("CLAY", "SILT", "SAND"), total = 100)
d.alr <- alr(d.comp) # additive logratio transform!
d.21$CLAYalr <- d.alr[,1]; d.21$SILTalr <- d.alr[,2]
```

```
210
211 d.comp <- acomp(d.33@data, parts = c("CLAY", "SILT", "SAND"), total = 100)
212 d.alr <- alr(d.comp) # additive logratio transform!
213 d.33$CLAYalr <- d.alr[,1]; d.33$SILTalr <- d.alr[,2]
214
215 d.comp <- acomp(d21.B@data, parts = c("CLAY", "SILT", "SAND"), total = 100)
216 d.alr <- alr(d.comp) # additive logratio transform!
217 d21.B$CLAYalr <- d.alr[,1]; d21.B$SILTalr <- d.alr[,2]
218
219 rm(d.alr,d.comp)
220
221 d <- list(d.21, d.33, d21.B)
222 j <- c(
223   "ELEV","SAGAWI","SLOPE","ASPECT","PROFC","PLANC","CONVG","TWI","STREAMP","LS")
224 for (q in 1:length(d)) {
225   ov <- over(d[[q]], demt); d[[q]]@data[,j] <- ov[,j]
226 }
227 d.21 <- d[[1]]; d.33 <- d[[2]]; d21.B <- d[[3]]
228 rm(ov,d,dem1,myenv,j,q)
229
230 save.image(paste(getwd(), "phd_calc_input/az_basisDTA.RData", sep = "/"))
231
232 # end of script, azienda_dta2k5_vDiss.R, 27.01.2012
```

.4.3 azienda_geophysics2k5_vDiss.R

```
 1 #
 2 ##############################################################################
 3 #### Geophysical sensing data for digital soil mapping, Azienda San Michele ####
 4 ##############################################################################
 5 #
 6 ## last update: 02.12.2014
 7
 8 # preparation of geophysical covariates at field scale
 9
10 # author: Michael Blaschek (blaschek@geographie.uni-kiel.de)
11
12 rm(list = ls())
13
14 library(rgdal); library(gstat); library(RColorBrewer); library(ggplot2)
15 library(grid); library(gridExtra); library(gtable)
16 # R v3.0.2, rgdal_0.8-15, gstat_1.0-16, RColorBrewer_1.0-5, ggplot2_0.9.3.1
17 #   gridExtra_0.9.1, gtable_0.1.2
18
19
20 #-----------------------------------------------------------------------------
21 ## Load geophysical sensor data
22 #-----------------------------------------------------------------------------
23
```

```
# geophysical measurements, UFZ:
dgo <- read.table(paste(
  getwd(), "phd_calc_input/ussana_gamma_1009-1010_gesamtfl che.txt", sep = "/")
      ,
  sep = "\t", dec = ".", header = T)
coordinates(dgo) <- ~ E + N
writeOGR(dgo, paste(getwd(), "phd_calc_input/emi_shps", sep = "/"),
  "dgo", driver = "ESRI Shapefile")
# (d)ata (g)amma (o)riginal
# received from Ulrike Werban, UFZ, 14.12.2010 via www.transfer.ufz.de

# problem: missing points in the North of field33
# manually completed from gamma_all.txt (using ArcGIS 10 - Select)
# clipped to field 21/33-boundaries (in ArcGIS 10)
dgo.21 <- readOGR(paste(
  getwd(), "phd_calc_input/emi_shps", sep = "/"), "dgo2_21")
dgo.33 <- readOGR(paste(
  getwd(), "phd_calc_input/emi_shps", sep = "/"), "dgo2_33")

demio.2 <- read.table(paste(getwd(), "phd_calc_input/ussana_EM38_all_2.txt",
  sep = "/"), sep = "\t", dec = ".", header = T)
#coordinates(demio.2) <- ~ E + N
#writeOGR(demio.2, paste(getwd(), "phd_calc_input/emi_shps", sep = "/"),
#  "demio2", driver = "ESRI Shapefile")

demio.5 <- read.table(paste(getwd(), "phd_calc_input/ussana_EM38_all_5.txt",
  sep = "/"), sep = "\t", dec = ".", header = T)
#coordinates(demio.5) <- ~ E + N
#writeOGR(demio.5, paste(getwd(), "phd_calc_input/emi_shps", sep = "/"),
#  "demio5", driver = "ESRI Shapefile")

demio.6 <- read.table(paste(getwd(), "phd_calc_input/ussana_EM38_all_6.txt",
  sep = "/"), sep = "\t", dec = ".", header = T)
#coordinates(demio.6) <- ~ E + N
#writeOGR(demio.6, paste(getwd(), "phd_calc_input/emi_shps", sep = "/"),
#  "demio6", driver = "ESRI Shapefile")

# (d)ata (emi) (o)riginal
# received from Ulrike Werban, UFZ, 14.12.2010 via www.transfer.ufz.de
# ussana_EM38_all.txt with different col-length
# --> divided into subfiles using Notepad++
# relevant for field 33:
#  ussana_EM38_all_2.txt, ussana_EM38_all_5.txt, ussana_EM38_all_6.txt
# NA = NA, before loading: #WERT! replaced by NA in Notepad++

# delete those points with NA-values in H and V column:
demio.2 <- demio.2[-which(is.na(demio.2$H)),]
demio.2 <- demio.2[-which(is.na(demio.2$V)),]

# same as for demio.2, but delete also columns: Hr, Hkor, Hur, H2, H.1
```

```
demio.5 <- demio.5[-which(is.na(demio.5$H)),1:13]
demio.5 <- demio.5[-which(is.na(demio.5$V)),]

# same as for demio.2, but delete also columns: V.1, H.1, V2, H2, V.2, H.2
demio.6 <- demio.6[-which(is.na(demio.6$H)),1:13]
demio.6 <- demio.6[-which(is.na(demio.6$V)),]

demio.2$SET <- 2; demio.5$SET <- 5; demio.6$SET <- 6

demio <- rbind(demio.2, demio.5, demio.6)
coordinates(demio) <- ~ E + N
proj4string(demio) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
writeOGR(demio, paste(getwd(), "phd_calc_input/emi_shps", sep = "/"),
  "demio_33", driver = "ESRI Shapefile")

# demio1 is simply field 21, renamed to demio_21.shp
demio.21 <- readOGR(paste(
  getwd(), "phd_calc_input/emi_shps", sep = "/"), "demio_21")
demio.33 <- readOGR(paste(
  getwd(), "phd_calc_input/emi_shps", sep = "/"), "demio_33")

# --> work with dgo.21/dgo.33 and demio.21/demio.33
# --> concluded to fields2133_geophysics_raw_v1.RData
rm(demio,demio.2,demio.5,demio.6,dgo)
save.image(paste(
  getwd(), "phd_calc_input/fields2133_geophysics_raw_v1.RData", sep = "/"))


#-------------------------------------------------------------------------------
## Generate geophysical covariates
#-------------------------------------------------------------------------------

# load DEMs:
# calculation is done for each field separately
dem.21 <- readGDAL(paste(
  getwd(), "phd_calc_input/50000_548wgs84_field21_2k5tinlin.tif", sep = "/"))
dem.33 <- readGDAL(paste(
  getwd(), "phd_calc_input/50000_548wgs84_field33_2k5tinlin.tif", sep = "/"))
names(dem.21)[1] <- "ELEV"; names(dem.33)[1] <- "ELEV"

zerodist(demio.21); zerodist(demio.33); zerodist(dgo.33); zerodist(dgo.21)
# --> only dgo.21 contains duplicate points, removed prior to interpolation:
dgo.21 <- dgo.21[-as.vector(zerodist(dgo.21)),]

# field 21:
# OK prediction of the EMI-covariates:
n <- dim(demio.21@data)[1]
# random selection of validation points, 1/3:
v.id <- sample(1:dim(demio.21@data)[1], dim(demio.21@data)[1]/3)
```

```
dco.v <- demio.21[v.id,]
dco.c <- demio.21[!is.element(1:dim(demio.21@data)[1], v.id),]
n.v <- length(v.id); n.c <- n - n.v

diag.bbox <- sqrt((bbox(demio.21)[1,1] -
  bbox(demio.21)[1,2])^2 + (bbox(demio.21)[2,1] - bbox(demio.21)[2,2])^2)
diag.bbox/2; diag.bbox/3 # --> distance of reliability between 144 and 217m
va.21.H <- variogram(H ~ 1, loc = dco.c, cutoff = 180, width = 2.5)
vm.21.H <- vgm(34, "Sph", 110, 21)
vmf.21.H <- fit.variogram(va.21.H, vm.21.H, fit.method = 7)
va.21.V <- variogram(V ~ 1, loc = dco.c, cutoff = 180, width = 2.5)
vm.21.V <- vgm(55, "Sph", 100, 8)
vmf.21.V <- fit.variogram(va.21.V, vm.21.V, fit.method = 7)

ok.out.21.H <- krige(H ~ 1, loc = dco.c, newdata = dem.21, model = vmf.21.H)
ok.out.21.V <- krige(V ~ 1, loc = dco.c, newdata = dem.21, model = vmf.21.V)

dem.21$EMIH <- ok.out.21.H$var1.pred; dem.21$EMIV <- ok.out.21.V$var1.pred

ok.cv.21.H <- krige(H ~ 1, loc = dco.c, newdata = dco.v, model = vmf.21.H)
ok.cv.21.V <- krige(V ~ 1, loc = dco.c, newdata = dco.v, model = vmf.21.V)

obs.vs.pred <- lm((dco.v$H ~ ok.cv.21.H$var1.pred))
sum.lm <- summary(obs.vs.pred)
r2.21.H <- sum.lm$r.squared # 0.55
rmse.21.H <- sqrt(mean((ok.cv.21.H$var1.pred - dco.v$H)^2)); rmse.21.H # 4.528
obs.vs.pred <- lm((dco.v$V ~ ok.cv.21.V$var1.pred))
sum.lm <- summary(obs.vs.pred)
r2.21.V <- sum.lm$r.squared; r2.21.V # 0.89
rmse.21.V <- sqrt(mean((ok.cv.21.V$var1.pred - dco.v$V)^2)); rmse.21.V # 2.425

# field 33:
# OK prediction of the EMI-covariates:
n <- dim(demio.33@data)[1]
# random selection of validation points, 1/3:
v.id <- sample(1:dim(demio.33@data)[1], dim(demio.33@data)[1]/3)
dco.v <- demio.33[v.id,]
dco.c <- demio.33[!is.element(1:dim(demio.33@data)[1], v.id),]
n.v <- length(v.id); n.c <- n - n.v

diag.bbox <- sqrt((bbox(demio.33)[1,1] -
  bbox(demio.33)[1,2])^2 + (bbox(demio.33)[2,1] - bbox(demio.33)[2,2])^2)
diag.bbox/2; diag.bbox/3 # --> distance of reliability between 210 and 317m
va.33.H <- variogram(H ~ 1, loc = dco.c, cutoff = 260, width = 2.5)
vm.33.H <- vgm(50, "Sph", 50, 0)
vmf.33.H <- fit.variogram(va.33.H, vm.33.H)
vm.33.H <- vgm(20, "Sph", 200, add.to = vmf.33.H)
vmf2.33.H <- fit.variogram(va.33.H, vm.33.H, fit.method = 6)
va.33.V <- variogram(V ~ 1, loc = dco.c, cutoff = 260, width = 2.5)
vm.33.V <- vgm(50, "Sph", 50, 0)
```

```
vmf.33.V <- fit.variogram(va.33.V, vm.33.V)
vm.33.V <- vgm(20, "Sph", 200, add.to = vmf.33.V)
vmf2.33.V <- fit.variogram(va.33.V, vm.33.V, fit.method = 6)

ok.out.33.H <- krige(H ~ 1, loc = dco.c, newdata = dem.33, model = vmf2.33.H)
ok.out.33.V <- krige(V ~ 1, loc = dco.c, newdata = dem.33, model = vmf2.33.V)

dem.33$EMIH <- ok.out.33.H$var1.pred; dem.33$EMIV <- ok.out.33.V$var1.pred

ok.cv.33.H <- krige(H ~ 1, loc = dco.c, newdata = dco.v, model = vmf2.33.H)
ok.cv.33.V <- krige(V ~ 1, loc = dco.c, newdata = dco.v, model = vmf2.33.V)

obs.vs.pred <- lm((dco.v$H ~ ok.cv.33.H$var1.pred))
sum.lm <- summary(obs.vs.pred)
r2.33.H <- sum.lm$r.squared; r2.33.H # 0.986
rmse.33.H <- sqrt(mean((ok.cv.33.H$var1.pred - dco.v$H)^2)); rmse.33.H # 0.917
obs.vs.pred <- lm((dco.v$V ~ ok.cv.33.V$var1.pred))
sum.lm <- summary(obs.vs.pred)
r2.33.V <- sum.lm$r.squared; r2.33.V # 0.996
rmse.33.V <- sqrt(mean((ok.cv.33.V$var1.pred - dco.v$V)^2)); rmse.33.V # 0.792

# field 21:
# OK prediction of the gamma-ray-covariates:
n <- dim(dgo.21@data)[1]
# random selection of validation points, 1/3:
v.id <- sample(1:dim(dgo.21@data)[1], dim(dgo.21@data)[1]/3)
dco.v <- dgo.21[v.id,]
dco.c <- dgo.21[!is.element(1:dim(dgo.21@data)[1], v.id),]
n.v <- length(v.id); n.c <- n - n.v

diag.bbox <- sqrt((bbox(dgo.21)[1,1] -
  bbox(dgo.21)[1,2])^2 + (bbox(dgo.21)[2,1] - bbox(dgo.21)[2,2])^2)
diag.bbox/2; diag.bbox/3 # --> distance of reliability between 144 and 217m
va.21.K <- variogram(K ~ 1, loc = dco.c, cutoff = 150, width = 2.5)
vm.21.K <- vgm(.3, "Sph", 90, 0.02)
vmf.21.K <- fit.variogram(va.21.K, vm.21.K, fit.method = 7)
va.21.U <- variogram(U ~ 1, loc = dco.c, cutoff = 150, width = 2.5)
vm.21.U <- vgm(.25, "Sph", 90, 0.4)
vmf.21.U <- fit.variogram(va.21.U, vm.21.U, fit.method = 6)
va.21.TH <- variogram(TH ~ 1, loc = dco.c, cutoff = 150, width = 2.5)
vm.21.TH <- vgm(3, "Sph", 90, 1)
vmf.21.TH <- fit.variogram(va.21.TH, vm.21.TH, fit.method = 7)
va.21.DR <- variogram(DR ~ 1, loc = dco.c, cutoff = 150, width = 2.5)
vm.21.DR <- vgm(200, "Sph", 90, 5)
vmf.21.DR <- fit.variogram(va.21.DR, vm.21.DR, fit.method = 7)

ok.out.21.K <- krige(K ~ 1, loc = dco.c, newdata = dem.21, model = vmf.21.K)
ok.out.21.U <- krige(U ~ 1, loc = dco.c, newdata = dem.21, model = vmf.21.U)
ok.out.21.TH <- krige(TH ~ 1, loc = dco.c, newdata = dem.21, model = vmf.21.TH)
ok.out.21.DR <- krige(DR ~ 1, loc = dco.c, newdata = dem.21, model = vmf.21.DR)
```

```
223
224 dem.21$GAMMAK <- ok.out.21.K$var1.pred; dem.21$GAMMATH <- ok.out.21.TH$var1.pred
225 dem.21$GAMMAU <- ok.out.21.U$var1.pred; dem.21$GAMMADR <- ok.out.21.DR$var1.pred
226
227 ok.cv.21.K <- krige(K ~ 1, loc = dco.c, newdata = dco.v, model = vmf.21.K)
228 ok.cv.21.U <- krige(U ~ 1, loc = dco.c, newdata = dco.v, model = vmf.21.U)
229 ok.cv.21.TH <- krige(TH ~ 1, loc = dco.c, newdata = dco.v, model = vmf.21.TH)
230 ok.cv.21.DR <- krige(DR ~ 1, loc = dco.c, newdata = dco.v, model = vmf.21.DR)
231
232 obs.vs.pred <- lm((dco.v$K ~ ok.cv.21.K$var1.pred))
233 sum.lm <- summary(obs.vs.pred)
234 r2.21.K <- sum.lm$r.squared; r2.21.K # 0.83
235 rmse.21.K <- sqrt(mean((ok.cv.21.K$var1.pred - dco.v$K)^2)); rmse.21.K # 0.180
236 obs.vs.pred <- lm((dco.v$U ~ ok.cv.21.U$var1.pred))
237 sum.lm <- summary(obs.vs.pred)
238 r2.21.U <- sum.lm$r.squared; r2.21.U # 0.30
239 rmse.21.U <- sqrt(mean((ok.cv.21.U$var1.pred - dco.v$U)^2)); rmse.21.U # 0.635
240 obs.vs.pred <- lm((dco.v$TH ~ ok.cv.21.TH$var1.pred))
241 sum.lm <- summary(obs.vs.pred)
242 r2.21.TH <- sum.lm$r.squared; r2.21.TH # 0.63
243 rmse.21.TH <- sqrt(mean((ok.cv.21.TH$var1.pred - dco.v$TH)^2)) # 1.046
244 obs.vs.pred <- lm((dco.v$DR ~ ok.cv.21.DR$var1.pred))
245 sum.lm <- summary(obs.vs.pred)
246 r2.21.DR <- sum.lm$r.squared; r2.21.DR # 0.88
247 rmse.21.DR <- sqrt(mean((ok.cv.21.DR$var1.pred - dco.v$DR)^2)) # 4.030
248
249 # field 33:
250 # OK prediction of the gamma-ray-covariates:
251 n <- dim(dgo.33@data)[1]
252 # random selection of validation points, 1/3:
253 v.id <- sample(1:dim(dgo.33@data)[1], dim(dgo.33@data)[1]/3)
254 dco.v <- dgo.33[v.id,]
255 dco.c <- dgo.33[!is.element(1:dim(dgo.33@data)[1], v.id),]
256 n.v <- length(v.id); n.c <- n - n.v
257
258 diag.bbox <- sqrt((bbox(dgo.33)[1,1] -
259   bbox(dgo.33)[1,2])^2 + (bbox(dgo.33)[2,1] - bbox(dgo.33)[2,2])^2)
260 diag.bbox/2; diag.bbox/3 # --> distance of reliability between 207 and 311m
261 va.33.K <- variogram(K ~ 1, loc = dco.c, cutoff = 125, width = 2.5)
262 vm.33.K <- vgm(.1, "Sph", 70, 0.03)
263 vmf.33.K <- fit.variogram(va.33.K, vm.33.K, fit.method = 7)
264 va.33.U <- variogram(U ~ 1, loc = dco.c, cutoff = 125, width = 2.5)
265 vm.33.U <- vgm(.2, "Sph", 60, 0.3)
266 vmf.33.U <- fit.variogram(va.33.U, vm.33.U, fit.method = 7)
267 va.33.TH <- variogram(TH ~ 1, loc = dco.c, cutoff = 125, width = 2.5)
268 vm.33.TH <- vgm(1.5, "Sph", 80, 1)
269 vmf.33.TH <- fit.variogram(va.33.TH, vm.33.TH, fit.method = 7)
270 va.33.DR <- variogram(DR ~ 1, loc = dco.c, cutoff = 125, width = 2.5)
271 vm.33.DR <- vgm(65, "Sph", 80, 5)
272 vmf.33.DR <- fit.variogram(va.33.DR, vm.33.DR, fit.method = 7)
```

```
273
274 ok.out.33.K <- krige(K ~ 1, loc = dco.c, newdata = dem.33, model = vmf.33.K)
275 ok.out.33.U <- krige(U ~ 1, loc = dco.c, newdata = dem.33, model = vmf.33.U)
276 ok.out.33.TH <- krige(TH ~ 1, loc = dco.c, newdata = dem.33, model = vmf.33.TH)
277 ok.out.33.DR <- krige(DR ~ 1, loc = dco.c, newdata = dem.33, model = vmf.33.DR)
278
279 dem.33$GAMMAK <- ok.out.33.K$var1.pred; dem.33$GAMMATH <- ok.out.33.TH$var1.pred
280 dem.33$GAMMAU <- ok.out.33.U$var1.pred; dem.33$GAMMADR <- ok.out.33.DR$var1.pred
281
282 ok.cv.33.K <- krige(K ~ 1, loc = dco.c, newdata = dco.v, model = vmf.33.K)
283 ok.cv.33.U <- krige(U ~ 1, loc = dco.c, newdata = dco.v, model = vmf.33.U)
284 ok.cv.33.TH <- krige(TH ~ 1, loc = dco.c, newdata = dco.v, model = vmf.33.TH)
285 ok.cv.33.DR <- krige(DR ~ 1, loc = dco.c, newdata = dco.v, model = vmf.33.DR)
286
287 obs.vs.pred <- lm((dco.v$K ~ ok.cv.33.K$var1.pred))
288 sum.lm <- summary(obs.vs.pred)
289 r2.33.K <- sum.lm$r.squared; r2.33.K # 0.70
290 rmse.33.K <- sqrt(mean((ok.cv.33.K$var1.pred - dco.v$K)^2)); rmse.33.K # 0.195
291 obs.vs.pred <- lm((dco.v$U ~ ok.cv.33.U$var1.pred))
292 sum.lm <- summary(obs.vs.pred)
293 r2.33.U <- sum.lm$r.squared; r2.33.U # 0.16
294 rmse.33.U <- sqrt(mean((ok.cv.33.U$var1.pred - dco.v$U)^2)); rmse.33.U # 0.644
295 obs.vs.pred <- lm((dco.v$TH ~ ok.cv.33.TH$var1.pred))
296 sum.lm <- summary(obs.vs.pred)
297 r2.33.TH <- sum.lm$r.squared; r2.33.TH # 0.44
298 rmse.33.TH <- sqrt(mean((ok.cv.33.TH$var1.pred - dco.v$TH)^2)) # 1.055
299 obs.vs.pred <- lm((dco.v$DR ~ ok.cv.33.DR$var1.pred))
300 sum.lm <- summary(obs.vs.pred)
301 r2.33.DR <- sum.lm$r.squared; r2.33.DR # 0.80
302 rmse.33.DR <- sqrt(mean((ok.cv.33.DR$var1.pred - dco.v$DR)^2)) # 3.889
303
304 rm(sum.lm,v.id,obs.vs.pred,n.v,n.c,n,diag.bbox,dco.v,dco.c)
305
306 rm(list = ls(pattern = "va")); rm(list = ls(pattern = "vm"))
307 rm(list = ls(pattern = "ok"))
308 rm(list = ls(pattern = "r2")); rm(list = ls(pattern = "rmse"))
309
310 # K was in %, TH and U in ppm
311 dem.21$THKratio <- (dem.21$GAMMATH/1000000)/(dem.21$GAMMAK/100) * 1000000
312 dem.33$THKratio <- (dem.33$GAMMATH/1000000)/(dem.33$GAMMAK/100) * 1000000
313
314 dem.21$KTHratio <- (dem.21$GAMMAK/100)/(dem.21$GAMMATH/1000000) * 1000000
315 dem.33$KTHratio <- (dem.33$GAMMAK/100)/(dem.33$GAMMATH/1000000) * 1000000
316
317 dem.21$UKratio <- (dem.21$GAMMAU/1000000)/(dem.21$GAMMAK/100) * 1000000
318 dem.33$UKratio <- (dem.33$GAMMAU/1000000)/(dem.33$GAMMAK/100) * 1000000
319
320 dem.21$THUratio <- dem.21$GAMMATH/dem.21$GAMMAU
321 dem.33$THUratio <- dem.33$GAMMATH/dem.33$GAMMAU
322
```

```
323
324 # combine results from DTA and geophysical interpolation:
325 dem33.save <- dem.33; dem21.save <- dem.21
326 demt.save <- demt; d21.save <- d.21; d33.save <- d.33
327
328 catchm <- readOGR(dsn = "phd_calc_input", layer = "fields2133_diss_v2")
329
330 proj4string(catchm) <- CRS(NA); proj4string(dem.33) <- CRS(NA)
331 proj4string(dem.21) <- CRS(NA); proj4string(demt) <- CRS(NA)
332 proj4string(catchm) <-
333   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
334 proj4string(dem.33) <-
335   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
336 proj4string(dem.21) <-
337   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
338 proj4string(demt) <-
339   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
340
341 # from digital terrain analysis:
342 j <- c(
343   "ELEV","SAGAWI","SLOPE","ASPECT","PROFC","PLANC","CONVG","TWI","STREAMP","LS")
344 dem33.ov <- over(dem.33, demt); dem.33@data[,j] <- dem33.ov[,j]
345 dem21.ov <- over(dem.21, demt); dem.21@data[,j] <- dem21.ov[,j]
346
347 # set non-field values to NA:
348 asdf <- which(!is.na(over(dem.33, catchm[1,1]))); dem.33 <- dem.33[-asdf,]
349 asdf <- which(!is.na(over(dem.21, catchm[2,1]))); dem.21 <- dem.21[-asdf,]
350
351 # from geophysical sensors:
352 j.g <- c("EMIH", "EMIV", "GAMMAK", "GAMMATH", "GAMMAU", "GAMMADR", "THKratio",
353   "KTHratio", "UKratio", "THUratio")
354 dem33.ov <- over(d.33, dem.33); d.33@data[,j.g] <- dem33.ov[,j.g]
355 dem21.ov <- over(d.21, dem.21); d.21@data[,j.g] <- dem21.ov[,j.g]
356 dem21.ov <- over(d21.B, dem.21); d21.B@data[,j.g] <- dem21.ov[,j.g]
357
358
359 # plot EMI-covariates:
360 pts.21 <- list("sp.points", d.21, pch = 19, col = "black", cex = .6, alpha = 1)
361 pts.33 <- list("sp.points", d.33, pch = 19, col = "black", cex = .6, alpha = 1)
362
363 prof.21 <- readOGR(dsn = "phd_calc_input", layer = "az_profiles11_pnts21")
364 prof.33 <- readOGR(dsn = "phd_calc_input", layer = "az_profiles13_pnts33")
365
366 cat <- fortify(catchm, region = "DISS")
367 pt21 <- as.data.frame(pts.21); pt33 <- as.data.frame(pts.33)
368 prof21 <- as.data.frame(prof.21); prof33 <- as.data.frame(prof.33)
369 dem21.df <- as.data.frame(dem.21); dem33.df <- as.data.frame(dem.33)
370 dem <- rbind(dem21.df, dem33.df)
371 v.interest <- c("EMIH","EMIV", "x", "y")
372 dem.emiH <- dem[,v.interest]
```

```
dem.emiH$VARS <- "EMI-H"; dem.emiH$VALUE <- dem.emiH$EMIH
dem.emiV <- dem[,v.interest]
dem.emiV$VARS <- "EMI-V"; dem.emiV$VALUE <- dem.emiV$EMIV
dem.emiH <- dem.emiH[,-c(1:2)]; dem.emiV <- dem.emiV[,-c(1:2)]
demm <- rbind(dem.emiH, dem.emiV)

emi.min <- min(demm$VALUE, na.rm = TRUE); emi.min # 8.4
emi.max <- max(demm$VALUE, na.rm = TRUE); emi.max # 85.9

demm$CUTV <- cut(demm$VALUE, breaks = seq(0,90,10))
levels(demm$CUTV) <- c("0 - 10", "10 - 20", "20 - 30", "30 - 40", "40 - 50",
  "50 - 60", "60 - 70", "70 - 80", "80 - 90")

map <- ggplot(cat, aes(long, lat)) +
  geom_raster(aes(x, y, fill = CUTV), data = demm) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1,
    data = pt21, shape = 21) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1,
    data = pt33, shape = 21) +
  geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 2,
    data = prof21, shape = 18) +
  geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 2,
    data = prof33, shape = 18) +
  geom_polygon(size = .5, linetype = "dashed", color = "black",
    fill = "grey40", alpha = 0) +
  coord_equal() +
  theme_bw(base_size = 9, base_family = "Helvetica") +
  scale_fill_manual(name = "ECa in mS/m", values = rev(brewer.pal(9,"Spectral"))
      ) +
  labs(x = "\nUTM-E/m", y = "UTM-N/m\n") +
  scale_y_continuous(breaks = c(4362600,4362800,4363000)) +
  scale_x_continuous(breaks = c(508800,509000,509200)) +
  facet_wrap(~ VARS, ncol = 2) +
  theme(axis.text.y = element_text(angle = 90, hjust = .5),
    legend.text = element_text(size = 10),
    legend.title = element_text(size = 10),
    strip.text.x = element_text(size = 12),
    axis.text = element_text(size = 10), axis.title = element_text(size = 10))

ggsave(paste(path.fig, "azienda_emi_interpolOK_v1.pdf", sep = "/"), map,
  width = 8.27, height = 4.59)
print(map)
dev.off()
rm(emi.max,emi.min,map,v.interest,demm,dem33.df,
  dem21.df,dem.emiH,dem.emiV,dem21.ov,dem33.ov,j,j.g)


# plot gamma-ray-covariates:
#dgo.21 <- readOGR(paste(
#  getwd(), "phd_calc_input/emi_shps", sep = "/"), "dgo2_21")
```

```
#dgo.33 <- readOGR(paste(
#  getwd(), "phd_calc_input/emi_shps", sep = "/"), "dgo2_33")
#dgo21 <- as.data.frame(dgo.21); dgo33 <- as.data.frame(dgo.33)
v.interest <- c("GAMMATH","GAMMAK", "GAMMAU", "GAMMADR", "x", "y")
dem.gamma <- dem[,v.interest]

gammaTH.min <- min(dem.gamma$GAMMATH, na.rm = TRUE); gammaTH.min # 2.7
gammaTH.max <- max(dem.gamma$GAMMATH, na.rm = TRUE); gammaTH.max # 9.1

dem.gamma$THCUTV <- cut(dem.gamma$GAMMATH, breaks = seq(2.7, 9.45, by = .75))
levels(dem.gamma$THCUTV) <- c("2.70 - 3.45", "3.45 - 4.20", "4.20 - 4.95",
  "4.95 - 5.70", "5.70 - 6.45", "6.45 - 7.20", "7.20 - 7.95",
  "7.95 - 8.70", "8.70 +")

map.TH <- ggplot(cat, aes(long, lat)) +
  geom_raster(aes(x, y, fill = THCUTV), data = dem.gamma) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.4,
    data = pt21, shape = 21) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.4,
    data = pt33, shape = 21) +
  geom_polygon(size = .5, linetype = "dashed", color = "black",
    fill = "grey40", alpha = 0) +
  geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 3,
    data = prof21, shape = 18) +
  geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 3,
    data = prof33, shape = 18) +
  coord_equal() +
  theme_bw(base_size = 9, base_family = "Helvetica") +
  scale_fill_manual(
    name = "Thorium\nin ppm", values = rev(brewer.pal(9,"Spectral"))) +
  labs(x = "\nUTM-E/m", y = "UTM-N/m\n") +
  scale_y_continuous(breaks = c(4362600,4362800,4363000)) +
  scale_x_continuous(breaks = c(508800,509000,509200)) +
  theme(axis.text.y = element_text(angle = 90, hjust = .5),
    legend.text = element_text(size = 11), legend.key.size = unit(18, "pt"),
    legend.title = element_text(size = 11),
    axis.text = element_text(size = 11), axis.title = element_text(size = 11))


gammaK.min <- min(dem.gamma$GAMMAK, na.rm = TRUE); gammaK.min # 0.6
gammaK.max <- max(dem.gamma$GAMMAK, na.rm = TRUE); gammaK.max # 2.5

dem.gamma$KCUTV <- cut(dem.gamma$GAMMAK, breaks = c(seq(.6, 2.2, by = .2), 2.5))
levels(dem.gamma$KCUTV) <- c("0.6 - 0.8", "0.8 - 1.0", "1.0 - 1.2",
  "1.2 - 1.4", "1.4 - 1.6", "1.6 - 1.8", "1.8 - 2.0",
  "2.0 - 2.2", "2.2 +")

map.K <- ggplot(cat, aes(long, lat)) +
  geom_raster(aes(x, y, fill = KCUTV), data = dem.gamma) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.4,
```

```
    data = pt21, shape = 21) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.4,
    data = pt33, shape = 21) +
  geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 3,
    data = prof21, shape = 18) +
  geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 3,
    data = prof33, shape = 18) +
  geom_polygon(size = .5, linetype = "dashed", color = "black",
    fill = "grey40", alpha = 0) +
  coord_equal() +
  theme_bw(base_size = 9, base_family = "Helvetica") +
  scale_fill_manual(
    name = "Potassium\nin %", values = rev(brewer.pal(9,"Spectral"))) +
  labs(x = "\nUTM-E/m", y = "UTM-N/m\n") +
  scale_y_continuous(breaks = c(4362600,4362800,4363000)) +
  scale_x_continuous(breaks = c(508800,509000,509200)) +
  theme(axis.text.y = element_text(angle = 90, hjust = .5),
    legend.text = element_text(size = 11), legend.key.size = unit(18, "pt"),
    legend.title = element_text(size = 11),
    axis.text = element_text(size = 11), axis.title = element_text(size = 11))


gammaU.min <- min(dem.gamma$GAMMAU, na.rm = TRUE); gammaU.min # 0.9
gammaU.max <- max(dem.gamma$GAMMAU, na.rm = TRUE); gammaU.max # 2.8

dem.gamma$UCUTV <- cut(dem.gamma$GAMMAU, breaks = c(seq(.9, 2.5, by = .2), 2.8))
levels(dem.gamma$UCUTV) <- c("0.9 - 1.1", "1.1 - 1.3", "1.3 - 1.5",
  "1.5 - 1.7", "1.7 - 1.9", "1.9 - 2.1", "2.1 - 2.3", "2.3 - 2.5", "2.5 +")

map.U <- ggplot(cat, aes(long, lat)) +
  geom_raster(aes(x, y, fill = UCUTV), data = dem.gamma) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.4,
    data = pt21, shape = 21) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.4,
    data = pt33, shape = 21) +
  geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 3,
    data = prof21, shape = 18) +
  geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 3,
    data = prof33, shape = 18) +
  geom_polygon(size = .5, linetype = "dashed", color = "black",
    fill = "grey40", alpha = 0) +
  coord_equal() +
  theme_bw(base_size = 9, base_family = "Helvetica") +
  scale_fill_manual(
    name = "Uranium\nin ppm", values = rev(brewer.pal(9,"Spectral"))) +
  labs(x = "\nUTM-E/m", y = "UTM-N/m\n") +
  scale_y_continuous(breaks = c(4362600,4362800,4363000)) +
  scale_x_continuous(breaks = c(508800,509000,509200)) +
  theme(axis.text.y = element_text(angle = 90, hjust = .5),
    legend.text = element_text(size = 11), legend.key.size = unit(18, "pt"),
```

```
    legend.title = element_text(size = 11),
    axis.text = element_text(size = 11), axis.title = element_text(size = 11))


gammaDR.min <- min(dem.gamma$GAMMADR, na.rm = TRUE); gammaDR.min # 20
gammaDR.max <- max(dem.gamma$GAMMADR, na.rm = TRUE); gammaDR.max # 71

dem.gamma$DRCUTV <- cut(dem.gamma$GAMMADR, breaks = seq(20, 74, by = 6))
levels(dem.gamma$DRCUTV) <- c("20 - 26", "26 - 32", "32 - 38",
  "38 - 44", "44 - 50", "50 - 56", "56 - 62", "62 - 68", "68 +")

map.DR <- ggplot(cat, aes(long, lat)) +
  geom_raster(aes(x, y, fill = DRCUTV), data = dem.gamma) +
  #geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = .35,
  #  data = dgo21, shape = 19) +
  #geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = .35,
  #  data = dgo33, shape = 19) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.4,
    data = pt21, shape = 21) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.4,
    data = pt33, shape = 21) +
  geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 3,
    data = prof21, shape = 18) +
  geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 3,
    data = prof33, shape = 18) +
  geom_polygon(size = .5, linetype = "dashed", color = "black",
    fill = "grey40", alpha = 0) +
  coord_equal() +
  theme_bw(base_size = 9, base_family = "Helvetica") +
  scale_fill_manual(
    name = "Dose rate\nin nG/h", values = rev(brewer.pal(9,"Spectral"))) +
  labs(x = "\nUTM-E/m", y = "UTM-N/m\n") +
  scale_y_continuous(breaks = c(4362600,4362800,4363000)) +
  scale_x_continuous(breaks = c(508800,509000,509200)) +
  theme(axis.text.y = element_text(angle = 90, hjust = .5),
    legend.text = element_text(size = 11), legend.key.size = unit(18, "pt"),
    legend.title = element_text(size = 11),
    axis.text = element_text(size = 11), axis.title = element_text(size = 11))


v.interest <- c("THKratio", "UKratio", "THUratio", "ELEV", "x", "y")
dem.gamma <- dem[,v.interest]

gammaELEV.min <- min(dem.gamma$ELEV, na.rm = TRUE); gammaELEV.min # 115
gammaELEV.max <- max(dem.gamma$ELEV, na.rm = TRUE); gammaELEV.max # 140

dem.gamma$ELEVCUTV <- cut(dem.gamma$ELEV, breaks = seq(115, 142, by = 3))
levels(dem.gamma$ELEVCUTV) <- c("115 - 118", "118 - 121", "121 - 124",
  "124 - 127", "127 - 130", "130 - 133", "133 - 136", "136 - 139", "139 +")

```

```
572 map.ELEV <- ggplot(cat, aes(long, lat)) +
573   geom_raster(aes(x, y, fill = ELEVCUTV), data = dem.gamma) +
574   geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.5,
575     data = pt21, shape = 19) +
576   geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.5,
577     data = pt33, shape = 19) +
578   geom_polygon(size = .5, linetype = "dashed", color = "black",
579     fill = "grey40", alpha = 0) +
580   coord_equal() +
581   theme_bw(base_size = 9, base_family = "Helvetica") +
582   scale_fill_manual(#guide = guide_legend(reverse = TRUE),
583     name = "Elevation in m", values = rev(brewer.pal(9,"Spectral"))) +
584   labs(x = "\nUTM-E/m", y = "UTM-N/m\n") +
585   scale_y_continuous(breaks = c(4362600,4362800,4363000)) +
586   scale_x_continuous(breaks = c(508800,509000,509200)) +
587   theme(axis.text.y = element_text(angle = 90, hjust = .5),
588     legend.text = element_text(size = 14), legend.key.size = unit(21, "pt"),
589     legend.title = element_text(size = 14),
590     axis.text = element_text(size = 14), axis.title = element_text(size = 14))
591
592 ggsave(paste(path.fig, "azienda_ELEV_v3.pdf", sep = "/"), map.ELEV,
593   width = 8.27, height = 7.57)
594 print(map.ELEV)
595 dev.off()
596
597 # prepare 4 gamma-plots for coupled export (vertical and legend harmonized):
598 gTH <- ggplot_gtable(ggplot_build(map.TH))
599 gK <- ggplot_gtable(ggplot_build(map.K))
600 gU <- ggplot_gtable(ggplot_build(map.U))
601 gDR <- ggplot_gtable(ggplot_build(map.DR))
602
603 legTH <- with(gTH$grobs[[8]], grobs[[1]]$widths[[4]])
604 legK <- with(gK$grobs[[8]], grobs[[1]]$widths[[4]])
605 legU <- with(gU$grobs[[8]], grobs[[1]]$widths[[4]])
606 legDR <- with(gDR$grobs[[8]], grobs[[1]]$widths[[4]])
607
608 gTH$widths; gK$widths; gU$widths; gDR$widths # --> max. legend width: Th
609
610 # set the widths to max. (= Th):
611 gK$widths <- gTH$widths; gU$widths <- gTH$widths; gDR$widths <- gTH$widths
612
613 # add an empty column of "abs(diff(widths)) mm"
614 #   width on the right of legend box for the smaller legend box:
615 gK$grobs[[8]] <-
616   gtable_add_cols(gK$grobs[[8]], unit(abs(diff(c(legTH, legK))), "mm"))
617 gU$grobs[[8]] <-
618   gtable_add_cols(gU$grobs[[8]], unit(abs(diff(c(legTH, legU))), "mm"))
619 gDR$grobs[[8]] <-
620   gtable_add_cols(gDR$grobs[[8]], unit(abs(diff(c(legTH, legDR))), "mm"))
621
```

```
gg <- arrangeGrob(gTH, gK, gU, gDR, nrow = 2)

# clone ggsave and bypass the class check:
ggsave <- ggplot2::ggsave; body(ggsave) <- body(ggplot2::ggsave)[-2]

ggsave(paste(path.fig, "azienda_gamma_interpolOK_v4.pdf", sep = "/"), gg,
  width = 10.79, height = 8.69)
dev.off()

rm(emi.max,emi.min,map,pts.21,pts.33,v.interest,pt21,pt33,demm,dem33.df,
  dem21.df,dem.emiH,dem.emiV,dem,cat,dem21.ov,dem33.ov,j,j.g)

rm(demio.21,demio.33,dgo.21,dgo.33,asdf,catchm,d21.save,d33.save,demt.save,
  dem21.save,dem33.save)

save.image(paste(getwd(), "phd_calc_input/az_basis.RData", sep = "/"))

# end of script, azienda_geophysics2k5_vDiss.R, 23.09.2013
```

.4.4 azienda_eda_vDiss.R

```
#
################################################################################
#### EDA of soil textural fractions at field 21 and 33, Azienda San Michele ####
################################################################################
#
## last update: 30.11.2014

# numerical and graphical summaries (of target variables)

# author: Michael Blaschek (blaschek@geographie.uni-kiel.de)

rm(list = ls())

library(xtable); library(lattice); library(RColorBrewer); library(rgdal)
# R v3.0.2, xtable_1.7-3, lattice_0.20-24, RColorBrewer_1.0-5, rgdal_0.8-15

load(paste(getwd(), "phd_calc_input/az_basis.RData", sep = "/"))
path.fig <- paste(getwd(), "phd_calc_figures", sep = "/")

n <- c(dim(d.21)[1], dim(d.33)[1]) # 43, 64


#-------------------------------------------------------------------------------
## Summary statistics
#-------------------------------------------------------------------------------

source("descr_statistics.R")

```

```
datasets <- list(d.21@data, d.33@data)
z <- c("CLAY", "SILT", "SAND", "CLAYalr", "SILTalr")

descr.d <- summary_stats(data = datasets, vars = z, export.latex = TRUE,
  path.to = c(paste(getwd(), "phd_calc_out", paste0(
    "az_descr", n[1], "_v1", ".tex"), sep = "/"),
    paste(getwd(), "phd_calc_out", paste0(
      "az_descr", n[2], "_v1", ".tex"), sep = "/")))

# Shapiro-Wilk normality test:
j <- c("W", "p"); shap.d <- NULL
for (q in 1:length(datasets)) {
 dd <- datasets[[q]]; k <- 1
 shap <- data.frame(z, 1:length(z), 1:length(z))
 names(shap) <- c("Target variables", j) # variable names
 for (i in z) {
  shap[k,"W"] <- round(shapiro.test(dd[,i])$statistic, 2)
  shap[k,"p"] <- round(shapiro.test(dd[,i])$p.value, 3)
  k <- k + 1
 }
 shap.d[[q]] <- shap
}
rm(i,j,k,q,dd,shap,datasets) # new objects from this section: descr.d, shap.d


#-------------------------------------------------------------------------------
## Check the representativity of d.21 and d.33 for each other
#-------------------------------------------------------------------------------

x <- as.matrix(d.21@data[c("CLAY","SILT","SAND")])
y <- as.matrix(d.33@data[c("CLAY","SILT","SAND")])
mx <- apply(x, 2, mean); my <- apply(y, 2, mean)
sx <- cov(x); sy <- cov(y); p <- dim(x)[2]
s.pooled <- ((n[1] - 1) * sx + (n[2] - 1) * sy) / (n[1] - 1 + n[2] - 1)

# two-sample Hotelling's T-squared test for differences in 2 multivariate means:
D2 <- t(mx - my) %*% solve(s.pooled) %*% (mx - my) * n[1] * n[2]/(n[1] + n[2])
# transforming Hotelling's T-square statistic into an F-statistic:
m <- (n[1] + n[2] - dim(x)[2] - 1) / (dim(x)[2] * (n[1] + n[2] - 2)); m
hF <- m * D2; hF
# calculating p-value for the given F-statistic and degrees of freedom:
p.value <- pf(hF, dim(x)[2], n[1] + n[2] - dim(x)[2] - 1, lower.tail = FALSE)

# h0:   1 =  2 , F-statistic = 15.331, p.value (X > hF) = 2.56e-08

# Bartlett's test of homogeneity of variance-covariance matrices:
bcf <- 1 + (2*p^2 + 3*p - 1)/(6*(p + 1)) *
  ((1/(n[1] - 1)) + (1/(n[2] - 1)) - (1/(n[1] + n[2] - 2))); bcf
bL <- (1/bcf) * ((n[1] + n[2] - 2) * log(det(s.pooled)) -
  (n[1] - 1) * log(det(sx)) - (n[2] - 1) * log(det(sy))); bL
```

```
79
80 p.value2 <- pchisq(bL, p * (p + 1)/2, lower.tail = FALSE); p.value2
81 # h0: Sigma1 = Sigma2, test-statistic = 9.925, p.value = 0.1278
82
83 rm(D2,s.pooled,sx,sy,x,y,bcf,m,mx,my,p)
84
85
86 #-------------------------------------------------------------------------------
87 ## Exploratory graphics
88 #-------------------------------------------------------------------------------
89
90 # density histograms, untransformed contents of fractions:
91 # class width based on Scott's rule (1979)
92 pdf(paste(path.fig, "azienda_histogr_soilsep_v1.pdf", sep = "/"),
93   width = 11.69, height = 8.27, pointsize = 18)
94 par(mfrow = c(2,3), las = 1, cex.main = 1, font.main = 2,
95   mar = c(4,3,2,0)+.1, oma = c(.5,.5,.5,.5), mgp = c(2,1,0))
96
97 # CLAY for field 21 (n = 43), Azienda San Michele:
98 hk <- 3.49 * sd(d.21$CLAY) * n[2]^(-1/3) # hk = 5.740
99 hd <- hist(d.21$CLAY, freq = F, col = "grey", xlim = c(0,55), ylim = c(0,0.1),
100   breaks = seq(18,54,6), xlab = "Clay content in %", axes = FALSE, ylab = "",
101   main = "a) Clay, N = 43", sub = paste("Shapiro-Wilk test: W = ",
102     shap.d[[2]][1,2], ", p = ", shap.d[[2]][1,3], sep = ""))
103 curve(dnorm(x, mean(d.21$CLAY), sd(d.21$CLAY)), add = T, lwd = 2, lty = 2)
104 axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
105 axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
106 axis(side = 1, lwd = 0, line = -.35); axis(side = 2, lwd = 0, line = -.35)
107 box(which = "plot", lty = "solid", col = "black")
108
109 # SILT for field 21 (n = 43), Azienda San Michele:
110 hk <- 3.49 * sd(d.21$SILT) * n[2]^(-1/3) # hk = 3.871
111 hd <- hist(d.21$SILT, freq = F, col = "grey", xlim = c(0,40), ylim = c(0,0.1),
112   breaks = seq(16,36,4), xlab = "Silt content in %", axes = FALSE, ylab = "",
113   main = "b) Silt, N = 43", sub = paste("Shapiro-Wilk test: W = ",
114     shap.d[[2]][2,2], ", p = ", shap.d[[2]][2,3], sep = ""))
115 curve(dnorm(x, mean(d.21$SILT), sd(d.21$SILT)), add = T, lwd = 2, lty = 2)
116 axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
117 axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
118 axis(side = 1, lwd = 0, line = -.35); axis(side = 2, lwd = 0, line = -.35)
119 box(which = "plot", lty = "solid", col = "black")
120
121 # SAND for field 21 (n = 43), Azienda San Michele:
122 hk <- 3.49 * sd(d.21$SAND) * n[2]^(-1/3) # hk = 7.230
123 hd <- hist(d.21$SAND, freq = F, col = "grey", xlim = c(0,60), ylim = c(0,0.1),
124   breaks = seq(28,56,7), xlab = "Sand content in %", axes = F, ylab = "",
125   main = "c) Sand, N = 43", sub = paste("Shapiro-Wilk test: W = ",
126     shap.d[[2]][3,2], ", p = ", shap.d[[2]][3,3], sep = ""))
127 curve(dnorm(x, mean(d.21$SAND), sd(d.21$SAND)), add = T, lwd = 2, lty = 2)
128 axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
```

```
axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 1, lwd = 0, line = -.35); axis(side = 2, lwd = 0, line = -.35)
box(which = "plot", lty = "solid", col = "black")

# CLAY for field 33 (n = 64), Azienda San Michele:
hk <- 3.49 * sd(d.33$CLAY) * n[3]^(-1/3) # hk = 5.661
hd <- hist(d.33$CLAY, freq = F, col = "grey", xlim = c(0,55), ylim = c(0,0.1),
  breaks = seq(18,54,6), xlab = "Clay content in %", axes = FALSE, ylab = "",
  main = "d) Clay, N = 64", sub = paste("Shapiro-Wilk test: W = ",
    shap.d[[3]][1,2], ", p = ", shap.d[[3]][1,3], sep = ""))
curve(dnorm(x, mean(d.33$CLAY), sd(d.33$CLAY)), add = T, lwd = 2, lty = 2)
axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 1, lwd = 0, line = -.35); axis(side = 2, lwd = 0, line = -.35)
box(which = "plot", lty = "solid", col = "black")

# SILT for field 33 (n = 64), Azienda San Michele:
hk <- 3.49 * sd(d.33$SILT) * n[3]^(-1/3); hk # hk = 4.283
hd <- hist(d.33$SILT, freq = F, col = "grey", xlim = c(0,40), ylim = c(0,0.1),
  breaks = seq(8,36,4), xlab = "Silt content in %", axes = FALSE, ylab = "",
  main = "e) Silt, N = 64", sub = paste("Shapiro-Wilk test: W = ",
    shap.d[[3]][2,2], ", p = ", shap.d[[3]][2,3], sep = ""))
curve(dnorm(x, mean(d.33$SILT), sd(d.33$SILT)), add = T, lwd = 2, lty = 2)
axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 1, lwd = 0, line = -.35); axis(side = 2, lwd = 0, line = -.35)
box(which = "plot", lty = "solid", col = "black")

# SAND for field 33 (n = 64), Azienda San Michele:
hk <- 3.49 * sd(d.33$SAND) * n[3]^(-1/3) # hk = 5.779
hd <- hist(d.33$SAND, freq = F, col = "grey", xlim = c(0,60), ylim = c(0,0.1),
  breaks = seq(24,60,6), xlab = "Sand content in %", axes = F, ylab = "",
  main = "f) Sand, N = 64", sub = paste("Shapiro-Wilk test: W = ",
    shap.d[[3]][3,2], ", p = ", shap.d[[3]][3,3], sep = ""))
curve(dnorm(x, mean(d.33$SAND), sd(d.33$SAND)), add = T, lwd = 2, lty = 2)
axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 1, lwd = 0, line = -.35); axis(side = 2, lwd = 0, line = -.35)
box(which = "plot", lty = "solid", col = "black")
dev.off()

rm(hk,hd)


# box-and-whisker plot, Azienda San Michele:
# build upon code published under CC BY-SA by
# Tim Appelhans: Creating publication quality graphs in R
# http://teachpress.environmentalinformatics-marburg.de/2013/07/
#   creating-publication-quality-graphs-in-r-7/ (visited on 06/10/14)
d <- rbind(d.21,d.33)
```

```
d$FIELDS <- factor(d$FIELD, labels = c("Field 21", "Field 33"))
x <- d@data[c("FIELDS","CLAY","SILT","SAND")]
names(x) <- c("field", "text1", "text2", "text3")
t = reshape(x, direction = "long",  varying = 2:4, sep = "")
t$textcl[which(t$time == 1)] <- "Clay"; t$textcl[which(t$time == 2)] <- "Silt"
t$textcl[which(t$time == 3)] <- "Sand"

pdf(paste(path.fig, "azienda_boxplot_soilsep_v1.pdf", sep = "/"),
  width = 6.75, height = 4, pointsize = 14)
bpl <- bwplot(text ~ factor(textcl, levels = c("Clay","Silt","Sand")) |
    as.character(field), data = t, layout = c(2,1),
  main = "", xlab = "Soil textural fractions", ylab = "Content in %", asp = 1,
  ylim = c(0,100), coef = 1.5, par.strip.text = list(cex = 1),
  scales = list(y = list(at = c(0,20,40,60,80,100))))

th <- trellis.par.get()
th$box.dot$pch <- "|"
th$box.rectangle$col <- "black"; th$box.rectangle$lwd <- 2
th$box.rectangle$fill <- brewer.pal(3, "Dark2")
th$box.umbrella$lty <- 1; th$box.umbrella$col <- "black"
th$plot.symbol$col <- "grey40"; th$plot.symbol$pch <- "*"
th$plot.symbol$cex <- 2; th$strip.background$col <- "grey80"
th$par.xlab.text$cex <- 1; th$par.ylab.text$cex <- 1
th$fontsize$text <- 14; th$axis.text$cex <- 1
th$axis.components$left$tck <- .75; th$axis.components$right$tck <- .75
th$layout.widths$left.padding <- 0; th$layout.widths$right.padding <- 0
th$layout.heights$top.padding <- 0; th$layout.heights$bottom.padding <- 0

bpl.upd <- update(bpl, par.settings = th)
print(bpl.upd)
dev.off()
rm(t,x,th,bpl,bpl.upd,n,z)
# remaining objects: d, d.c, d.v, dem1, descr.d, shap.d, path.fig

# end of script, azienda_eda_vDiss.R, 11.09.2014
```

.4.5 azienda_ternaryd_vDiss.R

```
#
##############################################################################
####  Ternary plot of soil texture at field 21 and 33, Azienda San Michele  ####
##############################################################################
#
## last update: 30.11.2014

# plot textural soil classifications

# author: Michael Blaschek (blaschek@geographie.uni-kiel.de)

```

```
12 rm(list = ls())
13
14 library(soiltexture); library(compositions)
15 # R v3.0.2, soiltexture_1.2.13, compositions_1.40-1
16
17 load(paste(getwd(), "phd_calc_input/az_basis.RData", sep = "/"))
18 path.fig <- paste(getwd(), "phd_calc_figures", sep = "/")
19
20 n <- c(dim(d.21)[1], dim(d.33)[1])
21
22
23 #------------------------------------------------------------------------------
24 ## Compare own measured values with previous studies at field 21
25 #------------------------------------------------------------------------------
26
27 # from Massimo Melis, AGRIS, 2005:
28 cly5a <- c(30,24,30,34,26,40,38,38,33); slt5a <- c(18,16,18,15,38,24,21,20,21)
29 snd5a <- c(53,61,52,52,37,36,42,42,46)
30
31 agris21.x <- c(508959,508824,508900,508826,508852,508821,508767,508824,508695)
32 agris21.y <- c(
33   4362593,4362468,4362598,4362520,4362615,4362663,4362602,4362736,4362606)
34
35 text.agris.21 <- data.frame(agris21.x,agris21.y,cly5a,slt5a,snd5a)
36 coordinates(text.agris.21) <- ~ agris21.x + agris21.y
37 proj4string(text.agris.21) <-
38   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
39
40 # from Ulrike Werban, UFZ, 2010:
41 text.ufz <- read.csv2(
42   "/home/mibla/Dokumente/phd_14/phd_calc/phd_calc_input/feld21_ufz.csv")
43 text.ufz.A <- subset(text.ufz, text.ufz$HORIZON == "A")
44
45 coordinates(text.ufz.A) <- ~ EAST + NORTH
46 proj4string(text.ufz.A) <-
47   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
48 proj4string(dem.21) <-
49   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
50
51 # take only those UFZ-measurements that are located at field 21:
52 dem21ufz.ov <- over(text.ufz.A, dem.21)
53 text.ufz.A@data[,"EMIH"] <- dem21ufz.ov[,"EMIH"]
54 text.ufz.A21 <- text.ufz.A[!is.na(text.ufz.A$EMIH),]
55
56 d.21.comp <- acomp(d.21@data, parts = c("CLAY", "SILT", "SAND"), total = 100)
57 d.33.comp <- acomp(d.33@data, parts = c("CLAY", "SILT", "SAND"), total = 100)
58 ufz21.comp <-
59   acomp(text.ufz.A21@data, parts = c("CLAY", "SILT", "SAND"), total = 100)
60 agris21.comp <-
61   acomp(text.agris.21@data, parts = c("cly5a", "slt5a", "snd5a"), total = 100)
```

```
62
63 #writeOGR(text.ufz.A21[1:10], driver = "ESRI Shapefile",
64 #  dsn = paste(getwd(), "phd_calc_out", sep = "/"), layer = "f21_samplingUFZ10")
65
66 #pnts.ufz = list(
67 #  "sp.points", text.ufz.A21, pch = 19, col = "black", cex = .6, alpha = 1)
68 #pnts.agris = list(
69 #  "sp.points", text.agris.21, pch = 19, col = "red", cex = .6, alpha = 1)
70 #spplot(dem.21["EMIH"],
71 #  scales = list(draw = T), sp.layout = list(pnts.ufz, pnts.agris))
72
73
74 #-------------------------------------------------------------------------
75 ## Ternary plot based on USDA/FAO classification
76 #-------------------------------------------------------------------------
77
78 # calculate compositional means:
79 st.cm.21 <- mean(d.21.comp)
80 st.cm.21 <- as.data.frame(rbind(st.cm.21[1:3] * 100))
81 st.cm.33 <- mean(d.33.comp)
82 st.cm.33 <- as.data.frame(rbind(st.cm.33[1:3] * 100))
83
84 names(st.cm.21) <- c("CLAY", "SILT", "SAND")
85 names(st.cm.33) <- c("CLAY", "SILT", "SAND")
86
87 st.cm.ufz <- mean(ufz21.comp)
88 st.cm.ufz <- as.data.frame(rbind(st.cm.ufz[1:3] * 100))
89 st.cm.agris <- mean(agris21.comp)
90 st.cm.agris <- as.data.frame(rbind(st.cm.agris[1:3] * 100))
91
92 names(st.cm.ufz) <- c("CLAY", "SILT", "SAND")
93 names(st.cm.agris) <- c("CLAY", "SILT", "SAND")
94
95 st.cm.agris; st.cm.ufz; st.cm.21
96 # 32.6, 20.7, 46.7 = Melis 2005 = sandy clay loam, n = 9
97 # 27.9, 29.3, 42.8 = UFZ 2010 = clay loam, n = 5
98 # 33.4, 26.1, 40.5 = CAU/LMU = clay loam, n = 43
99
100 # using the 2000-63-2 system for particle-size fractions
101 # FAO 06: Guidelines for soil description
102 pdf(paste(path.fig, "azienda_ternary_soilsep_v1.pdf", sep = "/"),
103   width = 8.5, height = 9.25, pointsize = 14)
104 stp <- TT.plot(class.sys = "USDA.TT", tri.data = d.33@data,
105   pch = "*", col = "blue", cex = 1.3, cex.axis = 1, cex.lab = 1, main = "",
106   class.lab.show = "abr", class.p.bg.col = F, class.line.col = "gray50",
107   new.mar = c(2,1,0,0)+.1, grid.show = F, frame.bg.col = "white",
108   lwd.lab = 1.2, lwd.axis = 1.2, font.lab = 1, font.axis = 1,
109   css.lab = c("% Clay 0-2   m ", "% Silt 2-63   m ", "% Sand 63-2000   m "))
110 stp.21 = TT.points(
111   tri.data = d.21@data, geo = stp, pch = "*", cex = 1.3, col = "red")
```

```
stp.m.21 = TT.points(
  tri.data = st.cm.21, geo = stp, pch = 18, cex = 1.3, col = "red")
stp.m.33 = TT.points(
  tri.data = st.cm.33, geo = stp, pch = 18, cex = 1.3, col = "blue")
# --> on average: clay loam (equals German: Lts = sandig-toniger Lehm)
dev.off()

# extract soil texture class for each sample location:
stc.21 <- TT.points.in.classes(d.21@data, class.sys = "USDA.TT", PiC.type = "n")
stcl.21 <- 0
for (i in 1:n[1])
  stcl.21[i] <- attributes(stc.21)$dimnames[[2]][which(stc.21[i,] == 1)]

d.21$USDATEXTCL <- stcl.21
table(d.21$USDATEXTCL) # --> 25 points inside ClLo = clay loam class

stc.33 <- TT.points.in.classes(d.33@data, class.sys = "USDA.TT", PiC.type = "n")
stcl.33 <- 0
for (i in 1:n[2])
  stcl.33[i] <- attributes(stc.33)$dimnames[[2]][which(stc.33[i,] == 1)]

d.33$USDATEXTCL <- stcl.33
table(d.33$USDATEXTCL) # --> 33 points inside Cl = clay class

rm(dem21ufz.ov,stc.21,stc.33,st.cm.21,st.cm.33,st.cm.agris,st.cm.ufz,stp.m.21,
  stp.m.33,stp.21,text.ufz,agris21.comp,agris21.x,agris21.y,cly5a,d.21.comp,
  d.33.comp,i,n,slt5a,snd5a,stcl.21,stcl.33,stp,text.agris.21,text.ufz.A,
  text.ufz.A21,ufz21.comp)

# end of script, costara_ternaryd_vDiss.R, 14.12.2011
```

.4.6 azienda_corPCA_vDiss.R

```
#
################################################################################
####   Correlation and principal component analysis, Azienda San Michele    ####
################################################################################
#
## last update: 30.11.2014

# scatterplot matrix of target variables and relief parameter
# principal component analysis

# author: Michael Blaschek (blaschek@geographie.uni-kiel.de)

rm(list = ls())

library(caret); library(plotrix); library(xtable)
# R v3.0.2, caret_6.0-22, plotrix_3.5-3, xtable_1.7-3
```

```
17
18 load(paste(getwd(), "phd_calc_input/az_basis.RData", sep = "/"))
19 path.fig <- paste(getwd(), "phd_calc_figures", sep = "/")
20
21 n <- c(dim(d.21)[1], dim(d.33)[1]) # 43, 64
22
23
24 #-------------------------------------------------------------------------------
25 ## Correlation coefficients and scatterplot-matrix
26 #-------------------------------------------------------------------------------
27
28 source("pearsons_cor_coeffs.R")
29
30 j <- c("EMIH", "EMIV", "GAMMAK", "GAMMATH", "GAMMAU", "THKratio",
31   "ELEV", "SLOPE", "SAGAWI")
32 z <- c("CLAY", "SILT", "SAND")
33
34 #cor21.ll <- pearsons_corr(data = d.21@data, covars = j, targets = z)
35 cor21.ll <- pearsons_corr(data = d.21@data, covars = j, targets = z,
36   scatterplot.matrix = TRUE, cex.lab = 1.2,
37   path.to = paste(path.fig, "field21_scatterpl_v99.pdf", sep = "/"))
38
39 #cor33.ll <- pearsons_corr(data = d.33@data, covars = j, targets = z)
40 cor33.ll <- pearsons_corr(data = d.33@data, covars = j, targets = z,
41   scatterplot.matrix = TRUE, cex.lab = 1.2,
42   path.to = paste(path.fig, "field33_scatterpl_v99.pdf", sep = "/"))
43
44 d33.c <- d.33[which(!is.na(d.33$CORG)),]
45 cor(d33.c$CaCO3, d33.c$GAMMADR) # -0.9
46 # --> CaCO3 negatively related to gamma-ray (most signif.: DR = -0.9)
47
48 cor(d.33$GRAVEL, d.33$GAMMADR) # 0.37
49 # --> GRAVEL positively related to gamma-ray (appr. 0.4)
50
51 cor(d33.c$CORG, d33.c$GAMMAU) # -0.36
52 # --> CORG negatively related to gamma-ray (most signif.: U = -0.36)
53
54
55 #-------------------------------------------------------------------------------
56 ## Principal component analysis (PCA)
57 #-------------------------------------------------------------------------------
58
59 # remove ASPECT and CONVG (too many NA-values):
60 j <- c("ELEV","SAGAWI","SLOPE","PROFC","PLANC","TWI","STREAMP","LS",
61   "EMIH","EMIV","GAMMAK","GAMMATH","GAMMAU","THKratio","THUratio","UKratio")
62
63 # field 33:
64 dat1 <- dem.33@data[,j]
65
66 # check that all covariates have the same number of NA-values:
```

```
for (i in 1:length(j)) {
  print(length(which(is.na(dat1[i]))))
}
dat1 <- dat1[which(!is.na(dat1$ELEV)),]

pc3 <- prcomp(dat1, scale. = T)

# selection of the right number of factors:
# critical value after Karlis, Saporta and Spinakis (2003):
p <- sum((pc3$sdev)^2) # the first 16 factors explain 100%
ni <- dim(dat1)[1] # number of instances
cv <- 1 + 1.65 * sqrt((p - 1)/(ni - 1)) # 1.03

nf <- which((pc3$sdev)^2 >= cv); nf # --> first six factors are significant

# convert the significant PCs to grids:
pc.comps <- as.data.frame(pc3$x)

# insert grid index:
dem.33$nrs <- seq(1, length(dem.33@data[[1]]))
dem33.pnt <- as(dem.33["nrs"], "SpatialPointsDataFrame")

maskpoints <- as.numeric(attr(pc3$x, "dimnames")[[1]]) # mask NA grid nodes

# attach coordinates:
pc.comps$X <- dem33.pnt@coords[maskpoints, 1]
pc.comps$Y <- dem33.pnt@coords[maskpoints, 2]
coordinates(pc.comps) <- ~ X + Y

# overlay with existing dem.33:
proj4string(d.33) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
proj4string(dem.33) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
proj4string(pc.comps) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")

dem33.ov <- over(dem.33, pc.comps)
dem.33@data$PC1 <- dem33.ov$PC1; dem.33@data$PC2 <- dem33.ov$PC2
dem.33@data$PC3 <- dem33.ov$PC3; dem.33@data$PC4 <- dem33.ov$PC4
dem.33@data$PC5 <- dem33.ov$PC5; dem.33@data$PC6 <- dem33.ov$PC6

# overlay with d.33:
dem33.ov <- over(d.33, dem.33)
d.33@data$PC1 <- dem33.ov$PC1; d.33@data$PC2 <- dem33.ov$PC2
d.33@data$PC3 <- dem33.ov$PC3; d.33@data$PC4 <- dem33.ov$PC4
d.33@data$PC5 <- dem33.ov$PC5; d.33@data$PC6 <- dem33.ov$PC6

# loadings:
pc3$sdev^2
```

```
pc3$rotation[,1:length(nf)]
aload <- abs(pc3$rotation[,1:length(nf)])
sweep(aload, 2, colSums(aload), "/") # --> relative contribution
# --> dim.1 influenced by SAGAWI, SLOPE, TWI, LS
# --> dim.2 influenced by THKratio, ELEV, GAMMAK, GAMMAU
# --> dim.3 influenced by UKratio, THUratio
# --> dim.4 influenced by EMIH, EMIV
# --> dim.5 influenced by EMIH, EMIV, THUratio, UKratio
# --> dim.6 influenced by PROFC, PLANC, THKratio

# export loading coefficients:
pcs <- (pc3$sdev * sqrt(max(1, nrow(dat1) - 1)))^2
ss.total <-  sum((pc3$sdev * sqrt(max(1, nrow(dat1) - 1)))^2)
pcs <- pcs/ss.total * 100 # --> explained variance by each PC
# only for those PCs that cumulatively exceed half of the total variance:
loads.val <- round(pc3$rotation[,1:which(cumsum(pcs) >= 50)[1]], digits = 3)
loads.out <- xtable(loads.val, caption = "PC loadings", label = "tab:PCL33")
print.xtable(loads.out,
  file = paste(getwd(), "phd_calc_out", "f33_loadings_v1.tex", sep = "/"))

pca.33 <- pc3


# field 21:
dat1 <- dem.21@data[,j]

# check that all covariates have the same number of NA-values:
for (i in 1:length(j)) {
  print(length(which(is.na(dat1[i]))))
}
dat1 <- dat1[which(!is.na(dat1$ELEV)),]

pc3 <- prcomp(dat1, scale. = T)

# selection of the right number of factors:
# critical value after Karlis, Saporta and Spinakis (2003):
p <- sum((pc3$sdev)^2) # the first 16 factors explain 100%
ni <- dim(dat1)[1] # number of instances
cv <- 1 + 1.65 * sqrt((p - 1)/(ni - 1)) # 1.05

nf <- which((pc3$sdev)^2 >= cv); nf # --> first five factors are significant

# convert the significant PCs to grids:
pc.comps <- as.data.frame(pc3$x)

# insert grid index:
dem.21$nrs <- seq(1, length(dem.21@data[[1]]))
dem21.pnt <- as(dem.21["nrs"], "SpatialPointsDataFrame")

maskpoints <- as.numeric(attr(pc3$x, "dimnames")[[1]]) # mask NA grid nodes
```

```
167
168 # attach coordinates:
169 pc.comps$X <- dem21.pnt@coords[maskpoints, 1]
170 pc.comps$Y <- dem21.pnt@coords[maskpoints, 2]
171 coordinates(pc.comps) <- ~ X + Y
172
173 # overlay with existing dem.21:
174 proj4string(d.21) <-
175   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
176 proj4string(dem.21) <-
177   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
178 proj4string(pc.comps) <-
179   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
180
181 dem21.ov <- over(dem.21, pc.comps)
182 dem.21@data$PC1 <- dem21.ov$PC1; dem.21@data$PC2 <- dem21.ov$PC2
183 dem.21@data$PC3 <- dem21.ov$PC3; dem.21@data$PC4 <- dem21.ov$PC4
184 dem.21@data$PC5 <- dem21.ov$PC5; dem.21@data$PC6 <- dem21.ov$PC6
185
186 # overlay with d.21:
187 dem21.ov <- over(d.21, dem.21)
188 d.21@data$PC1 <- dem21.ov$PC1; d.21@data$PC2 <- dem21.ov$PC2
189 d.21@data$PC3 <- dem21.ov$PC3; d.21@data$PC4 <- dem21.ov$PC4
190 d.21@data$PC5 <- dem21.ov$PC5; d.21@data$PC6 <- dem21.ov$PC6
191
192 # loadings:
193 pc3$sdev^2
194 pc3$rotation[,1:5]
195 aload <- abs(pc3$rotation[,1:5])
196 sweep(aload, 2, colSums(aload), "/") # --> relative contribution
197 # --> dim.1 influenced by GAMMAK, GAMMATH, GAMMAU, EMIV, EMIH
198 # --> dim.2 influenced by TWI, SAGAWI
199 # --> dim.3 influenced by LS, STREAMP, SLOPE, THUratio
200 # --> dim.4 influenced by THUratio, ELEV, UKratio
201 # --> dim.5 influenced by PROFC, PLANC
202
203 pcs <- (pc3$sdev * sqrt(max(1, nrow(dat1) - 1)))^2
204 ss.total <-  sum((pc3$sdev * sqrt(max(1, nrow(dat1) - 1)))^2)
205 pcs <- pcs/ss.total * 100 # --> explained variance by each PC
206 # only for those PCs that cumulatively exceed half of the total variance:
207 loads.val <- round(pc3$rotation[,1:which(cumsum(pcs) >= 50)[1]], digits = 3)
208 loads.out <- xtable(loads.val, caption = "PC loadings", label = "tab:PCL21")
209 print.xtable(loads.out,
210   file = paste(getwd(), "phd_calc_out", "f21_loadings_v1.tex", sep = "/"))
211
212 pca.21 <- pc3
213
214
215 dat1 <- dem.33@data[,j]
216 dat1 <- dat1[which(!is.na(dat1$ELEV)),]
```

```
m33.means <- apply(dat1, 2, mean)
m33.y <- sweep(dat1, 2, m33.means)
m33.y <- m33.y/sqrt(nrow(m33.y) * ncol(m33.y)) # scaling (standardization)

dat1 <- dem.21@data[,j]
dat1 <- dat1[which(!is.na(dat1$ELEV)),]
m21.means <- apply(dat1, 2, mean)
m21.y <- sweep(dat1, 2, m21.means)
m21.y <- m21.y/sqrt(nrow(m21.y) * ncol(m21.y)) # scaling (standardization)

# bi- and embedded screeplots:
pcs <- (pca.33$sdev * sqrt(max(1, nrow(dem.33@data[,j]) - 1)))^2
ss.total <-  sum((pca.33$sdev * sqrt(max(1, nrow(dem.33@data[,j]) - 1)))^2)
pcs.p33 <- pcs/ss.total * 100 # --> explained variance by each PC
npcs <- length(pca.33$sdev)
xp.33 <- seq_len(npcs)
plot(xp.33, pcs.p33, type = "b", axes = T,
  main = "Eigenvalues", xlab = "eigenvalue number", ylab = "variance [in %]")

pcs <- (pca.21$sdev * sqrt(max(1, nrow(dem.21@data[,j]) - 1)))^2
ss.total <-  sum((pca.21$sdev * sqrt(max(1, nrow(dem.21@data[,j]) - 1)))^2)
pcs.p21 <- pcs/ss.total * 100 # --> explained variance by each PC
npcs <- length(pca.21$sdev)
xp.21 <- seq_len(npcs)
plot(xp.21, pcs.p21, type = "b", axes = T,
  main = "Eigenvalues", xlab = "eigenvalue number", ylab = "variance [in %]")

pdf(paste(path.fig, "azienda_biplot_v33.pdf", sep = "/"))
par(mar = c(4,4,3,3) + 0.1)
biplot(pca.33, col = c("gray50", "red"), cex = c(.65,.85), scale = 1,
  xlim = c(-.0205, .0205), ylim = c(-.0205, .0205),
  xlabs = rep(".", times = length(pca.33$x[,1])),
  ylabs = rep("+", times = length(pca.33$sdev)))
#draw.circle(0, 0, 1, border = "blue", col = "transparent", lty = 2)
text(c(.54,1,1,1.15,1,1.125,1.2,1,1,1,1.1,1,1,1,1,1.37) *
  pca.33$rotation %*% diag(pca.33$sdev *
   sqrt(max(1, nrow(m33.y))))[,1],
    c(1,.85,1.13,.7,1.3,1,1.39,.81,.68,1.2,.92,1.14,1.09,1.07,1.13,1) *
     pca.33$rotation %*% diag(pca.33$sdev *
      sqrt(max(1, nrow(m33.y))))[,2],
       labels = colnames(dem.33@data[,j]), col = 2, cex = 1.15)
boxed.labels(-27.5, -117, bg = "white", border = NA, cex = 1,
  "centred, scaled to unit variance, partitioning factor = 1")
box()
par(fig = c(.66,.91,.66,.91), new = TRUE, mar = c(2,2,.5,.5), mgp = c(3,.25,0))
plot(xp.33, pcs.p33, axes = FALSE, ann = FALSE,
  type = "p", pch = 16, tcl = -.1, cex = .65, cex.axis = .5)
box(which = "plot")
points(xp.33[6], pcs.p33[6], type = "p", cex = .65, pch = 16, col = "red")
axis(side = 2, cex.axis = .5, padj = -.15, tcl = -.1)
```

```
267 axis(side = 1, cex.axis = .5, padj = -1, tcl = -.1)
268 mtext(side = 2, text = "expl. variance", line = 1, cex = .65)
269 mtext(side = 1, text = "eigenvalue number", line = .5, cex = .65)
270 dev.off()
271
272 pdf(paste(path.fig, "azienda_biplot_v21.pdf", sep = "/"))
273 par(mar = c(4,4,3,3) + 0.1)
274 biplot(pca.21, col = c("gray50", "red"), cex = c(.65,1.2), scale = 1,
275   xlabs = rep(".", times = length(pca.21$x[,1])),
276   ylabs = rep("+", times = length(pca.21$sdev)))
277 text(c(-12,1,1,2.75,1,.63,1,1.11,.98,.98,1,.785,1,1.38,1.63,1) *
278   pca.21$rotation %*% diag(pca.21$sdev *
279    sqrt(max(1, nrow(m21.y))))[,1],
280     c(1,.93,1.13,1.32,1.12,1,1.1,1,.75,1.2,1.21,1.06,.7,1,1.15,1.11) *
281      pca.21$rotation %*% diag(pca.21$sdev *
282       sqrt(max(1, nrow(m21.y))))[,2],
283        labels = colnames(dem.21@data[,j]), col = 2, cex = 1.15)
284 boxed.labels(-8, -68, bg = "white", border = NA, cex = 1,
285   "centred, scaled to unit variance, partitioning factor = 1")
286 box()
287 par(fig = c(.66,.91,.66,.91), new = TRUE, mar = c(2,2,.5,.5), mgp = c(3,.25,0))
288 plot(xp.21, pcs.p21, axes = FALSE, ann = FALSE,
289   type = "p", pch = 16, tcl = -.1, cex = .65, cex.axis = .5)
290 box(which = "plot")
291 points(xp.21[5], pcs.p21[5], type = "p", cex = .65, pch = 16, col = "red")
292 axis(side = 2, cex.axis = .5, padj = -.15, tcl = -.1)
293 axis(side = 1, cex.axis = .5, padj = -1, tcl = -.1)
294 mtext(side = 2, text = "expl. variance", line = 1, cex = .65)
295 mtext(side = 1, text = "eigenvalue number", line = .5, cex = .65)
296 dev.off()
297
298 rm(aload,dat1,dem21.ov,dem33.ov,m21.y,m33.y,cv,dem21.pnt,dem33.pnt,i,j,
299   m21.means,m33.means,maskpoints,nf,ni,npcs,p,pc.comps,pc3,pcs.21,pcs.33,
300   pcs.p21,pcs.p33,ss.total,xp.21,xp.33)
301
302 save.image(paste(getwd(), "phd_calc_input/az_basisPCs.RData", sep = "/"))
303
304 # end of script, azienda_corPCA_vDiss.R, 02.11.2014
```

.4.7 azienda_explor_sp_vDiss.R

```
1 #
2 ##############################################################################
3 #### ESDA of soil textural fractions at field 21 and 33, Azienda S. Michele ####
4 ##############################################################################
5 #
6 ## last update: 30.11.2014
7
8 # trend detection and variography (of target variables)
```

```
9
10 # author: Michael Blaschek (blaschek@geographie.uni-kiel.de)
11
12 rm(list = ls())
13
14 library(rgdal); library(RColorBrewer); library(sp); library(gstat)
15 # R v3.0.2, rgdal_0.8-15, RColorBrewer_1.0-5, sp_1.0-14, gstat_1.0-16
16
17 load(paste(getwd(), "phd_calc_input/az_basis.RData", sep = "/"))
18 path.fig <- paste(getwd(), "phd_calc_figures", sep = "/")
19
20 n <- c(dim(d.21)[1], dim(d.33)[1]) # 43, 64
21
22
23 #-------------------------------------------------------------------------------
24 ## Trend detection - Postplots
25 #-------------------------------------------------------------------------------
26
27 # build upon code published by
28 # Richard E. Plant: Spatial data analysis in ecology and agriculture using R
29 # Taylor & Francis Group, 2012
30
31 catchm <- readOGR(dsn = "phd_calc_input", layer = "fields2133_diss_v2")
32
33 d <- rbind(d.21,d.33)
34 soil.az <- readOGR(dsn = "phd_calc_input", layer = "soilmap_azienda_mb")
35
36 proj4string(soil.az) <- CRS(NA); proj4string(d.21) <- CRS(NA)
37 proj4string(d.33) <- CRS(NA); proj4string(d) <- CRS(NA)
38 proj4string(soil.az) <-
39   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
40 proj4string(d.21) <-
41   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
42 proj4string(d.33) <-
43   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
44 proj4string(d) <-
45   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
46
47 soil.az.ov <- over(d.21, soil.az); d.21$SOILUNIT <- soil.az.ov$soilunit
48 soil.az.ov <- over(d.33, soil.az); d.33$SOILUNIT <- soil.az.ov$soilunit
49 soil.az.ov <- over(d, soil.az); d$SOILUNIT <- soil.az.ov$soilunit
50
51 d$color <- "#1B9E77"
52 d$color[which(d$SOILUNIT == 6)] <- "#D95F02"
53 d$color[which(d$SOILUNIT == 7)] <- "#7570B3"
54 d$color[which(d$SOILUNIT == 8)] <- "#E7298A"
55 d$color[which(d$SOILUNIT == 9)] <- "#66A61E"
56
57 soil.units <- c("Brown soils on Miocenic marls", "Regosols on Miocenic marls",
58   "Soils on subrecent alluvial deposits ... with intact profiles",
```

```
  "... after moderate erosion", "... after intense erosion, with crusts")

pdf(paste(path.fig, "azienda_postpl_soilsep_v1.pdf", sep = "/"),
  width = 15, height = 14, pointsize = 24)
#layout(matrix(c(4,1,1,5,2,2,3,3), 2, 4, byrow = TRUE))
layout(matrix(c(1,1,2,2,3,3,4,4), 2, 4, byrow = TRUE))
# layout.show(3)
par(mar = c(3,3,2,0)+.1, oma = c(.5,.5,.5,.5), mgp = c(2,1,0),
  cex.main = 1, font.main = 2)

# clay, n = 107:
plot(catchm, bg = NA, lwd = 1, lty = 2, axes = FALSE, ylim = c(4362450,4363065))
abline(h = seq(4362500,4363100,250),
  v = seq(508750,509300,250), lty = 1, col = "grey80")
plot(d, pch = 20, col = d$color, asp = 1, axes = FALSE, add = TRUE,
  cex = d$CLAY * 3/max(d$CLAY))
axis(side = 1, tck = -.02, at = seq(508750,509300,250), labels = NA)
axis(side = 2, tck = -.02, at = seq(4362500,4363100,250), labels = NA)
axis(side = 1, lwd = 0, line = -.35, at = seq(508750,509300,250))
axis(side = 2, lwd = 0, line = -.35, at = seq(4362500,4363100,250))
title(main = "a) Clay content in %", xlab = "UTM-E/m", ylab = "UTM-N/m")
box(which = "plot", lty = "solid", col = "black")
SpatialPolygonsRescale(layout.scale.bar(height = 0.065), plot.grid = FALSE,
  offset = c(509200,4362480), scale = 100, fill = c("white", "black"))
text(509200,4362500, "0", cex = .75); text(509300,4362500, "100 m", cex = .75)

# silt, n = 197:
plot(catchm, bg = NA, lwd = 1, lty = 2, axes = FALSE, ylim = c(4362450,4363065))
abline(h = seq(4362500,4363100,250),
  v = seq(508750,509300,250), lty = 1, col = "grey80")
plot(d, pch = 20, col = d$color, asp = 1, axes = FALSE, add = TRUE,
  cex = d$SILT * 3/max(d$SILT))
axis(side = 1, tck = -.02, at = seq(508750,509300,250), labels = NA)
axis(side = 2, tck = -.02, at = seq(4362500,4363100,250), labels = NA)
axis(side = 1, lwd = 0, line = -.35, at = seq(508750,509300,250))
axis(side = 2, lwd = 0, line = -.35, at = seq(4362500,4363100,250))
title(main = "b) Silt content in %", xlab = "UTM-E/m")
box(which = "plot", lty = "solid", col = "black")
SpatialPolygonsRescale(layout.scale.bar(height = 0.065), plot.grid = FALSE,
  offset = c(509200,4362480), scale = 100, fill = c("white", "black"))
text(509200,4362500, "0", cex = .75); text(509300,4362500, "100 m", cex = .75)

# sand, n = 197:
plot(catchm, bg = NA, lwd = 1, lty = 2, axes = FALSE, ylim = c(4362450,4363065))
abline(h = seq(4362500,4363100,250),
  v = seq(508750,509300,250), lty = 1, col = "grey80")
plot(d, pch = 20, col = d$color, asp = 1, axes = FALSE, add = TRUE,
  cex = d$SAND * 3/max(d$SAND))
axis(side = 1, tck = -.02, at = seq(508750,509300,250), labels = NA)
axis(side = 2, tck = -.02, at = seq(4362500,4363100,250), labels = NA)
```

```
109 axis(side = 1, lwd = 0, line = -.35, at = seq(508750,509300,250))
110 axis(side = 2, lwd = 0, line = -.35, at = seq(4362500,4363100,250))
111 title(main = "c) Sand content in %", xlab = "UTM-E/m", ylab = "UTM-N/m")
112 box(which = "plot", lty = "solid", col = "black")
113 SpatialPolygonsRescale(layout.scale.bar(height = 0.065), plot.grid = FALSE,
114   offset = c(509200,4362480), scale = 100, fill = c("white", "black"))
115 text(509200,4362500, "0", cex = .75); text(509300,4362500, "100 m", cex = .75)
116
117 plot(0, 0, type = "n", bty = "n", axes = FALSE, xlab = "", ylab = "")
118 legend("center", soil.units, cex = 1,
119   fill = brewer.pal(5, "Dark2"), bg = "white", ncol = 1)
120 dev.off()
121 rm(soil.units,soil.az,soil.az.ov)
122
123
124 #------------------------------------------------------------------------------
125 ## Variography
126 #------------------------------------------------------------------------------
127
128 # field 21:
129 diag.bbox <- sqrt(
130   (bbox(d.21)[1,1] - bbox(d.21)[1,2])^2 + (bbox(d.21)[2,1] - bbox(d.21)[2,2])^2)
131 diag.bbox/2; diag.bbox/3
132 # 2 after Journel and Huijbregts 78, 3 = gstat's default
133 # --> distance of reliability between 127 and 191m
134
135 # using gstat (package) and methods-of-moments estimation:
136 # Clay:
137 va.cutoff <- 195; va.width <- 15
138 va.cutoff2 <- 135
139 cl.21.va <- variogram(
140   CLAY ~ 1, loc = d.21, cutoff = va.cutoff, width = va.width)
141 cl.21.va2 <- variogram(
142   CLAY ~ 1, loc = d.21, cutoff = va.cutoff2, width = va.width)
143 # initial variogram (s. Hengl09, p.130): nugget = measurement error,
144 #  sill = sampled variance, range = 1/4 of the diagonal of the bounding box
145 a.ini <- sqrt(diff(d.21@bbox["EAST",])^2 + diff(d.21@bbox["NORTH",])^2)/4
146 psill.ini <- var(d.21$CLAY); nug.ini <- 0
147 cl.21.vm <- vgm(psill.ini, "Sph", a.ini, nug.ini)
148 cl.21.vmf <- fit.variogram(cl.21.va2, cl.21.vm, fit.method = 7)
149 # Silt:
150 va.cutoff <- 195; va.width <- 15
151 si.21.va <- variogram(
152   SILT ~ 1, loc = d.21, cutoff = va.cutoff, width = va.width)
153 a.ini <- sqrt(diff(d.21@bbox["EAST",])^2 + diff(d.21@bbox["NORTH",])^2)/4
154 psill.ini <- var(d.21$SILT); nug.ini <- 0
155 si.21.vm <- vgm(psill.ini, "Sph", a.ini, nug.ini)
156 si.21.vmf <- fit.variogram(si.21.va, si.21.vm, fit.method = 6)
157 # Sand:
158 va.cutoff <- 195; va.width <- 15
```

```
va.cutoff2 <- 135
sa.21.va <- variogram(
  SAND ~ 1, loc = d.21, cutoff = va.cutoff, width = va.width)
sa.21.va2 <- variogram(
  SAND ~ 1, loc = d.21, cutoff = va.cutoff2, width = va.width)
a.ini <- sqrt(diff(d.21@bbox["EAST",])^2 + diff(d.21@bbox["NORTH",])^2)/4
psill.ini <- var(d.21$SAND); nug.ini <- 0
sa.21.vm <- vgm(psill.ini, "Exp", a.ini, nug.ini)
sa.21.vmf <- fit.variogram(sa.21.va2, sa.21.vm, fit.method = 0)

# field 33:
diag.bbox <- sqrt(
  (bbox(d.33)[1,1] - bbox(d.33)[1,2])^2 + (bbox(d.33)[2,1] - bbox(d.33)[2,2])^2)
diag.bbox/2; diag.bbox/3
# --> distance of reliability between 178 and 268m
# Clay:
va.cutoff <- 255; va.width <- 15
cl.33.va <- variogram(
  CLAY ~ 1, loc = d.33, cutoff = va.cutoff, width = va.width)
a.ini <- sqrt(diff(d.33@bbox["EAST",])^2 + diff(d.33@bbox["NORTH",])^2)/4
psill.ini <- var(d.33$CLAY); nug.ini <- 0
cl.33.vm <- vgm(psill.ini, "Sph", a.ini, nug.ini)
cl.33.vmf <- fit.variogram(cl.33.va, cl.33.vm, fit.method = 7)
# Silt:
va.cutoff <- 255; va.width <- 15
si.33.va <- variogram(
  SILT ~ 1, loc = d.33, cutoff = va.cutoff, width = va.width)
a.ini <- sqrt(diff(d.33@bbox["EAST",])^2 + diff(d.33@bbox["NORTH",])^2)/4
psill.ini <- var(d.33$SILT); nug.ini <- 0
si.33.vm <- vgm(psill.ini, "Sph", a.ini, nug.ini)
si.33.vmf <- fit.variogram(si.33.va, si.33.vm, fit.method = 7)
nsr <- si.33.vmf[1,2]/(si.33.vmf[1,2] + si.33.vmf[2,2]) * 100 # 62.55
# Sand:
va.cutoff <- 270; va.width <- 15
sa.33.va <- variogram(
  SAND ~ 1, loc = d.33, cutoff = va.cutoff, width = va.width)
a.ini <- 135; psill.ini <- 55; nug.ini <- 8
sa.33.vm <- vgm(psill.ini, "Exp", a.ini, nug.ini)
sa.33.vmf <- fit.variogram(sa.33.va, sa.33.vm, fit.method = 0)
plot(sa.33.va, model = sa.33.vmf)

pdf(paste(path.fig, "azienda_vario_soilsep_v2.pdf", sep = "/"),
  width = 8.27, height = 11.69, pointsize = 21)
par(mar = c(3,3,2,0)+.1, oma = c(.5,.5,.5,.5), mgp = c(2,1,0),
  cex.main = 1, font.main = 2, las = 1, mfrow = c(3,2))

plot(cl.21.va$gamma ~ cl.21.va$dist, pch = 20, cex = 1.25, col = "black",
  xlim = c(0, max(cl.21.va$dist)*1.05), ylim = c(0, max(cl.21.va$gamma)*1.1),
  axes = FALSE, ann = FALSE, xaxs = "i", yaxs = "i")
axis(side = 1, tck = -.02, labels = NA, at = seq(0,200,50))
```

```
209 axis(side = 2, tck = -.02, labels = NA, at = seq(0,60,20))
210 axis(side = 1, lwd = 0, line = -.35, at = seq(0,200,50),
211   labels = c("0.1", seq(50,200,50)))
212 axis(side = 2, lwd = 0, line = -.35, at = seq(0,60,20))
213 title(xlab = "Lag distance in m",
214   ylab = "Variance", main = "a) Clay, Field 21", sub = "")
215 lines(variogramLine(cl.21.vmf, maxdist = max(cl.21.va2$dist)),
216   col = "black", lwd = 1.25)
217 legend("bottomright", "Nug + Sph", bty = "n")
218 box(which = "plot", lty = "solid", col = "black")
219
220 plot(cl.33.va$gamma ~ cl.33.va$dist, pch = 20, cex = 1.25, col = "black",
221   xlim = c(0, max(cl.33.va$dist)*1.05), ylim = c(0, max(cl.21.va$gamma)*1.1),
222   axes = FALSE, ann = FALSE, xaxs = "i", yaxs = "i")
223 axis(side = 1, tck = -.02, labels = NA, at = seq(0,250,50))
224 axis(side = 2, tck = -.02, labels = NA, at = seq(0,60,20))
225 axis(side = 1, lwd = 0, line = -.35, at = seq(0,250,50),
226   labels = c("0.1", seq(50,250,50)))
227 axis(side = 2, lwd = 0, line = -.35, at = seq(0,60,20))
228 title(
229   xlab = "Lag distance in m", ylab = "", main = "b) Clay, Field 33", sub = "")
230 lines(variogramLine(cl.33.vmf, maxdist = max(cl.33.va$dist)),
231   col = "black", lwd = 1.25)
232 legend("bottomright", "Nug + Sph", bty = "n")
233 box(which = "plot", lty = "solid", col = "black")
234
235 plot(si.21.va$gamma ~ si.21.va$dist, pch = 20, cex = 1.25, col = "black",
236   xlim = c(0, max(si.21.va$dist)*1.05), ylim = c(0, max(si.33.va$gamma)*1.1),
237   axes = FALSE, ann = FALSE, xaxs = "i", yaxs = "i")
238 axis(side = 1, tck = -.02, labels = NA, at = seq(0,200,50))
239 axis(side = 2, tck = -.02, labels = NA, at = seq(0,30,5))
240 axis(side = 1, lwd = 0, line = -.35, at = seq(0,200,50),
241   labels = c("0.1", seq(50,200,50)))
242 axis(side = 2, lwd = 0, line = -.35, at = seq(0,30,5))
243 title(xlab = "Lag distance in m",
244   ylab = "Variance", main = "c) Silt, Field 21", sub = "")
245 lines(variogramLine(si.21.vmf, maxdist = max(si.21.va$dist)),
246   col = "black", lwd = 1.25)
247 legend("bottomright", "Nug + Sph", bty = "n")
248 box(which = "plot", lty = "solid", col = "black")
249
250 plot(si.33.va$gamma ~ si.33.va$dist, pch = 20, cex = 1.25, col = "black",
251   xlim = c(0, max(si.33.va$dist)*1.05), ylim = c(0, max(si.33.va$gamma)*1.1),
252   axes = FALSE, ann = FALSE, xaxs = "i", yaxs = "i")
253 axis(side = 1, tck = -.02, labels = NA, at = seq(0,250,50))
254 axis(side = 2, tck = -.02, labels = NA, at = seq(0,30,5))
255 axis(side = 1, lwd = 0, line = -.35, at = seq(0,250,50),
256   labels = c("0.1", seq(50,250,50)))
257 axis(side = 2, lwd = 0, line = -.35, at = seq(0,30,5))
258 title(
```

```
259   xlab = "Lag distance in m", ylab = "", main = "d) Silt, Field 33", sub = "")
260 lines(variogramLine(si.33.vmf, maxdist = max(si.33.va$dist)),
261   col = "black", lwd = 1.25)
262 legend("bottomright", "Nug + Sph", bty = "n")
263 box(which = "plot", lty = "solid", col = "black")
264
265 plot(sa.21.va$gamma ~ sa.21.va$dist, pch = 20, cex = 1.25, col = "black",
266   xlim = c(0, max(sa.21.va$dist)*1.05), ylim = c(0, max(sa.21.va$gamma)*1.1),
267   axes = FALSE, ann = FALSE, xaxs = "i", yaxs = "i")
268 axis(side = 1, tck = -.02, labels = NA, at = seq(0,200,50))
269 axis(side = 2, tck = -.02, labels = NA, at = seq(0,120,20))
270 axis(side = 1, lwd = 0, line = -.35, at = seq(0,200,50),
271   labels = c("0.1", seq(50,200,50)))
272 axis(side = 2, lwd = 0, line = -.35, at = seq(0,120,20))
273 title(xlab = "Lag distance in m",
274   ylab = "Variance", main = "e) Sand, Field 21", sub = "")
275 lines(variogramLine(sa.21.vmf, maxdist = max(sa.21.va2$dist)),
276   col = "black", lwd = 1.25)
277 legend("bottomright", "Exp", bty = "n")
278 box(which = "plot", lty = "solid", col = "black")
279
280 plot(sa.33.va$gamma ~ sa.33.va$dist, pch = 20, cex = 1.25, col = "black",
281   xlim = c(0, max(sa.33.va$dist)*1.05), ylim = c(0, max(sa.21.va$gamma)*1.1),
282   axes = FALSE, ann = FALSE, xaxs = "i", yaxs = "i")
283 axis(side = 1, tck = -.02, labels = NA, at = seq(0,250,50))
284 axis(side = 2, tck = -.02, labels = NA, at = seq(0,120,20))
285 axis(side = 1, lwd = 0, line = -.35, at = seq(0,250,50),
286   labels = c("0.1", seq(50,250,50)))
287 axis(side = 2, lwd = 0, line = -.35, at = seq(0,120,20))
288 title(
289   xlab = "Lag distance in m", ylab = "", main = "f) Sand, Field 33", sub = "")
290 lines(variogramLine(sa.33.vmf, maxdist = max(sa.33.va$dist)),
291   col = "black", lwd = 1.25)
292 legend("bottomright", "Exp", bty = "n")
293 box(which = "plot", lty = "solid", col = "black")
294 dev.off()
295
296 rm(h.max,cl.21.va,cl.21.va2,cl.21.vm,cl.21.vmf,cl.33.va,cl.33.vm,cl.33.vmf,
297   sa.21.va,sa.21.va2,sa.21.vm,sa.21.vmf,sa.33.va,sa.33.vm,sa.33.vmf,a.ini,
298   si.21.va,si.21.vm,si.21.vmf,si.33.va,si.33.vm,si.33.vmf,nug.ini,psill.ini,
299   diag.bbox,va.cutoff,va.cutoff2,va.width)
300
301 # end of script, azienda_explor_sp_vDiss.R, 11.06.2014
```

.4.8 azienda_rcok2k5_vDiss.R

```
1 #
2 ##############################################################################
3 #### Regression cokriging soil text. at field 21 and 33, Azienda S. Michele ####
```

```
4 ##############################################################################
5 #
6 ## last update: 09.04.2015
7
8 # regression modelling and residual cokriging
9 #   of alr-transformed soil separates at field scale
10
11 # author: Michael Blaschek (blaschek@geographie.uni-kiel.de)
12
13 rm(list = ls())
14
15 library(rgdal); library(gstat); library(RColorBrewer); library(compositions)
16 library(caret); library(ggplot2)
17 library(grid); library(gridExtra); library(gtable)
18 # R v3.0.2, gstat_1.0-16, rgdal_0.8-15, RColorBrewer_1.0-5, compositions_1.40-1
19 #  caret_6.0-22, ggplot2_0.9.3.1, gridExtra_0.9.1, gtable_0.1.2
20
21 load(paste(getwd(), "phd_calc_input/az_basisPCs.RData", sep = "/"))
22 #path.fig <- paste(getwd(), "phd_calc_figures", sep = "/")
23
24 #n <- c(dim(d.21)[1], dim(d.33)[1]) # 43, 64
25
26
27 #-----------------------------------------------------------------------------
28 ## Define target grid size:
29 #-----------------------------------------------------------------------------
30
31 # as proposed by T. Hengl 2006 - Finding the right pixel size:
32 # based on inspection density/working scale:
33 ezg.area <- c(.0486,.1074) # km^2
34 obs <- 2.5 # observations per 1cm^2 of the map = recommended compromise
35
36 sn <- sqrt(obs * ezg.area * 1e6/n) * 100 # working scale = appr. 1:5.000
37
38 # the scale number can be used to estimate the grid resolution:
39 p <- sqrt(obs * ezg.area * 1e6/n) * 100 * .0005; p
40 # --> grid resolution = appr. 2.5m (f21: 2.7 vs. f33: 3.2)
41
42 rm(ezg.area,obs,sn,p)
43
44
45 #-----------------------------------------------------------------------------
46 ## Maps of first four PCs:
47 #-----------------------------------------------------------------------------
48
49 catchm <- readOGR(dsn = "phd_calc_input", layer = "fields2133_diss_v2")
50
51 proj4string(catchm) <- CRS(NA); proj4string(dem.33) <- CRS(NA)
52 proj4string(dem.21) <- CRS(NA)
53 proj4string(catchm) <-
```

```
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
proj4string(dem.33) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
proj4string(dem.21) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")

pts.21 <- list("sp.points", d.21, pch = 19, col = "black", cex = .6, alpha = 1)
pts.33 <- list("sp.points", d.33, pch = 19, col = "black", cex = .6, alpha = 1)

prof.21 <- readOGR(dsn = "phd_calc_input", layer = "az_profiles11_pnts21")
prof.33 <- readOGR(dsn = "phd_calc_input", layer = "az_profiles13_pnts33")

cat <- fortify(catchm, region = "DISS")
pt21 <- as.data.frame(pts.21); pt33 <- as.data.frame(pts.33)
prof21 <- as.data.frame(prof.21); prof33 <- as.data.frame(prof.33)
dem21.df <- as.data.frame(dem.21); dem33.df <- as.data.frame(dem.33)
dem <- rbind(dem21.df, dem33.df)
v.interest <- c("PC1", "PC2", "PC3", "x", "y")

pc1.min <- min(dem$PC1, na.rm = TRUE); pc1.min # -6.9
pc1.max <- max(dem$PC1, na.rm = TRUE); pc1.max # 6.6

dem$PC1CUTV <- cut(dem$PC1, breaks = c(-8.25, seq(-6.6,6.6,1.65)))
levels(dem$PC1CUTV) <- c("below -6.6", "-6.6 to -4.95", "-4.95 to -3.30",
  "-3.30 to -1.65", "-1.65 to 0", "0 to 1.65", "1.65 to 3.30",
   "3.30 to 4.95", "4.95 to 6.60")

map.PC1 <- ggplot(cat, aes(long, lat)) +
  geom_raster(aes(x, y, fill = PC1CUTV), data = dem) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.4,
    data = pt21, shape = 21) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.4,
    data = pt33, shape = 21) +
  geom_polygon(size = .5, linetype = "dashed", color = "black",
    fill = "grey40", alpha = 0) +
  geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 3,
    data = prof21, shape = 18) +
  geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 3,
    data = prof33, shape = 18) +
  coord_equal() +
  theme_bw(base_size = 9, base_family = "Helvetica") +
  scale_fill_manual(
    name = "PC1", values = rev(brewer.pal(9,"Spectral"))) +
  labs(x = "\nUTM-E/m", y = "UTM-N/m\n") +
  scale_y_continuous(breaks = c(4362600,4362800,4363000)) +
  scale_x_continuous(breaks = c(508800,509000,509200)) +
  theme(axis.text.y = element_text(angle = 90, hjust = .5),
    legend.text = element_text(size = 11), legend.key.size = unit(18, "pt"),
    legend.title = element_text(size = 11),
    axis.text = element_text(size = 11), axis.title = element_text(size = 11))
```

```
104
105 pc2.min <- min(dem$PC2, na.rm = TRUE); pc2.min # -5.6
106 pc2.max <- max(dem$PC2, na.rm = TRUE); pc2.max # 7.3
107
108 dem$PC2CUTV <- cut(dem$PC2, breaks = c(seq(-5.6,5.6,1.4), 7.3))
109 levels(dem$PC2CUTV) <- c("-5.6 to -4.2", "-4.2 to -2.8", "-2.8 to -1.4",
110   "-1.4 to 0", "0 to 1.4", "1.4 to 2.8", "2.8 to 4.2",
111    "4.2 to 5.6", "5.6 +")
112
113 map.PC2 <- ggplot(cat, aes(long, lat)) +
114   geom_raster(aes(x, y, fill = PC2CUTV), data = dem) +
115   geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.4,
116     data = pt21, shape = 21) +
117   geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.4,
118     data = pt33, shape = 21) +
119   geom_polygon(size = .5, linetype = "dashed", color = "black",
120     fill = "grey40", alpha = 0) +
121   geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 3,
122     data = prof21, shape = 18) +
123   geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 3,
124     data = prof33, shape = 18) +
125   coord_equal() +
126   theme_bw(base_size = 9, base_family = "Helvetica") +
127   scale_fill_manual(
128     name = "PC2", values = rev(brewer.pal(9,"Spectral"))) +
129   labs(x = "\nUTM-E/m", y = "UTM-N/m\n") +
130   scale_y_continuous(breaks = c(4362600,4362800,4363000)) +
131   scale_x_continuous(breaks = c(508800,509000,509200)) +
132   theme(axis.text.y = element_text(angle = 90, hjust = .5),
133     legend.text = element_text(size = 11), legend.key.size = unit(18, "pt"),
134     legend.title = element_text(size = 11),
135     axis.text = element_text(size = 11), axis.title = element_text(size = 11))
136
137 pc3.min <- min(dem$PC3, na.rm = TRUE); pc3.min # -10.9
138 pc3.max <- max(dem$PC3, na.rm = TRUE); pc3.max # 11.5
139
140 dem$PC3CUTV <- cut(dem$PC3, breaks = c(seq(-11,11,2.75),13.75))
141 levels(dem$PC3CUTV) <- c(
142   "-11 to -8.25", "-8.25 to -5.50", "-5.50 to -2.75", "-2.75 to 0",
143     "0 to 2.75", "2.75 to 5.50", "5.50 to 8.25", "8.25 to 11", "11 +")
144
145 map.PC3 <- ggplot(cat, aes(long, lat)) +
146   geom_raster(aes(x, y, fill = PC3CUTV), data = dem) +
147   geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.4,
148     data = pt21, shape = 21) +
149   geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.4,
150     data = pt33, shape = 21) +
151   geom_polygon(size = .5, linetype = "dashed", color = "black",
152     fill = "grey40", alpha = 0) +
153   geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 3,
```

```
    data = prof21, shape = 18) +
  geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 3,
    data = prof33, shape = 18) +
  coord_equal() +
  theme_bw(base_size = 9, base_family = "Helvetica") +
  scale_fill_manual(
    name = "PC3", values = rev(brewer.pal(9,"Spectral"))) +
  labs(x = "\nUTM-E/m", y = "UTM-N/m\n") +
  scale_y_continuous(breaks = c(4362600,4362800,4363000)) +
  scale_x_continuous(breaks = c(508800,509000,509200)) +
  theme(axis.text.y = element_text(angle = 90, hjust = .5),
    legend.text = element_text(size = 11), legend.key.size = unit(18, "pt"),
    legend.title = element_text(size = 11),
    axis.text = element_text(size = 11), axis.title = element_text(size = 11))

pc4.min <- min(dem$PC4, na.rm = TRUE); pc4.min # -8.3
pc4.max <- max(dem$PC4, na.rm = TRUE); pc4.max # 5.9

dem$PC4CUTV <- cut(dem$PC4, breaks = c(-8.3, seq(-6,6,1.5)))
levels(dem$PC4CUTV) <- c("below -6.0", "-6.0 to -4.5", "-4.5 to -3.0",
  "-3.0 to -1.5", "-1.5 to 0", "0 to 1.5", "1.5 to 3.0",
  "3.0 to 4.5", "4.5 to 6.0")

map.PC4 <- ggplot(cat, aes(long, lat)) +
  geom_raster(aes(x, y, fill = PC4CUTV), data = dem) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.4,
    data = pt21, shape = 21) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.4,
    data = pt33, shape = 21) +
  geom_polygon(size = .5, linetype = "dashed", color = "black",
    fill = "grey40", alpha = 0) +
  geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 3,
    data = prof21, shape = 18) +
  geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 3,
    data = prof33, shape = 18) +
  coord_equal() +
  theme_bw(base_size = 9, base_family = "Helvetica") +
  scale_fill_manual(
    name = "PC4", values = rev(brewer.pal(9,"Spectral"))) +
  labs(x = "\nUTM-E/m", y = "UTM-N/m\n") +
  scale_y_continuous(breaks = c(4362600,4362800,4363000)) +
  scale_x_continuous(breaks = c(508800,509000,509200)) +
  theme(axis.text.y = element_text(angle = 90, hjust = .5),
    legend.text = element_text(size = 11), legend.key.size = unit(18, "pt"),
    legend.title = element_text(size = 11),
    axis.text = element_text(size = 11), axis.title = element_text(size = 11))

# prepare 4 PC-plots for coupled export (vertical and legend harmonized):
gPC1 <- ggplot_gtable(ggplot_build(map.PC1))
gPC2 <- ggplot_gtable(ggplot_build(map.PC2))
```

```
gPC3 <- ggplot_gtable(ggplot_build(map.PC3))
gPC4 <- ggplot_gtable(ggplot_build(map.PC4))

legPC1 <- with(gPC1$grobs[[8]], grobs[[1]]$widths[[4]])
legPC2 <- with(gPC2$grobs[[8]], grobs[[1]]$widths[[4]])
legPC3 <- with(gPC3$grobs[[8]], grobs[[1]]$widths[[4]])
legPC4 <- with(gPC4$grobs[[8]], grobs[[1]]$widths[[4]])

gPC1$widths; gPC2$widths; gPC3$widths; gPC4$widths # --> max. legend width: Th

# set the widths to max. (= PC1):
gPC2$widths <- gPC1$widths; gPC3$widths <-gPC1$widths; gPC4$widths <-gPC1$widths

# add an empty column of "abs(diff(widths)) mm"
#   width on the right of legend box for the smaller legend box:
gPC2$grobs[[8]] <-
  gtable_add_cols(gPC2$grobs[[8]], unit(abs(diff(c(legPC1, legPC2))), "mm"))
gPC3$grobs[[8]] <-
  gtable_add_cols(gPC3$grobs[[8]], unit(abs(diff(c(legPC1, legPC3))), "mm"))
gPC4$grobs[[8]] <-
  gtable_add_cols(gPC4$grobs[[8]], unit(abs(diff(c(legPC1, legPC4))), "mm"))

gg <- arrangeGrob(gPC1, gPC2, gPC3, gPC4, nrow = 2)

# clone ggsave and bypass the class check:
ggsave <- ggplot2::ggsave; body(ggsave) <- body(ggplot2::ggsave)[-2]

ggsave(paste(path.fig, "azienda_PCmaps4_v1.pdf", sep = "/"), gg,
  width = 10.79, height = 8.69)
dev.off()


#------------------------------------------------------------------------------
## (Stepwise) regression modelling
#------------------------------------------------------------------------------

# field 33:
explana <- c("PC1", "PC2", "PC3", "PC4", "PC5", "PC6")
# stepwise regression:
regr.start <- lm(
  paste0("CLAYalr ~ ", paste(explana, collapse = "+")), data = d.33)
regr <- step(regr.start, direction = "backward")
regr.fwd <- step(
  lm(CLAYalr ~ 1, d.33), scope = list(lower = CLAYalr ~ 1, upper = paste0(
    "CLAYalr ~ ", paste(explana, collapse = "+"))), direction = "forward")
# --> forward and backward lead to the same result: CLAYalr ~ PC1 + PC2 + PC3

regr.start2 <- lm(CLAYalr ~ PC3 + PC1 + PC5 + PC2 + PC6 + PC4, data = d.33)
regr2 <- step(regr.start2, direction = "backward")
# --> another order yields the same result as above --> very good
```

```
254 rm(regr.start,regr.fwd,regr.start2,regr2); summary(regr) # r2 = 0.32
255
256 regr.start2 <- lm(
257   paste0("SILTalr ~ ", paste(explana, collapse = "+")), data = d.33)
258 regr2 <- step(regr.start2, direction = "backward")
259 regr.fwd <- step(
260   lm(SILTalr ~ 1, d.33), scope = list(lower = SILTalr ~ 1, upper = paste0(
261     "SILTalr ~ ", paste(explana, collapse = "+"))), direction = "forward")
262 # forward and backward lead to the same result: SILTalr ~ PC1 + PC2
263
264 regr.start3 <- lm(SILTalr ~ PC3 + PC1 + PC5 + PC2 + PC6 + PC4, data = d.33)
265 regr3 <- step(regr.start3, direction = "backward")
266 # --> another order yields the same result as above --> very good
267 rm(regr.start2,regr.fwd,regr.start3,regr3); summary(regr2) # r2 = 0.17
268
269 regr.33.cl <- regr; regr.33.si <- regr2
270
271
272 # field 21:
273 d.21$x <- d.21$EAST; d.21$y <- d.21$NORTH
274
275 explana <- c("PC1", "PC2", "PC3", "PC4", "PC5", "x", "y")
276 # stepwise regression:
277 regr.start <- lm(
278   paste0("CLAYalr ~ ", paste(explana, collapse = "+")), data = d.21)
279 regr <- step(regr.start, direction = "backward")
280 regr.fwd <- step(
281   lm(CLAYalr ~ 1, d.21), scope = list(lower = CLAYalr ~ 1, upper = paste0(
282     "CLAYalr ~ ", paste(explana, collapse = "+"))), direction = "forward")
283 # --> forward and backward lead to the same result: CLAYalr ~ PC1 + x + y
284
285 regr.start2 <- lm(CLAYalr ~ PC3 + y + PC1 + PC5 + PC2 + x + PC4, data = d.21)
286 regr2 <- step(regr.start2, direction = "backward")
287 # --> another order yields the same result as above --> very good
288 rm(regr.start,regr.fwd,regr.start2,regr2); summary(regr) # r2 = 0.67
289
290 regr.start2 <- lm(
291   paste0("SILTalr ~ ", paste(explana, collapse = "+")), data = d.21)
292 regr2 <- step(regr.start2, direction = "backward")
293 regr.fwd <- step(
294   lm(SILTalr ~ 1, d.21), scope = list(lower = SILTalr ~ 1, upper = paste0(
295     "SILTalr ~ ", paste(explana, collapse = "+"))), direction = "forward")
296 # forward and backward lead to the same result: SILTalr ~ PC1 + x + y
297
298 regr.start3 <- lm(SILTalr ~ PC3 + y + PC1 + PC5 + PC2 + x + PC4, data = d.21)
299 regr3 <- step(regr.start3, direction = "backward")
300 # --> another order yields the same result as above --> very good
301 rm(regr.start2,regr.fwd,regr.start3,regr3); summary(regr2) # r2 = 0.74
302 rm(explana)
303
```

```
304 regr.21.cl <- regr; regr.21.si <- regr2; rm(regr,regr2)
305
306
307 #------------------------------------------------------------------------------
308 ## Regression diagnostics
309 #------------------------------------------------------------------------------
310
311 datasets.21 <- list(regr.21.cl, regr.21.si)
312 datasets.33 <- list(regr.33.cl, regr.33.si)
313
314 z <- c("CLAYalr", "SILTalr")
315
316 source("regr_diagnostics.R")
317 rd.21 <- regr_diagn(lm.obj = datasets.21, sp.obj = d.21, vars = z, nk = 4)
318 rd.33 <- regr_diagn(lm.obj = datasets.33, sp.obj = d.33, vars = z, nk = 4)
319 rd.21; rd.33
320 rm(z,regr_diagn,datasets.21,datasets.33)
321
322
323 #------------------------------------------------------------------------------
324 ## Regression Cokriging (RCOK)
325 #------------------------------------------------------------------------------
326
327 d.33.rm <- d.33[c(
328   which(d.33$ID == 68), which(d.33$ID == 77), which(d.33$ID == 92)),]
329
330 d.33 <- d.33[-c(
331   which(d.33$ID == 68), which(d.33$ID == 77), which(d.33$ID == 92)),]
332 n <- c(dim(d.21)[1], dim(d.33)[1]); n # 43, 61
333
334 # field 33:
335 # to derive the beta-gls by debug.level = 32 from predict.gstat-function:
336 # just one unknown location, for better overview
337 d33.pt <- as.data.frame(dem.33)[14000, c("PC1", "PC2", "PC3", "x", "y")]
338 coordinates(d33.pt) <- ~ x + y
339 proj4string(d33.pt) <-
340   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
341 proj4string(d.33) <-
342   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
343 proj4string(dem.33) <-
344   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
345
346 rva <- variogram(formula(regr.33.cl), loc = d.33, cutoff = 270, width = 30)
347 rvm <- vgm(.1, "Sph", 240, 0)
348 rvmf <- fit.variogram(rva, rvm, fit.method = 7)
349
350 g.33 <- gstat(NULL, id = "CLAYalr", form = formula(regr.33.cl), data = d.33)
351 g.33 <- gstat(g.33, id = "SILTalr", form = formula(regr.33.si), data = d.33)
352
353 va.cross.33 <- variogram(g.33, cutoff = 270, width = 30)
```

```
g.33 <- gstat(g.33, id = "CLAYalr", model = rvm, fill.all = T)
g.33 <- fit.lmc(va.cross.33, g.33)
plot(va.cross.33, g.33)
# nugget, partial sill and (constant) range parameter:
g.33$model$CLAYalr[1,2];g.33$model$SILTalr[1,2];g.33$model$CLAYalr.SILTalr[1,2]
g.33$model$CLAYalr[2,2];g.33$model$SILTalr[2,2];g.33$model$CLAYalr.SILTalr[2,2]
g.33$model$CLAYalr.SILTalr[2,3]
# Nugget-to-sill-ratios (NSR) after Cambardella94
g.33$model$CLAYalr[1,2] /
  (g.33$model$CLAYalr[1,2] + g.33$model$CLAYalr[2,2]) * 100
g.33$model$SILTalr[1,2] /
  (g.33$model$SILTalr[1,2] + g.33$model$SILTalr[2,2]) * 100

pt.ucok <- predict.gstat(g.33, d33.pt, debug.level = 32) # debug.level = 32
# --> GLS-parameter for field 33:
beta.gls.33 <- c(-0.0406837499, 0.0509900831, 0.0561748569, 0.0395800627,
  -0.626983275 , 0.0180864224, 0.0639881452)

# derive the GLS residuals:
cl.est.gls33 <- beta.gls.33[1] + beta.gls.33[2] * d.33$PC1 +
  beta.gls.33[3] * d.33$PC2 + beta.gls.33[4] * d.33$PC3
cl.res.gls33 <- d.33$CLAYalr - cl.est.gls33
si.est.gls33 <-
  beta.gls.33[5] + beta.gls.33[6] * d.33$PC1 + beta.gls.33[7] * d.33$PC2
si.res.gls33 <- d.33$SILTalr - si.est.gls33
cor.t <- cor.test(
  cl.res.gls33, si.res.gls33, method = "pearson", alternative = "two.sided")
round(cor.t$estimate, 2); round(cor.t$p.value, 2) # 0.18/0.16

ucok.out <- predict.gstat(g.33, dem.33)

dem.33$CLAY_UCOKALR2 <- ucok.out$CLAYalr.pred
dem.33$SILT_UCOKALR2 <- ucok.out$SILTalr.pred

# backtranform (biased!!!) = additive generalized logistic (agl) transform:
dem.33.alr <- matrix(c(dem.33$CLAY_UCOKALR2, dem.33$SILT_UCOKALR2),
  nrow = length(dem.33$CLAY_UCOKALR2), ncol = 2,
  dimnames = list(NULL, c("CLAY", "SILT")))

dem.comp.back <- alrInv(dem.33.alr)

dem.33$CLAY_UCOKALRBACK2 <- dem.comp.back[,1] * 100
dem.33$SILT_UCOKALRBACK2 <- dem.comp.back[,2] * 100
dem.33$SAND_UCOKALRBACK2 <- dem.comp.back[,3] * 100


# field 21:
d21.pt <- as.data.frame(dem.21)[14000, c("PC1", "x", "y")]
coordinates(d21.pt) <- ~ x + y
proj4string(d21.pt) <-
```

```
404   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
405 proj4string(d.21) <-
406   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
407 proj4string(dem.21) <-
408   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
409
410 rva <- variogram(formula(regr.21.cl), loc = d.21, cutoff = 189, width = 21)
411 rvm <- vgm(.035, "Sph", 75, 0.007)
412 rvmf <- fit.variogram(rva, rvm, fit.method = 7)
413
414 g.21 <- gstat(NULL, id = "CLAYalr", form = formula(regr.21.cl), data = d.21)
415 g.21 <- gstat(g.21, id = "SILTalr", form = formula(regr.21.si), data = d.21)
416
417 va.cross.21 <- variogram(g.21, cutoff = 189, width = 21)
418 g.21 <- gstat(g.21, id = "CLAYalr", model = rvmf, fill.all = T)
419 g.21 <- fit.lmc(va.cross.21, g.21)
420 plot(va.cross.21, g.21)
421 # nugget, partial sill and (constant) range parameter:
422 g.21$model$CLAYalr[1,2];g.21$model$SILTalr[1,2];g.21$model$CLAYalr.SILTalr[1,2]
423 g.21$model$CLAYalr[2,2];g.21$model$SILTalr[2,2];g.21$model$CLAYalr.SILTalr[2,2]
424 g.21$model$CLAYalr.SILTalr[2,3]
425 # Nugget-to-sill-ratios (NSR) after Cambardella94
426 g.21$model$CLAYalr[1,2] /
427   (g.21$model$CLAYalr[1,2] + g.21$model$CLAYalr[2,2]) * 100
428 g.21$model$SILTalr[1,2] /
429   (g.21$model$SILTalr[1,2] + g.21$model$SILTalr[2,2]) * 100
430
431 pt.ucok <- predict.gstat(g.21, d21.pt, debug.level = 32) # debug.level = 32
432 # --> GLS-parameter for field 21:
433 beta.gls.21 <- c(-10392.3092, -0.028691496, -0.0019463484, 0.00260909128,
434   -7906.86297, -0.049241787, -0.00155911801, 0.0019941555)
435
436 # derive the GLS residuals:
437 cl.est.gls21 <- beta.gls.21[1] + beta.gls.21[2] * d.21$PC1 +
438   beta.gls.21[3] * d.21$x + beta.gls.21[4] * d.21$y
439 cl.res.gls21 <- d.21$CLAYalr - cl.est.gls21
440 si.est.gls21 <- beta.gls.21[5] + beta.gls.21[6] * d.21$PC1 +
441   beta.gls.21[7] * d.21$x + beta.gls.21[8] * d.21$y
442 si.res.gls21 <- d.21$SILTalr - si.est.gls21
443 cor.t <- cor.test(
444   cl.res.gls21, si.res.gls21, method = "pearson", alternative = "two.sided")
445 round(cor.t$estimate, 2); round(cor.t$p.value, 2) # 0.28/0.07
446
447 ucok.out <- predict.gstat(g.21, dem.21)
448
449 dem.21$CLAY_UCOKALR2 <- ucok.out$CLAYalr.pred
450 dem.21$SILT_UCOKALR2 <- ucok.out$SILTalr.pred
451
452 dem.21.alr <- matrix(c(dem.21$CLAY_UCOKALR2, dem.21$SILT_UCOKALR2),
453   nrow = length(dem.21$CLAY_UCOKALR2), ncol = 2,
```

```
454   dimnames = list(NULL, c("CLAY", "SILT")))
455
456 dem.comp.back <- alrInv(dem.21.alr)
457
458 dem.21$CLAY_UCOKALRBACK2 <- dem.comp.back[,1] * 100
459 dem.21$SILT_UCOKALRBACK2 <- dem.comp.back[,2] * 100
460 dem.21$SAND_UCOKALRBACK2 <- dem.comp.back[,3] * 100
461
462
463 #------------------------------------------------------------------------------
464 ## Plot regression cokriging estimates at field 21 and 33
465 #------------------------------------------------------------------------------
466
467 # residual (cross-)variograms of alr-transformed targets:
468 pdf(paste(path.fig, "azienda_crossvario_MLRres_v1.pdf", sep = "/"),
469   width = 15, height = 7, pointsize = 21)
470 par(mfrow = c(1,2), mar = c(3,3,2,0)+.1, oma = c(.5,.5,.5,.5), mgp = c(2,1,0),
471   cex.main = 1, font.main = 2, las = 1)
472
473 plot(va.cross.21$gamma[19:27] ~ va.cross.21$dist[19:27], pch = 16, cex = .75,
474   xlim = c(0,max(va.cross.21$dist)*1.05), col = "black", xaxs = "i", yaxs = "i",
475   ylim = c(0,0.057),
476   xlab = "lag distance in m", ylab = "(cross)-variance", main = "a) field 21",
477   axes = FALSE)
478 axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
479 axis(side = 2, tck = -.02, cex.axis = 1, at = c("0.00", "0.01", ".02", ".03",
480   "0.04", "0.05"), labels = NA)
481 axis(side = 1, lwd = 0, line = -.35, at = seq(0,250,50),
482   labels = c("0.1", seq(50,250,50)))
483 axis(side = 2, lwd = 0, line = -.35, at = c("0.00", "0.01", ".02", ".03",
484   "0.04", "0.05"), labels = c("0.00", "0.01", ".02", ".03", ".04", "0.05"))
485 lines(variogramLine(g.21$model$CLAYalr, maxdist = max(va.cross.21$dist[19:27])),
486   col = "black", lwd = 1.35)
487 lines(variogramLine(g.21$model$SILTalr, maxdist = max(va.cross.21$dist[10:18])),
488   col = "black", lwd = 1.35, lty = 2)
489 lines(variogramLine(g.21$model$CLAYalr.SILTalr,
490   maxdist = max(va.cross.21$dist[1:9])), col = "black", lwd = 1.35, lty = 4)
491 points(va.cross.21$dist[10:18], va.cross.21$gamma[10:18],
492   col = "black", pch = 22, cex = .65)
493 points(va.cross.21$dist[1:9], va.cross.21$gamma[1:9],
494   col = "black", pch = 17, cex = .65)
495 legend("topleft", c("alr CLAY", "alr SILT", "Cross"),
496   col = c("black"), pch = c(16, 22, 17), cex = .75)
497 box(which = "plot", lty = "solid", col = "black")
498
499 plot(va.cross.33$gamma[19:27] ~ va.cross.33$dist[19:27], pch = 16, cex = .75,
500   xlim = c(0,max(va.cross.33$dist)*1.05), col = "black", xaxs = "i", yaxs = "i",
501   ylim = c(-0.01,max(va.cross.33$gamma)*1.2),
502   xlab = "lag distance in m", ylab = "", main = "b) field 33",
503   axes = FALSE)
```

```
axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 1, lwd = 0, line = -.35, at = seq(0,250,50),
  labels = c("0.1", seq(50,250,50)))
axis(side = 2, lwd = 0, line = -.35, at = c("0.00", "0.02", ".04", ".06",
  "0.08", "0.10"), labels = c("0.00", "0.02", "0.04", "0.06", "0.08", "0.10"))
lines(variogramLine(g.33$model$CLAYalr, maxdist = max(va.cross.33$dist[19:27])),
  col = "black", lwd = 1.35)
lines(variogramLine(g.33$model$SILTalr, maxdist = max(va.cross.33$dist[10:18])),
  col = "black", lwd = 1.35, lty = 2)
lines(variogramLine(g.33$model$CLAYalr.SILTalr,
  maxdist = max(va.cross.33$dist[1:9])), col = "black", lwd = 1.35, lty = 4)
points(va.cross.33$dist[10:18], va.cross.33$gamma[10:18],
  col = "black", pch = 22, cex = .75)
points(va.cross.33$dist[1:9], va.cross.33$gamma[1:9],
  col = "black", pch = 17, cex = .75)
legend("topleft", c("alr CLAY","alr SILT","Cross"),
  col = c("black"), lty = c(1,2,4), cex = .75)
box(which = "plot", lty = "solid", col = "black")
dev.off()

# Regression cokriging estimates of soil separates:
dem.21$SAND_UCOKALRBACK2[which(is.na(dem.21$EMIH))] <- NA
dem.33$SAND_UCOKALRBACK2[which(is.na(dem.33$EMIH))] <- NA

catchm <- readOGR(dsn = "phd_calc_input", layer = "fields2133_diss_v2")

proj4string(catchm) <- CRS(NA); proj4string(dem.33) <- CRS(NA)
proj4string(dem.21) <- CRS(NA); proj4string(demt) <- CRS(NA)
proj4string(catchm) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
proj4string(dem.33) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
proj4string(dem.21) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
proj4string(demt) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")

pts.21 <- list("sp.points", d.21, pch = 19, col = "black", cex = .6, alpha = 1)
pts.33 <- list("sp.points", d.33, pch = 19, col = "black", cex = .6, alpha = 1)
pts.33.rm <- list(
  "sp.points", d.33.rm, pch = 13, col = "black", cex = .6, alpha = 1)

prof.21 <- readOGR(dsn = "phd_calc_input", layer = "az_profiles11_pnts21")
prof.33 <- readOGR(dsn = "phd_calc_input", layer = "az_profiles13_pnts33")

cat <- fortify(catchm, region = "DISS")
pt21 <- as.data.frame(pts.21); pt33 <- as.data.frame(pts.33)
pt33.rm <- as.data.frame(pts.33.rm)
prof21 <- as.data.frame(prof.21); prof33 <- as.data.frame(prof.33)
```

```
dem21.df <- as.data.frame(dem.21); dem33.df <- as.data.frame(dem.33)
dem <- rbind(dem21.df, dem33.df)
v.interest <-
  c("CLAY_UCOKALRBACK2","SILT_UCOKALRBACK2", "SAND_UCOKALRBACK2", "x", "y")
dem.clay <- dem[,v.interest]
dem.clay$VARS <- "Clay"; dem.clay$VALUE <- dem.clay$CLAY_UCOKALRBACK2
dem.silt <- dem[,v.interest]
dem.silt$VARS <- "Silt"; dem.silt$VALUE <- dem.silt$SILT_UCOKALRBACK2
dem.sand <- dem[,v.interest]
dem.sand$VARS <- "Sand"; dem.sand$VALUE <- dem.sand $SAND_UCOKALRBACK2
dem.clay <- dem.clay[,-c(1:3)]; dem.silt <- dem.silt[,-c(1:3)]
dem.sand <- dem.sand[,-c(1:3)]
demm <- rbind(dem.clay, dem.silt, dem.sand)

ucok.min <- min(demm$VALUE, na.rm = TRUE); ucok.min # 13
ucok.max <- max(demm$VALUE, na.rm = TRUE); ucok.max # 55

demm$CUTV <- cut(demm$VALUE, breaks = c(seq(13,49,4.5),55))
levels(demm$CUTV) <- c(
  "13.0 - 17.5", "17.5 - 22.0", "22.0 - 26.5", "26.5 - 31.0",
    "31.0 - 35.5", "35.5 - 40.0", "40.0 - 44.5", "44.5 - 49.0", "49.0 +")

# change plot order of facet grid by changing the order of levels with factor():
demm$VARS <- factor(demm$VARS, levels = c("Clay", "Silt", "Sand"))

map <- ggplot(cat, aes(long, lat)) +
  geom_raster(aes(x, y, fill = CUTV), data = demm) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.3,
    data = pt21, shape = 21) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.3,
    data = pt33, shape = 21) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.3,
    data = pt33.rm, shape = 21, fill = "blue") +
  geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 2.3,
    data = prof21, shape = 18) +
  geom_point(mapping = aes(x = coords.x1, y = coords.x2), size = 2.3,
    data = prof33, shape = 18) +
  geom_polygon(size = .5, linetype = "solid", color = "black",
    fill = "grey40", alpha = 0) +
  coord_equal() +
  theme_bw(base_size = 9, base_family = "Helvetica") +
  scale_fill_brewer(name = "Content in %", palette = "YlOrBr") +
  labs(x = "\nUTM-E/m", y = "UTM-N/m\n") +
  scale_y_continuous(breaks = c(4362600,4362800,4363000)) +
  scale_x_continuous(breaks = c(508800,509000,509200)) +
  facet_wrap(~ VARS, ncol = 2) +
  theme(axis.text.y = element_text(angle = 90, hjust = .5, size = 12),
    axis.title = element_text(size = 12),
    axis.title.x = element_text(hjust = .25),
    axis.text = element_text(size = 12),
```

```
604     legend.position = c(.75,.25), legend.text = element_text(size = 12),
605     legend.title = element_text(size = 12),
606     strip.text.x = element_text(size = 14),
607     panel.grid.minor = element_blank())
608
609 ggsave(paste(path.fig, "azienda_ucok_predmaps_vPCA_v8.pdf", sep = "/"), map,
610   width = 7.02, height = 8.27)
611 print(map)
612 dev.off()
613
614 rm(regr.21.cl,regr.21.si,regr.33.cl,regr.33.si,ucok.out,pt.ucok,dem.comp.back,
615   va.cross.21,va.cross.33,rva,rvm,rvmf,dem.33.alr,dem.21.alr)
616 rm(v.interest,ucok.max,ucok.min,pts.21,pts.33,pts.33.rm,prof.21,prof.33,map,
617   d21.pt,d33.pt,catchm,pt21,pt33,pt33.rm,prof21,prof33,demm,dem21.df,
618   dem33.df,dem.silt,dem.sand,dem.clay,dem,cat)
619
620
621 #-------------------------------------------------------------------------------
622 ## Export:
623 #-------------------------------------------------------------------------------
624
625 # remove interim results from final objects:
626 drop.c <- c("CLAY_UCOKALR2","SILT_UCOKALR2")
627 dem.33@data <- dem.33@data[,!(names(dem.33) %in% drop.c)]
628 dem.21@data <- dem.21@data[,!(names(dem.21) %in% drop.c)]
629 rm(drop.c)
630
631 save.image(paste(getwd(), "phd_calc_out/az_rcokResults.RData", sep = "/"))
632
633 writeGDAL(dem.33["CLAY_UCOKALRBACK2"], drivername = "GTiff", type = "Float32",
634   paste(getwd(), "phd_calc_out/rcokclay33.tif", sep = "/"),
635   mvFlag = -9999, options = c("TFW = YES", "DECIMAL_PRECISION = 3"))
636
637 writeGDAL(dem.33["SILT_UCOKALRBACK2"], drivername = "GTiff", type = "Float32",
638   paste(getwd(), "phd_calc_out/rcoksilt33.tif", sep = "/"),
639   mvFlag = -9999, options = c("TFW = YES", "DECIMAL_PRECISION = 3"))
640
641 writeGDAL(dem.33["SAND_UCOKALRBACK2"], drivername = "GTiff", type = "Float32",
642   paste(getwd(), "phd_calc_out/rcoksand33.tif", sep = "/"),
643   mvFlag = -9999, options = c("TFW = YES", "DECIMAL_PRECISION = 3"))
644
645 writeGDAL(dem.21["CLAY_UCOKALRBACK2"], drivername = "GTiff", type = "Float32",
646   paste(getwd(), "phd_calc_out/rcokclay21.tif", sep = "/"),
647   mvFlag = -9999, options = c("TFW = YES", "DECIMAL_PRECISION = 3"))
648
649 writeGDAL(dem.21["SILT_UCOKALRBACK2"], drivername = "GTiff", type = "Float32",
650   paste(getwd(), "phd_calc_out/rcoksilt21.tif", sep = "/"),
651   mvFlag = -9999, options = c("TFW = YES", "DECIMAL_PRECISION = 3"))
652
653 writeGDAL(dem.21["SAND_UCOKALRBACK2"], drivername = "GTiff", type = "Float32",
```

```
654   paste(getwd(), "phd_calc_out/rcoksand21.tif", sep = "/"),
655   mvFlag = -9999, options = c("TFW = YES", "DECIMAL_PRECISION = 3"))
656
657 # end of script, azienda_ucok2k5_vDiss.R, 03.09.2012
```

.4.9 azienda_valid_vDiss.R

```
 1 #
 2 ##############################################################################
 3 #### Validation of regression cokriging at field scale, Azienda San Michele ####
 4 ##############################################################################
 5 #
 6 ## last update: 07.05.2015
 7
 8 # validate regression cokriging of alr-transformed soil separates
 9
10 # author: Michael Blaschek (blaschek@geographie.uni-kiel.de)
11
12 rm(list = ls())
13
14 library(rgdal); library(gstat); library(RColorBrewer); library(compositions)
15 library(caret); library(ggplot2)
16 # R v3.0.2, gstat_1.0-16, rgdal_0.8-15, RColorBrewer_1.0-5, compositions_1.40-1
17 # caret_6.0-22, ggplot2_0.9.3.1
18
19 load(paste(getwd(), "phd_calc_out/az_rcokResults.RData", sep = "/"))
20
21
22 #-----------------------------------------------------------------------------
23 ## Leave-one-out cross-validation:
24 #-----------------------------------------------------------------------------
25
26 # field 21:
27 ucok.cv <- gstat.cv(g.21, remove.all = T, all.residuals = T, verbose = T)
28
29 d.21$CLAY_UCOKALR2 <- d.21$CLAYalr - ucok.cv$CLAYalr
30 d.21$SILT_UCOKALR2 <- d.21$SILTalr - ucok.cv$SILTalr
31
32 d2.alr <- matrix(c(d.21$CLAY_UCOKALR2, d.21$SILT_UCOKALR2),
33   nrow = length(d.21$CLAYalr), ncol = 2,
34   dimnames = list(NULL, c("CLAY", "SILT")))
35
36 d.comp.back <- alrInv(d2.alr)
37
38 d.21$CLAY_UCOKALRBACK2 <- d.comp.back[,1] * 100
39 d.21$SILT_UCOKALRBACK2 <- d.comp.back[,2] * 100
40 d.21$SAND_UCOKALRBACK2 <- d.comp.back[,3] * 100
41
42 # field 33:
```

```
ucok.cv <- gstat.cv(g.33, remove.all = T, all.residuals = T, verbose = T)
# all.residuals = T --> residuals for all variables are returned

d.33$CLAY_UCOKALR2 <- d.33$CLAYalr - ucok.cv$CLAYalr
d.33$SILT_UCOKALR2 <- d.33$SILTalr - ucok.cv$SILTalr

# backtranform (biased!!!) = additive generalized logistic (agl) transform:
d2.alr <- matrix(c(d.33$CLAY_UCOKALR2, d.33$SILT_UCOKALR2),
  nrow = length(d.33$CLAYalr), ncol = 2,
  dimnames = list(NULL, c("CLAY", "SILT")))

d.comp.back <- alrInv(d2.alr)

d.33$CLAY_UCOKALRBACK2 <- d.comp.back[,1] * 100
d.33$SILT_UCOKALRBACK2 <- d.comp.back[,2] * 100
d.33$SAND_UCOKALRBACK2 <- d.comp.back[,3] * 100

rm(d.comp.back,d2.alr,ucok.cv)


#------------------------------------------------------------------------------
## Univariate validation (residual-based + association-based):
#------------------------------------------------------------------------------

source("valid_measures.R")

z <- c("CLAY", "SILT", "SAND")
methods <- "RCOK"; methods2 <- "UCOKALRBACK2"

val.msr.21 <- univar_error_metrics(data = d.21@data, targets = z,
  methods.nm = methods, methods.attr1 = methods2, sp.obj = d.21); val.msr.21

val.msr.33 <- univar_error_metrics(data = d.33@data, targets = z,
  methods.nm = methods, methods.attr1 = methods2, sp.obj = d.33); val.msr.33

rm(univar_error_metrics)


#------------------------------------------------------------------------------
## Combined goodness of estimation measure:
#------------------------------------------------------------------------------

nm <- length(methods); methods3 <- methods2
stress.msr.21 <- data.frame(STRESS = rep(1,nm))
rownames(stress.msr.21) <- methods

delta.ij <- NULL; delta.z0.ij <- NULL; j <- 2
for (q in 1:length(methods2)) {
  x <- acomp(d.21@data, parts = c("CLAY", "SILT", "SAND"), total = 100)
  y <- acomp(d.21@data, parts = c(paste(z[1], methods2[q], sep = "_"),
```

```
93      paste(z[2], methods2[q], sep = "_"),
94        paste(z[3], methods3[q], sep = "_")), total = 100)
95    delta.ij <- NULL; delta.z0.ij <- NULL; j <- 2
96    for (i in 1:(length(x[,1])-1)) {
97      delta.ij[i]  <- sum((clr(x[i,]) - clr(x[j,]))^2)
98      delta.z0.ij[i]  <- sum((clr(y[i,]) - clr(y[j,]))^2)
99      j <- j + 1
100   }
101   stress.msr.21[q,1] <- sqrt(sum((delta.ij - delta.z0.ij)^2)/sum(delta.ij^2))
102 }; stress.msr.21 # 0.82
103
104 stress.msr.33 <- data.frame(STRESS = rep(1,nm))
105 rownames(stress.msr.33) <- methods
106
107 delta.ij <- NULL; delta.z0.ij <- NULL; j <- 2
108 for (q in 1:length(methods2)) {
109   x <- acomp(d.33@data, parts = c("CLAY", "SILT", "SAND"), total = 100)
110   y <- acomp(d.33@data, parts = c(paste(z[1], methods2[q], sep = "_"),
111     paste(z[2], methods2[q], sep = "_"),
112       paste(z[3], methods3[q], sep = "_")), total = 100)
113   delta.ij <- NULL; delta.z0.ij <- NULL; j <- 2
114   for (i in 1:(length(x[,1])-1)) {
115     delta.ij[i]  <- sum((clr(x[i,]) - clr(x[j,]))^2)
116     delta.z0.ij[i]  <- sum((clr(y[i,]) - clr(y[j,]))^2)
117     j <- j + 1
118   }
119   stress.msr.33[q,1] <- sqrt(sum((delta.ij - delta.z0.ij)^2)/sum(delta.ij^2))
120 }; stress.msr.33 # 0.89
121 rm(i,j,q,x,y,delta.ij,delta.z0.ij,nm,methods3)
122
123
124 #-----------------------------------------------------------------------------
125 ## Scatterplots of obs vs. pred regression:
126 #-----------------------------------------------------------------------------
127
128 # field 21:
129 lr1 <- lm((d.21$CLAY ~ d.21$CLAY_UCOKALRBACK2))
130 lr2 <- lm((d.21$SILT ~ d.21$SILT_UCOKALRBACK2))
131 lr3 <- lm((d.21$SAND ~ d.21$SAND_UCOKALRBACK2))
132
133 yx <- NULL; yx.lm <- list(NULL); lines = list(NULL); y <- NULL; x <- NULL
134 for (i in 1:length(z)){
135   y <- d.21@data[,z[i]]
136   x <- d.21@data[,paste(z[i], methods2[1], sep = "_")]
137   yx[[i]] <- data.frame(Y = y, X = x, VARS = z[i])
138   yx[[i]]$VARS <- as.character(yx[[i]]$VARS)
139   yx.lm[[i]] <- lm(y ~ x)
140   # preparing elements for 1:1 and regression lines in scatterplots:
141   lines[[i]] <- as.data.frame(cbind(as.numeric(yx.lm[[i]]$coefficients[2]),
142     as.numeric(yx.lm[[i]]$coefficients[1]),
```

```
143     as.numeric(summary(yx.lm[[i]])$r.squared)))
144   colnames(lines[[i]]) <- c("SLOPE_LM", "INTERC_LM", "RSQUARED_LM")
145 }
146
147 yx3 <- rbind(yx[[1]], yx[[2]], yx[[3]])
148 line.e <- rbind(lines[[1]], lines[[2]], lines[[3]])
149
150 yx3[which(yx3$VARS == "CLAY"),"VARS"] <- "Clay"
151 yx3[which(yx3$VARS == "SILT"),"VARS"] <- "Silt"
152 yx3[which(yx3$VARS == "SAND"),"VARS"] <- "Sand"
153
154 line.e$VARS <- c("Clay", "Silt", "Sand")
155
156 yx3$VARS <- factor(yx3$VARS, levels = c("Clay", "Silt", "Sand"))
157 line.e$VARS <- factor(line.e$VARS, levels = c("Clay", "Silt", "Sand"))
158
159 regr.sc <- ggplot(data = yx3, aes(y = Y, x = X)) +
160   geom_point(shape = 1, size = 2) +
161   geom_abline(data = line.e,
162     mapping = aes(slope = SLOPE_LM, intercept = INTERC_LM)) +
163   geom_abline(
164     yintercept = 0, slope = 1, linetype = "dashed", colour = "gray50") +
165   scale_x_continuous(limits = c(0, 60)) +
166   scale_y_continuous(limits = c(0, 60)) +
167   labs(x = "\nPredicted value in %", y = "Observed value in %\n") +
168   facet_wrap(~ VARS, ncol = 3) +
169   coord_equal() +
170   theme_bw(base_size = 12, base_family = "Helvetica") +
171   theme(axis.text.y = element_text(size = 12),
172     axis.title = element_text(size = 13),
173     axis.text.x = element_text(size = 12),
174     strip.text.x = element_text(size = 12),
175     panel.grid.minor = element_blank())
176 print(regr.sc)
177
178 regr.sc2 <- regr.sc +
179   geom_text(data = line.e, aes(x = 2.5, y = 55,
180     label = paste0("y = ", round(SLOPE_LM, 2), "x+", round(INTERC_LM, 2))),
181     vjust = .5, hjust = 0, parse = FALSE, size = 4) +
182   geom_text(data = line.e, aes(x = 57.5, y = 5,
183     label = paste0("R^2 == ", round(RSQUARED_LM, 2))),
184     vjust = .5, hjust = 1, parse = TRUE, size = 4)
185 print(regr.sc2)
186
187 ggsave(paste(path.fig, "field21_valScatter_v2.pdf", sep = "/"),
188   regr.sc2, width = 8.27, height = 3.64)
189 print(regr.sc2)
190 dev.off()
191
192 # field 33:
```

```
lr1 <- lm((d.33$CLAY ~ d.33$CLAY_UCOKALRBACK2))
lr2 <- lm((d.33$SILT ~ d.33$SILT_UCOKALRBACK2))
lr3 <- lm((d.33$SAND ~ d.33$SAND_UCOKALRBACK2))

yx <- NULL; yx.lm <- list(NULL); lines = list(NULL); y <- NULL; x <- NULL
for (i in 1:length(z)){
  y <- d.33@data[,z[i]]
  x <- d.33@data[,paste(z[i], methods2[1], sep = "_")]
  yx[[i]] <- data.frame(Y = y, X = x, VARS = z[i])
  yx[[i]]$VARS <- as.character(yx[[i]]$VARS)
  yx.lm[[i]] <- lm(y ~ x)
  # preparing elements for 1:1 and regression lines in scatterplots:
  lines[[i]] <- as.data.frame(cbind(as.numeric(yx.lm[[i]]$coefficients[2]),
    as.numeric(yx.lm[[i]]$coefficients[1]),
    as.numeric(summary(yx.lm[[i]])$r.squared)))
  colnames(lines[[i]]) <- c("SLOPE_LM", "INTERC_LM", "RSQUARED_LM")
}

yx3 <- rbind(yx[[1]], yx[[2]], yx[[3]])
line.e <- rbind(lines[[1]], lines[[2]], lines[[3]])

yx3[which(yx3$VARS == "CLAY"),"VARS"] <- "Clay"
yx3[which(yx3$VARS == "SILT"),"VARS"] <- "Silt"
yx3[which(yx3$VARS == "SAND"),"VARS"] <- "Sand"

line.e$VARS <- c("Clay", "Silt", "Sand")

yx3$VARS <- factor(yx3$VARS, levels = c("Clay", "Silt", "Sand"))
line.e$VARS <- factor(line.e$VARS, levels = c("Clay", "Silt", "Sand"))

regr.sc <- ggplot(data = yx3, aes(y = Y, x = X)) +
  geom_point(shape = 1, size = 2) +
  geom_abline(data = line.e,
    mapping = aes(slope = SLOPE_LM, intercept = INTERC_LM)) +
  geom_abline(
    yintercept = 0, slope = 1, linetype = "dashed", colour = "gray50") +
  scale_x_continuous(limits = c(0, 60)) +
  scale_y_continuous(limits = c(0, 60)) +
  labs(x = "\nPredicted value in %", y = "Observed value in %\n") +
  facet_wrap(~ VARS, ncol = 3) +
  coord_equal() +
  theme_bw(base_size = 12, base_family = "Helvetica") +
  theme(axis.text.y = element_text(size = 12),
    axis.title = element_text(size = 13),
    axis.text.x = element_text(size = 12),
    strip.text.x = element_text(size = 12),
    panel.grid.minor = element_blank())
print(regr.sc)

regr.sc2 <- regr.sc +
```

```
  geom_text(data = line.e, aes(x = 2.5, y = 55,
    label = paste0("y = ", round(SLOPE_LM, 2), "x+", round(INTERC_LM, 2))),
    vjust = .5, hjust = 0, parse = FALSE, size = 4) +
  geom_text(data = line.e, aes(x = 57.5, y = 5,
    label = paste0("R^2 == ", round(RSQUARED_LM, 2))),
    vjust = .5, hjust = 1, parse = TRUE, size = 4)
print(regr.sc2)

ggsave(paste(path.fig, "field33_valScatter_v2.pdf", sep = "/"),
  regr.sc2, width = 8.27, height = 3.64)
print(regr.sc2)
dev.off()

rm(line.e,yx3,i,lines,lr1,lr2,lr3,regr.sc,regr.sc2,x,y,yx,yx.lm,methods,nm)


#------------------------------------------------------------------------------
## Bubble plots of validation residuals:
#------------------------------------------------------------------------------

catchm <- readOGR(dsn = "phd_calc_input", layer = "fields2133_diss_v2")
cat <- fortify(catchm, region = "DISS")

d.33$x <- d.33$EAST; d.33$y <- d.33$NORTH
dv <- rbind(d.21,d.33)
bmx <- dv; dv.id <- as.data.frame(dv[,1]); msc <- list(NULL)
for (i in 1:length(z)){
  bmx@data[paste(z[i], "RES", sep = "_")] <-
    dv@data[,z[i]] - dv@data[,paste(z[i], methods2[1], sep = "_")]
  msc[[i]] <- as.data.frame(
    cbind(bmx$ID, bmx@data[,paste(z[i], "RES", sep = "_")]))
  colnames(msc[[i]]) <- c("PNTID", "RES")
  msc[[i]]$EAST <- dv.id[,2]; msc[[i]]$NORTH <- dv.id[,3]
  # define whether the model over- or under-estimates:
  msc[[i]]$SIGN <- ifelse(msc[[i]]$RES > 0,
    msc[[i]]$SIGN <- "Underestimated", msc[[i]]$SIGN <- "Overestimated")
  msc[[i]]$VARS <- z[i] # needed for facets in ggplot2
}

msc3 <- rbind(msc[[1]], msc[[2]], msc[[3]])
msc3[which(msc3$VARS == "CLAY"),"VARS"] <- "Clay"
msc3[which(msc3$VARS == "SILT"),"VARS"] <- "Silt"
msc3[which(msc3$VARS == "SAND"),"VARS"] <- "Sand"
msc3$SIGN <- factor(msc3$SIGN)
msc3$VARS <- factor(msc3$VARS, levels = c("Clay", "Silt", "Sand"))

pr.br <- pretty(c(min(abs(msc3$RES)), max(abs(msc3$RES))))

val.r <- ggplot(cat, aes(long, lat)) +
  geom_point(data = msc3, aes(x = EAST, y = NORTH, size = abs(RES),
```

```
293     fill = SIGN), shape = 21, colour = "black") +
294   scale_size_area(breaks = pr.br, max_size = 6) +
295   geom_polygon(size = .5, linetype = "dashed", color = "black",
296     fill = "grey40", alpha = 0) +
297   facet_wrap(~ VARS, ncol = 2) +
298   coord_equal() +
299   theme_bw(base_size = 15, base_family = "Helvetica") +
300   guides(fill = guide_legend(order = 2, override.aes = list(size = 4,
301     shape = 21)), size = guide_legend(order = 1, override.aes = list(
302       colour = "black", shape = 21))) +
303   labs(x = "\nUTM-E/m", y = "UTM-N/m\n", fill = "Misprediction",
304     size = "Absolute residual\ncontent in %") +
305   scale_y_continuous(breaks = c(4362600,4362800,4363000)) +
306   scale_x_continuous(breaks = c(508800,509000,509200)) +
307   theme(axis.text.y = element_text(angle = 90, hjust = .5, size = 12),
308     axis.title = element_text(size = 12),
309     axis.title.x = element_text(hjust = .25),
310     axis.text.x = element_text(size = 12),
311     strip.text.x = element_text(size = 14),
312     legend.position = c(.75,.25), legend.box = "vertical",
313     panel.grid.minor = element_blank())
314
315 ggsave(paste(path.fig, "azienda_resBubble_v2.pdf", sep = "/"),
316   val.r, width = 7.02, height = 8.27)
317 print(val.r)
318 dev.off()
319 rm(msc3,bmx,i,methods2,z,msc,pr.br,val.r,dv,cat,dv.id,catchm)
320 save.image(paste(getwd(), "phd_calc_out/az_validResults.RData", sep = "/"))
321
322 # end of script, azienda_valid_vDiss.R, 27.07.2014
```

.4.10 costara_dta20_vDiss.R

```
 1 #
 2 ##############################################################################
 3 ####  Extraction of land-surface parameter using SAGA GIS, Rio di Costara   ####
 4 ##############################################################################
 5 #
 6 ## last update: 01.12.2014
 7
 8 # digital terrain analysis based on "phd_calc_input/dem_cos20.asc"
 9 # = aggregated from "dem_cos3cl.asc"
10 # = extracted (clip) from "ascii_50000_548.asc"
11 # = derived from http://www.sardegnageoportale.it/ --> Catalogo Dati -->
12 # Download --> Raccolte cartografiche --> Modello Digitale del Terreno SAR,
13 #   passo 10m --> Scarica il DTM (last access: 11.01.2011)
14
15 # author: Michael Blaschek (blaschek@geographie.uni-kiel.de)
16 # reference: Tomislav Hengl - Analysis of DEMs in R+ILWIS/SAGA
```

```
#   (from http://spatial-analyst.net/wiki/)

rm(list = ls())

library(rgdal); library(RSAGA); library(compositions)
library(ggplot2); library(grid); library(RColorBrewer)
# R v3.0.2, SAGA v2.0.8, rgdal_0.8-15, RSAGA v0.93-6, compositions_1.40-1
#   ggplot2_0.9.3.1, RColorBrewer_1.0-5

myenv <- rsaga.env(path = "/usr/bin") # --> define, where SAGA is located..
#myenv <- rsaga.env(path = "C:/Program Files (x86)/SAGA/saga_2.0.8_bin_msw_x64")

dem1 <- readGDAL(paste(getwd(), "phd_calc_input", "dem_cos20.asc", sep = "/"))
names(dem1)[1] <- "ELEV"


#-----------------------------------------------------------------------------
## Extraction of land-surface parameter
#-----------------------------------------------------------------------------

rsaga.esri.to.sgrd(env = myenv, in.grids = "phd_calc_input/dem_cos20.asc",
  out.sgrds = "phd_calc_out/demanalysis_r2/dem_cos20.sgrd", in.path = getwd())

# SAGA Wetness Index:
rsaga.geoprocessor(lib = "ta_hydrology", module = 15, env = myenv,
  param = list(DEM = "phd_calc_out/demanalysis_r2/dem_cos20.sgrd",
    C = "phd_calc_out/demanalysis_r2/dtaout_cos20/catcharea1.sgrd",
    GN = "phd_calc_out/demanalysis_r2/dtaout_cos20/catchslope1.sgrd",
    CS = "phd_calc_out/demanalysis_r2/dtaout_cos20/modcatcharea1.sgrd",
    SB = "phd_calc_out/demanalysis_r2/dtaout_cos20/sagawi1.sgrd", T = 10))

# primary attributes (local morphometry)
# after Zevenbergen & Thorne 1987 = METHOD 5:
# slope in rad (m/m):
rsaga.geoprocessor(lib = "ta_morphometry", module = 0, env = myenv,
  param = list(ELEVATION = "phd_calc_out/demanalysis_r2/dem_cos20.sgrd",
    SLOPE = "phd_calc_out/demanalysis_r2/dtaout_cos20/sloperad1.sgrd",
    ASPECT = "phd_calc_out/demanalysis_r2/dtaout_cos20/aspect1.sgrd",
    HCURV = "phd_calc_out/demanalysis_r2/dtaout_cos20/plancurv1.sgrd",
    VCURV = "phd_calc_out/demanalysis_r2/dtaout_cos20/profcurv1.sgrd",
    METHOD = 5))

# convert curvatures from 1/m to 1/100m:
rsaga.geoprocessor(lib = "grid_calculus", module = 1, env = myenv,
 param = list(GRIDS = "phd_calc_out/demanalysis_r2/dtaout_cos20/profcurv1.sgrd",
   RESULT = "phd_calc_out/demanalysis_r2/dtaout_cos20/profcurv1001.sgrd",
   FORMULA = "g1*100", FNAME = T))
rsaga.geoprocessor(lib = "grid_calculus", module = 1, env = myenv,
 param = list(GRIDS = "phd_calc_out/demanalysis_r2/dtaout_cos20/plancurv1.sgrd",
   RESULT = "phd_calc_out/demanalysis_r2/dtaout_cos20/plancurv1001.sgrd",
```

```
67    FORMULA = "g1*100", FNAME = T))
68
69 # convert aspect from rad to degree:
70 rsaga.geoprocessor(lib = "grid_calculus", module = 1, env = myenv,
71   param = list(GRIDS = "phd_calc_out/demanalysis_r2/dtaout_cos20/aspect1.sgrd",
72     RESULT = "phd_calc_out/demanalysis_r2/dtaout_cos20/aspectdeg1.sgrd",
73     FORMULA = "g1*180/pi()", FNAME = T))
74
75 # convergence/divergence index:
76 rsaga.geoprocessor(lib = "ta_morphometry", module = 2, env = myenv,
77   param = list(ELEVATION = "phd_calc_out/demanalysis_r2/dem_cos20.sgrd",
78     CONVERGENCE = "phd_calc_out/demanalysis_r2/dtaout_cos20/convg1.sgrd",
79     RADIUS = 3, DISTANCE_WEIGHTING_WEIGHTING = 0, SLOPE = T))
80
81
82 # secondary attributes (topographic indices):
83 # topographic wetness index:
84 rsaga.geoprocessor(lib = "ta_hydrology", module = 20, env = myenv,
85  param = list(SLOPE = "phd_calc_out/demanalysis_r2/dtaout_cos20/sloperad1.sgrd",
86    AREA = "phd_calc_out/demanalysis_r2/dtaout_cos20/catcharea1.sgrd",
87    TWI = "phd_calc_out/demanalysis_r2/dtaout_cos20/twi1.sgrd"))
88
89 # stream power index:
90 rsaga.geoprocessor(lib = "ta_hydrology", module = 21, env = myenv,
91  param = list(SLOPE = "phd_calc_out/demanalysis_r2/dtaout_cos20/sloperad1.sgrd",
92    AREA = "phd_calc_out/demanalysis_r2/dtaout_cos20/catcharea1.sgrd",
93    SPI = "phd_calc_out/demanalysis_r2/dtaout_cos20/streampow1.sgrd"))
94
95 # LS-Factor (Moore et al. 1991, Erosivity = 1):
96 rsaga.geoprocessor(lib = "ta_hydrology", module = 22, env = myenv,
97  param = list(SLOPE = "phd_calc_out/demanalysis_r2/dtaout_cos20/sloperad1.sgrd",
98    AREA = "phd_calc_out/demanalysis_r2/dtaout_cos20/catcharea1.sgrd",
99    LS = "phd_calc_out/demanalysis_r2/dtaout_cos20/ls1.sgrd",
100   CONV = 0, METHOD = 0, EROSIVITY = 1, STABILITY = 0))
101
102 # incoming solar radiation:
103 rsaga.geoprocessor(lib = "ta_lighting", module = 2, env = myenv,
104   param = list(GRD_DEM = "phd_calc_out/demanalysis_r2/dem_cos20.sgrd",
105     GRD_DIRECT = "phd_calc_out/demanalysis_r2/dtaout_cos20/directins1.sgrd",
106     GRD_DIFFUS = "phd_calc_out/demanalysis_r2/dtaout_cos20/diffusins1.sgrd",
107     GRD_TOTAL = "phd_calc_out/demanalysis_r2/dtaout_cos20/insolat1.sgrd",
108     LATITUDE = 39, DHOUR = 4, PERIOD = 2, DDAYS = 10, DAY_A = 20,
109     MON_A = 2, DAY_B = 19, MON_B = 2, METHOD = 2, LUMPED = 70))
110
111 # converting the resulting grids to ESRI-ASCII Grid:
112 rsaga.sgrd.to.esri(
113   in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_cos20/sloperad1.sgrd",
114   out.grids = "phd_calc_out/demanalysis_r2/dtaout_cos20/sloperad1.asc",
115   prec = 6, out.path = getwd(), env = myenv)
116 rsaga.sgrd.to.esri(
```

```
117   in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_cos20/sagawi1.sgrd",
118   out.grids = "phd_calc_out/demanalysis_r2/dtaout_cos20/sagawi1.asc",
119   prec = 2, out.path = getwd(), env = myenv)
120 rsaga.sgrd.to.esri(
121   in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_cos20/plancurv1001.sgrd",
122   out.grids = "phd_calc_out/demanalysis_r2/dtaout_cos20/plancurv1001.asc",
123   prec = 6, out.path = getwd(), env = myenv)
124 rsaga.sgrd.to.esri(
125   in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_cos20/profcurv1001.sgrd",
126   out.grids = "phd_calc_out/demanalysis_r2/dtaout_cos20/profcurv1001.asc",
127   prec = 6, out.path = getwd(), env = myenv)
128 rsaga.sgrd.to.esri(
129   in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_cos20/aspectdeg1.sgrd",
130   out.grids = "phd_calc_out/demanalysis_r2/dtaout_cos20/aspectdeg1.asc",
131   prec = 2, out.path = getwd(), env = myenv)
132 rsaga.sgrd.to.esri(
133   in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_cos20/convg1.sgrd",
134   out.grids = "phd_calc_out/demanalysis_r2/dtaout_cos20/convg1.asc",
135   prec = 4, out.path = getwd(), env = myenv)
136 rsaga.sgrd.to.esri(
137   in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_cos20/twi1.sgrd",
138   out.grids = "phd_calc_out/demanalysis_r2/dtaout_cos20/twi1.asc",
139   prec = 2, out.path = getwd(), env = myenv)
140 rsaga.sgrd.to.esri(
141   in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_cos20/streampow1.sgrd",
142   out.grids = "phd_calc_out/demanalysis_r2/dtaout_cos20/streampow1.asc",
143   prec = 2, out.path = getwd(), env = myenv)
144 rsaga.sgrd.to.esri(
145   in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_cos20/ls1.sgrd",
146   out.grids = "phd_calc_out/demanalysis_r2/dtaout_cos20/ls1.asc",
147   prec = 2, out.path = getwd(), env = myenv)
148 rsaga.sgrd.to.esri(
149   in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_cos20/directins1.sgrd",
150   out.grids = "phd_calc_out/demanalysis_r2/dtaout_cos20/directins1.asc",
151   prec = 4, out.path = getwd(), env = myenv)
152 rsaga.sgrd.to.esri(
153   in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_cos20/diffusins1.sgrd",
154   out.grids = "phd_calc_out/demanalysis_r2/dtaout_cos20/diffusins1.asc",
155   prec = 4, out.path = getwd(), env = myenv)
156 rsaga.sgrd.to.esri(
157   in.sgrds = "phd_calc_out/demanalysis_r2/dtaout_cos20/insolat1.sgrd",
158   out.grids = "phd_calc_out/demanalysis_r2/dtaout_cos20/insolat1.asc",
159   prec = 4, out.path = getwd(), env = myenv)
160
161
162 #-----------------------------------------------------------------------------
163 ## Creating dataset + target grid for digital soil mapping
164 #-----------------------------------------------------------------------------
165
166 # loading covariates:
```

```
j <- c("SAGAWI","SLOPE","ASPECT","PLANC","PROFC",
  "CONVG","TWI","STREAMP","LS","DIRECT","DIFFUS","INSOLAT")
k <- c("sagawi1","sloperad1","aspectdeg1","plancurv1001","profcurv1001",
  "convg1","twi1","streampow1","ls1","directins1","diffusins1","insolat1")
m = 1

for (i in k) {
  dem1@data[i] <- readGDAL(paste("phd_calc_out/demanalysis_r2/dtaout_cos20/",
    i, ".asc", sep = ""))$band1
  names(dem1)[m+1] <- j[m]; m <- m + 1
}
rm(i,j,k,m)

# extracting the overlaying geoltype for each grid cell of dem1:
geol <- readOGR(dsn = "phd_calc_input/costara_geology",
  layer = "mannu_geol_map_v2")
dem2 <- dem1
proj4string(geol) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
proj4string(dem2) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
dem1.ov <- over(dem2, geol)
dem1$GEOLTYPE <- dem1.ov$geoltype
dem1$GEOLTYPE[which(dem1$GEOLTYPE == -99)] <- NA

# create indicator variables of the four main geological units:
dem1$GEOL1 <- 0; dem1$GEOL2 <- 0; dem1$GEOL4 <- 0; dem1$GEOL5 <- 0
dem1$GEOL1[which(dem1$GEOLTYPE == 1)] <- 1
dem1$GEOL2[which(dem1$GEOLTYPE == 2)] <- 1
dem1$GEOL4[which(dem1$GEOLTYPE == 4)] <- 1
dem1$GEOL5[which(dem1$GEOLTYPE == 5)] <- 1

# extracting the overlaying GEOPPR for each grid cell of dem1:
dem1$GEOPPR <- dem1.ov$GEOPPR

# create indicator variables of the six most frequent GEOPPR units:
dem1$GEOPPR12 <- 0; dem1$GEOPPR18 <- 0; dem1$GEOPPR22 <- 0; dem1$GEOPPR470 <- 0
dem1$GEOPPR1233 <- 0; dem1$GEOPPR1465 <- 0
dem1$GEOPPR12[which(dem1$GEOPPR == 12)] <- 1
dem1$GEOPPR18[which(dem1$GEOPPR == 18)] <- 1
dem1$GEOPPR22[which(dem1$GEOPPR == 22)] <- 1
dem1$GEOPPR470[which(dem1$GEOPPR == 470)] <- 1
dem1$GEOPPR1233[which(dem1$GEOPPR == 1233)] <- 1
dem1$GEOPPR1465[which(dem1$GEOPPR == 1465)] <- 1

rm(dem2, geol, dem1.ov, myenv)

# load the laboratory data:
d.all <- read.csv2("phd_calc_input/costara_cau_soilp_lab_v4.csv",
  na.strings = "-999,000")
```

```
217
218 # restrict dataset to A-horizon:
219 d <- subset(d.all, d.all$HORIZON == "A"); str(d)
220 str(d.all)
221
222 # logratio transform to account for the compositional character of the targets:
223 d.comp <- acomp(d@data, parts = c("CLAY", "SILT", "SAND"), total = 100)
224 d.alr <- alr(d.comp) # additive logratio transform!
225 d$CLAYalr <- d.alr[,1]; d$SILTalr <- d.alr[,2]
226
227 # loading soil bulk density data (n.total = 83, n.mannu = 34):
228 n <- length(d$CORG)
229 d.db <- read.csv2("phd_calc_input/sardinia10_dB_importR_vers1.csv")
230 j <- 1
231 for (i in 1:n) {
232   ifelse(d$ID[i] == d.db$ID[j], {d$BULKDENSAV[i] <- d.db$BULKDENSAV[j];
233     j <- j + 1}, d$BULKDENSAV[i] <- NA)
234 }
235 # --> 49 of 197 locations were analysed for soil bulk density
236
237 coordinates(d) <- ~ EAST + NORTH
238
239 dem1.ov <- over(d, dem1)
240 d$ELEV <- dem1.ov$ELEV
241 d$SLOPE <- dem1.ov$SLOPE; d$PROFC <- dem1.ov$PROFC; d$PLANC <- dem1.ov$PLANC
242 d$ASPECT <- dem1.ov$ASPECT; d$CONVG <- dem1.ov$CONVG; d$LS <- dem1.ov$LS
243 d$SAGAWI <- dem1.ov$SAGAWI;d$TWI <- dem1.ov$TWI; d$INSOLAT <- dem1.ov$INSOLAT
244 d$DIRECT <- dem1.ov$DIRECT; d$DIFFUS <- dem1.ov$DIFFUS
245 d$STREAMP <- dem1.ov$STREAMP; d$GEOLTYPE <- dem1.ov$GEOLTYPE
246 d$GEOL1 <- dem1.ov$GEOL1; d$GEOL2 <- dem1.ov$GEOL2; d$GEOL4 <- dem1.ov$GEOL4
247 d$GEOL5 <- dem1.ov$GEOL5; d$GEOPPR <- dem1.ov$GEOPPR
248 d$GEOPPR12 <- dem1.ov$GEOPPR12; d$GEOPPR18 <- dem1.ov$GEOPPR18
249 d$GEOPPR22 <- dem1.ov$GEOPPR22; d$GEOPPR470 <- dem1.ov$GEOPPR470
250 d$GEOPPR1233 <- dem1.ov$GEOPPR1233; d$GEOPPR1465 <- dem1.ov$GEOPPR1465
251
252 rm(i,j,m,n,n.db,x,d.db,dem1.ov,d.all,d.alr,d.comp)
253
254 proj4string(d) <-
255   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
256 proj4string(dem1) <-
257   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
258
259 pts <- list("sp.points", d, pch = 19, col = "black", cex = .6, alpha = 1)
260
261 catchm <- readOGR(dsn = "phd_calc_input", layer = "costara_catchm_v1")
262 proj4string(catchm) <-
263   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
264
265 dem2 <- dem1; asdf <- which(!is.na(over(dem2, catchm))); dem3 <- dem2[asdf,]
266
```

```
cat <- fortify(catchm, region = "NAME")
pt <- as.data.frame(pts)
dem <- as.data.frame(dem3)

v.interest <- c("ELEV", "x", "y")
dem.gamma <- dem[,v.interest]

gammaELEV.min <- min(dem.gamma$ELEV, na.rm = TRUE); gammaELEV.min # 84
gammaELEV.max <- max(dem.gamma$ELEV, na.rm = TRUE); gammaELEV.max # 271

dem.gamma$ELEVCUTV <- cut(dem.gamma$ELEV, breaks = seq(84,273, by = 21))
levels(dem.gamma$ELEVCUTV) <- c("84 - 105", "105 - 126", "126 - 147",
  "147 - 168", "168 - 189", "189 - 210", "210 - 231", "231 - 252", "252 +")

map.ELEV <- ggplot(cat, aes(long, lat)) +
  geom_raster(aes(x, y, fill = ELEVCUTV), data = dem.gamma) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1.5,
    data = pt, shape = 19) +
  geom_polygon(size = .5, linetype = "dashed", color = "black",
    fill = "grey40", alpha = 0) +
  coord_equal() +
  theme_bw(base_size = 9, base_family = "Helvetica") +
  scale_fill_manual(
    guide = guide_legend(direction = "horizontal", title.position = "top",
      nrow = 2),
    name = "Elevation in m", values = rev(brewer.pal(9,"Spectral"))) +
  labs(x = "\nUTM-E/m", y = "UTM-N/m\n") +
  scale_y_continuous(breaks = c(4361000,4363000)) +
  scale_x_continuous(breaks = c(508000,510000,512000,514000)) +
  theme(axis.text.y = element_text(angle = 90, hjust = .5),
    legend.direction = "horizontal", legend.position = "bottom",
    legend.text = element_text(size = 14), #legend.key.width = unit(42, "pt"),
    legend.key.width = unit(21, "pt"), #legend.text.align = .5,
    legend.key.height = unit(21, "pt"),
    legend.title = element_text(size = 14),
    axis.text = element_text(size = 14), axis.title = element_text(size = 14))

ggsave(paste(path.fig, "costara_ELEV_v9.pdf", sep = "/"), map.ELEV,
  width = 8.27, height = 7.57)
print(map.ELEV)
dev.off()

save.image(paste(getwd(), "phd_calc_input/cos_basis20.RData", sep = "/"))

# end of script, costara_dta20_vDiss.R, 24.07.2012
```

.4.11 costara_eda_vDiss.R

```
#
```

```
################################################################################
####  Exploratory data analysis of soil textural fractions, Rio di Costara  ####
################################################################################
#
## last update: 30.11.2014

# numerical and graphical summaries (of target variables)

# author: Michael Blaschek (blaschek@geographie.uni-kiel.de)

rm(list = ls())

library(xtable); library(lattice); library(RColorBrewer)
# R v3.0.2, xtable_1.7-3, lattice_0.20-24, RColorBrewer_1.0-5

load(paste(getwd(), "phd_calc_input/cos_basis20.RData", sep = "/"))
path.fig <- paste(getwd(), "phd_calc_figures", sep = "/")

d.c <- d[-c(122:145,181:197),] # n = 156
d.v <- d[c(122:145,181:197),] # n = 41

n <- c(dim(d)[1], dim(d.c)[1], dim(d.v)[1]); n


#-------------------------------------------------------------------------------
## Summary statistics
#-------------------------------------------------------------------------------

source("descr_statistics.R")

datasets <- list(d@data, d.c@data, d.v@data)
z <- c("CLAY", "SILT", "SAND", "CLAYalr", "SILTalr")

descr.d <- summary_stats(data = datasets, vars = z, export.latex = TRUE,
  path.to = c(paste(getwd(), "phd_calc_out", paste0(
    "costara_descr", n[1], "_v1", ".tex"), sep = "/"),
    paste(getwd(), "phd_calc_out", paste0(
      "costara_descr", n[2], "_v1", ".tex"), sep = "/"),
     paste(getwd(), "phd_calc_out", paste0(
       "costara_descr", n[3], "_v1", ".tex"), sep = "/")))

# Shapiro-Wilk normality test:
j <- c("W", "p"); shap.d <- NULL
for (q in 1:length(datasets)) {
 dd <- datasets[[q]]; k <- 1
 shap <- data.frame(z, 1:length(z), 1:length(z))
 names(shap) <- c("Target variables", j) # variable names
 for (i in z) {
  shap[k,"W"] <- round(shapiro.test(dd[,i])$statistic, 2)
  shap[k,"p"] <- round(shapiro.test(dd[,i])$p.value, 3)
```

```
  k <- k + 1
 }
 shap.d[[q]] <- shap
}
rm(i,j,k,q,dd,shap,datasets) # new objects from this section: descr.d, shap.d


#-------------------------------------------------------------------------------
## Check the representativity of d.v and d.c for each other
#-------------------------------------------------------------------------------

x <- as.matrix(d.c@data[c("CLAY","SILT","SAND")])
y <- as.matrix(d.v@data[c("CLAY","SILT","SAND")])
mx <- apply(x, 2, mean); my <- apply(y, 2, mean)
sx <- cov(x); sy <- cov(y); p <- dim(x)[2]
s.pooled <- ((n[2] - 1) * sx + (n[3] - 1) * sy) / (n[2] - 1 + n[3] - 1)

# two-sample Hotelling's T-squared test for differences in 2 multivariate means:
D2 <- t(mx - my) %*% solve(s.pooled) %*% (mx - my) * n[2] * n[3]/(n[2] + n[3])
# transforming Hotelling's T-square statistic into an F-statistic:
m <- (n[2] + n[3] - dim(x)[2] - 1) / (dim(x)[2] * (n[2] + n[3] - 2)); m
hF <- m * D2; hF
# calculating p-value for the given F-statistic and degrees of freedom:
p.value <- pf(hF, dim(x)[2], n[2] + n[3] - dim(x)[2] - 1, lower.tail = FALSE)

# h0:  1 =  2 , F-statistic = 2.2925, p.value (X > hF) = 0.07942

# Bartlett's test of homogeneity of variance-covariance matrices:
bcf <- 1 + (2*p^2 + 3*p - 1)/(6*(p + 1)) *
  ((1/(n[2] - 1)) + (1/(n[3] - 1)) - (1/(n[2] + n[3] - 2))); bcf
bL <- (1/bcf) * ((n[2] + n[3] - 2) * log(det(s.pooled)) -
  (n[2] - 1) * log(det(sx)) - (n[3] - 1) * log(det(sy))); bL

p.value2 <- pchisq(bL, p * (p + 1)/2, lower.tail = FALSE); p.value2
# h0: Sigma1 = Sigma2, test-statistic = 6.425532, p.value = 0.3772458

rm(D2,s.pooled,sx,sy,x,y,bcf,m,mx,my,p)


#-------------------------------------------------------------------------------
## Exploratory graphics
#-------------------------------------------------------------------------------

# density histograms, untransformed contents of fractions:
# class width based on Scott's rule (1979)
pdf(paste(path.fig, "costara_histogr_soilsep_v1.pdf", sep = "/"),
  width = 8.75, height = 8.75, pointsize = 13)
par(mfrow = c(3,3), las = 1, cex.main = 1, font.main = 2,
  mar = c(4,3,2,0)+.1, oma = c(.5,.5,.5,.5), mgp = c(2,1,0))

```

```
# CLAY for all samples (n = 197), Rio di Costara:
hk <- 3.49 * sd(d$CLAY) * n[1]^(-1/3) # hk = 6.358
hd <- hist(d$CLAY, freq = F, col = "grey", xlim = c(0,70), ylim = c(0,0.05),
  breaks = seq(0,66,6), xlab = "Clay content in %", axes = FALSE, ylab = "",
  main = "a) Clay, N = 197", sub = paste("Shapiro-Wilk test: W = ",
    shap.d[[1]][1,2], ", p = ", shap.d[[1]][1,3], sep = ""))
curve(dnorm(x, mean(d$CLAY), sd(d$CLAY)), add = T, lwd = 2, lty = 2)
axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 1, lwd = 0, line = -.35); axis(side = 2, lwd = 0, line = -.35)
box(which = "plot", lty = "solid", col = "black")

# SILT for all samples (n = 197), Rio di Costara:
hk <- 3.49 * sd(d$SILT) * n[1]^(-1/3) # hk = 4.008
hd <- hist(d$SILT, freq = F, col = "grey", xlim = c(0,56), ylim = c(0,0.08),
  breaks = seq(4,52,4), xlab = "Silt content in %", axes = FALSE, ylab = "",
  main = "b) Silt, N = 197", sub = paste("Shapiro-Wilk test: W = ",
    shap.d[[1]][2,2], ", p = ", shap.d[[1]][2,3], sep = ""))
curve(dnorm(x, mean(d$SILT), sd(d$SILT)), add = T, lwd = 2, lty = 2)
axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 1, lwd = 0, line = -.35); axis(side = 2, lwd = 0, line = -.35)
box(which = "plot", lty = "solid", col = "black")

# SAND for all samples (n = 197), Rio di Costara:
hk <- 3.49 * sd(d$SAND) * n[1]^(-1/3) # hk = 7.327
hd <- hist(d$SAND, freq = F, col = "grey", xlim = c(0,100), ylim = c(0,0.04),
  breaks = seq(7.5,97.5,7.5), xlab = "Sand content in %", axes = F, ylab = "",
  main = "c) Sand, N = 197", sub = paste("Shapiro-Wilk test: W = ",
    shap.d[[1]][3,2], ", p = ", shap.d[[1]][3,3], sep = ""))
curve(dnorm(x, mean(d$SAND), sd(d$SAND)), add = T, lwd = 2, lty = 2)
axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 1, lwd = 0, line = -.35); axis(side = 2, lwd = 0, line = -.35)
box(which = "plot", lty = "solid", col = "black")

# CLAY for calibration set (n = 156), Rio di Costara:
hk <- 3.49 * sd(d.c$CLAY) * n[1]^(-1/3) # hk = 6.0765
hd <- hist(d.c$CLAY, freq = F, col = "grey", xlim = c(0,70), ylim = c(0,0.05),
  breaks = seq(0,66,6), xlab = "Clay content in %", axes = FALSE, ylab = "",
  main = "d) Clay, N = 156", sub = paste("Shapiro-Wilk test: W = ",
    shap.d[[2]][1,2], ", p = ", shap.d[[2]][1,3], sep = ""))
curve(dnorm(x, mean(d.c$CLAY), sd(d.c$CLAY)), add = T, lwd = 2, lty = 2)
axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 1, lwd = 0, line = -.35); axis(side = 2, lwd = 0, line = -.35)
box(which = "plot", lty = "solid", col = "black")

# SILT for calibration set (n = 156), Rio di Costara:
hk <- 3.49 * sd(d.c$SILT) * n[1]^(-1/3); hk # hk = 3.8683
```

```
hd <- hist(d.c$SILT, freq = F, col = "grey", xlim = c(0,56), ylim = c(0,0.08),
  breaks = seq(4,52,4), xlab = "Silt content in %", axes = FALSE, ylab = "",
  main = "e) Silt, N = 156", sub = paste("Shapiro-Wilk test: W = ",
    shap.d[[2]][2,2], ", p = ", shap.d[[2]][2,3], sep = ""))
curve(dnorm(x, mean(d.c$SILT), sd(d.c$SILT)), add = T, lwd = 2, lty = 2)
axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 1, lwd = 0, line = -.35); axis(side = 2, lwd = 0, line = -.35)
box(which = "plot", lty = "solid", col = "black")

# SAND for calibration set (n = 156), Rio di Costara:
hk <- 3.49 * sd(d.c$SAND) * n[1]^(-1/3) # hk = 6.7976
hd <- hist(d.c$SAND, freq = F, col = "grey", xlim = c(0,100), ylim = c(0,0.04),
  breaks = seq(7.5,97.5,7.5), xlab = "Sand content in %", axes = F, ylab = "",
  main = "f) Sand, N = 156", sub = paste("Shapiro-Wilk test: W = ",
    shap.d[[2]][3,2], ", p = ", shap.d[[2]][3,3], sep = ""))
curve(dnorm(x, mean(d.c$SAND), sd(d.c$SAND)), add = T, lwd = 2, lty = 2)
axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 1, lwd = 0, line = -.35); axis(side = 2, lwd = 0, line = -.35)
box(which = "plot", lty = "solid", col = "black")

# CLAY for validation set (n = 41), Rio di Costara:
hk <- 3.49 * sd(d.v$CLAY) * n[1]^(-1/3) # hk = 7.4094
hd <- hist(d.v$CLAY, freq = F, col = "grey", xlim = c(0,70), ylim = c(0,0.05),
  breaks = seq(0,66,6), xlab = "Clay content in %", axes = FALSE, ylab = "",
  main = "g) Clay, N = 41", sub = paste("Shapiro-Wilk test: W = ",
    shap.d[[3]][1,2], ", p = ", shap.d[[3]][1,3], sep = ""))
curve(dnorm(x, mean(d.v$CLAY), sd(d.v$CLAY)), add = T, lwd = 2, lty = 2)
axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 1, lwd = 0, line = -.35); axis(side = 2, lwd = 0, line = -.35)
box(which = "plot", lty = "solid", col = "black")

# SILT for validation set (n = 41), Rio di Costara:
hk <- 3.49 * sd(d.v$SILT) * n[1]^(-1/3) # hk = 4.4730
hd <- hist(d.v$SILT, freq = F, col = "grey", xlim = c(0,56), ylim = c(0,0.08),
  breaks = seq(4,52,4), xlab = "Silt content in %", axes = FALSE, ylab = "",
  main = "h) Silt, N = 41", sub = paste("Shapiro-Wilk test: W = ",
    shap.d[[3]][2,2], ", p = ", shap.d[[3]][2,3], sep = ""))
curve(dnorm(x, mean(d.v$SILT), sd(d.v$SILT)), add = T, lwd = 2, lty = 2)
axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 1, lwd = 0, line = -.35); axis(side = 2, lwd = 0, line = -.35)
box(which = "plot", lty = "solid", col = "black")

# SAND for validation set (n = 41), Rio di Costara:
hk <- 3.49 * sd(d.v$SAND) * n[1]^(-1/3) # hk = 9.0885
hd <- hist(d.v$SAND, freq = F, col = "grey", xlim = c(0,100), ylim = c(0,0.04),
  breaks = seq(7.5,97.5,7.5), xlab = "Sand content in %", axes = F, ylab = "",
```

```
  main = "i) Sand, N = 41", sub = paste("Shapiro-Wilk test: W = ",
    shap.d[[3]][3,2], ", p = ", shap.d[[3]][3,3], sep = ""))
curve(dnorm(x, mean(d.v$SAND), sd(d.v$SAND)), add = T, lwd = 2, lty = 2)
axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 1, lwd = 0, line = -.35); axis(side = 2, lwd = 0, line = -.35)
box(which = "plot", lty = "solid", col = "black")
dev.off(); rm(hk,hd)

# density histograms, alr-transformed contents of fractions:
pdf(paste(path.fig, "costara_histogr_alr156_v1.pdf", sep = "/"),
  width = 8.27, height = 4.19, pointsize = 13)
par(mfrow = c(1,2), las = 1, cex.main = 1, font.main = 2,
  mar = c(3,3,2,0)+.1, oma = c(.5,.5,.5,.5), mgp = c(2,1,0))

# alr CLAY for calibration set (n = 156), Rio di Costara:
hk <- 3.49 * sd(d.c$CLAYalr) * n[1]^(-1/3) # hk = 0.3675
#txt <- paste0("Min = ", descr.d[[2]][4,2], "\nMax = ", descr.d[[2]][4,3],
#  "\n1st Qu. = ", descr.d[[2]][4,4], "\nMedian = ", descr.d[[2]][4,5],
#  "\n3rd Qu. = ", descr.d[[2]][4,6])
hd <- hist(d.c$CLAYalr, freq = TRUE, col = "grey", xlim = c(-2.5,2),
  ylim = c(0,80), breaks = seq(-2.5,2,.5), axes = FALSE, ylab = "Counts",
  main = "a) alr Clay, N = 156", xlab = paste("Shapiro-Wilk test: W = ",
    shap.d[[2]][4,2], ", p = ", shap.d[[2]][4,3], sep = ""))
axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 1, lwd = 0, line = -.35); axis(side = 2, lwd = 0, line = -.35)
#text(.75, 45, txt, pos = 4, cex = .75)
box(which = "plot", lty = "solid", col = "black")

# alr SILT for calibration set (n = 156), Rio di Costara:
hk <- 3.49 * sd(d.c$SILTalr) * n[1]^(-1/3); hk # hk = 0.2811
#txt2 <- paste0("Min = ", descr.d[[2]][5,2], "\nMax = ", descr.d[[2]][5,3],
#  "\n1st Qu. = ", descr.d[[2]][5,4], "\nMedian = ", descr.d[[2]][5,5],
#  "\n3rd Qu. = ", descr.d[[2]][5,6])
hd <- hist(d.c$SILTalr, freq = TRUE, col = "grey", xlim = c(-2.5,2),
  ylim = c(0,80), breaks = seq(-2.5,2,.5), axes = FALSE, ylab = "",
  main = "b) alr Silt, N = 156", xlab = paste("Shapiro-Wilk test: W = ",
    shap.d[[2]][5,2], ", p = ", shap.d[[2]][5,3], sep = ""))
axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 1, lwd = 0, line = -.35); axis(side = 2, lwd = 0, line = -.35)
#text(.75, 45, txt2, pos = 4, cex = .75)
box(which = "plot", lty = "solid", col = "black")
dev.off(); rm(hk,hd)


# box-and-whisker plot, Rio di Costara:
# build upon code published under CC BY-SA by
# Tim Appelhans: Creating publication quality graphs in R
```

```
# http://teachpress.environmentalinformatics-marburg.de/2013/07/
#   creating-publication-quality-graphs-in-r-7/ (visited on 06/10/14)
d$GEOLTYP <- d$GEOLTYPE
d$GEOLTYP[which(d$GEOLTYPE == 4)] <- 5
d$GeolFac <- factor(d$GEOLTYP, labels = c("Quaternary deposits",
  "Oligo-Miocene sediments", "Paleozoic basement"))
x <- d@data[c("GeolFac","CLAY","SILT","SAND")]
names(x) <- c("geolt", "text1", "text2", "text3")
t = reshape(x, direction = "long",  varying = 2:4, sep = "")
t$textcl[which(t$time == 1)] <- "Clay"; t$textcl[which(t$time == 2)] <- "Silt"
t$textcl[which(t$time == 3)] <- "Sand"

pdf(paste(path.fig, "costara_boxplot_soilsep_v1.pdf", sep = "/"),
  width = 10, height = 4, pointsize = 14)
bpl <- bwplot(text ~ factor(textcl, levels = c("Clay","Silt","Sand")) |
    as.character(geolt), data = t, layout = c(3,1),
  main = "", xlab = "Soil textural fractions", ylab = "Content in %", asp = 1,
  ylim = c(0,100), coef = 1.5, par.strip.text = list(cex = 1),
  scales = list(y = list(at = c(0,20,40,60,80,100))))

th <- trellis.par.get()
th$box.dot$pch <- "|"
th$box.rectangle$col <- "black"; th$box.rectangle$lwd <- 2
th$box.rectangle$fill <- brewer.pal(3, "Dark2")
th$box.umbrella$lty <- 1; th$box.umbrella$col <- "black"
th$plot.symbol$col <- "grey40"; th$plot.symbol$pch <- "*"
th$plot.symbol$cex <- 2; th$strip.background$col <- "grey80"
th$par.xlab.text$cex <- 1; th$par.ylab.text$cex <- 1
th$fontsize$text <- 14; th$axis.text$cex <- 1
th$axis.components$left$tck <- .75; th$axis.components$right$tck <- .75
th$layout.widths$left.padding <- 0; th$layout.widths$right.padding <- 0
th$layout.heights$top.padding <- 0; th$layout.heights$bottom.padding <- 0

bpl.upd <- update(bpl, par.settings = th)
print(bpl.upd)
dev.off()
rm(t,x,th,bpl,bpl.upd,n,z)
# remaining objects: d, d.c, d.v, dem1, descr.d, shap.d, path.fig

# investigate distributions of target variables by geological unit:
j <- c("Median", "Skewness", "Octile skew")
z <- c("CLAY", "SILT", "SAND", "CLAYalr", "SILTalr")
g <- c("Oligo-Miocene sediments", "Paleozoic basement", "Quaternary deposits")
descr.dg <- NULL
for (i in 1:length(z)) {
 k <- 1; a <- 1
 descr <- data.frame(g, 1:length(g), 1:length(g), 1:length(g))
 names(descr) <- c("Target variables", j) # variable names
 for (q in g) {
  dd <- d@data[which(d$GeolFac == q),z[i]]
```

```
  descr[k,"Median"] <- round(median(dd), 2)
  descr[k,"Skewness"] <-
     round((sum((dd - mean(dd))^3) / length(dd)) / sd(dd)^3, 2)
  a <- quantile(
    dd, probs = c(.125,.5,.875), na.rm = FALSE, names = TRUE, type = 7)
  descr[k,"Octile skew"] <- round(((a[3]-a[2]) - (a[2]-a[1]))/(a[3]-a[1]), 2)
  k <- k + 1
 }
 descr.dg[[i]] <- descr
}
rm(a,dd,descr,i,j,k,q)

# end of script, costara_eda_vDiss.R, 10.06.2014
```

.4.12 costara_ternaryd_vDiss.R

```
#
################################################################################
####     Ternary diagram/Triangle plot of soil texture, Rio di Costara      ####
################################################################################
#
## last update: 04.09.2014

# plot textural soil classifications --> one figure: costara_ternary_soilsep_v..

# author: Michael Blaschek (blaschek@geographie.uni-kiel.de)

rm(list = ls())

library(soiltexture); library(compositions)
# R v3.0.2, soiltexture_1.2.13, compositions_1.40-1

load(paste(getwd(), "phd_calc_input/cos_basis20.RData", sep = "/"))
path.fig <- paste(getwd(), "phd_calc_figures", sep = "/")

d.comp <- acomp(d@data, parts = c("CLAY", "SILT", "SAND"), total = 100)
d.alr <- alr(d.comp) # additive logratio transform!
d$CLAYalr <- d.alr[,1]; d$SILTalr <- d.alr[,2]

n <- dim(d)[1]


#-------------------------------------------------------------------------------
## Ternary plot based on USDA/FAO classification
#-------------------------------------------------------------------------------

# using the 2000-63-2 system for particle-size fractions
# FAO 06: Guidelines for soil description
pdf(paste(path.fig, "costara_ternary_soilsep_v2.pdf", sep = "/"),
```

```
  width = 8.5, height = 9.25, pointsize = 14)
stp <- TT.plot(class.sys = "USDA.TT", tri.data = d@data,
  pch = "*", cex = 1.3, cex.axis = 1, cex.lab = 1, main = "",
  class.lab.show = "none", class.p.bg.col = F, class.line.col = "gray50",
  new.mar = c(2,1,0,0)+.1, grid.show = F, frame.bg.col = "white",
  lwd.lab = 1.2, lwd.axis = 1.2, font.lab = 1, font.axis = 1,
  css.lab = c("% Clay 0-2   m ", "% Silt 2-63   m ", "% Sand 63-2000   m "))

# add compositional mean:
st.cm <- mean(d.comp); st.cm <- as.data.frame(rbind(st.cm[1:3] * 100))
names(st.cm) <- c("CLAY", "SILT", "SAND")
stp.m =
  TT.points(tri.data = st.cm, geo = stp, pch = "*", cex = 1.3, col = "red")
# --> on average: clay loam (equals German: Lts = sandig-toniger Lehm)

# add compositional mahalanobis distances as contour lines:
mahal <- TT.mahalanobis(geo = stp, n = 100, tri.data = d@data, alr = TRUE)
stp.mahal <- TT.contour(x = mahal, geo = stp, main = "", lwd = 1, col = "blue",
  levels = c(0.5,1,2,4,8,16,32), add = TRUE, lty = 2, labcex = .8)
dev.off()

# extract soil texture class for each sample location:
stc <- TT.points.in.classes(d@data, class.sys = "USDA.TT", PiC.type = "n")
stcl <- 0
for (i in 1:n)
  stcl[i] <- attributes(stc)$dimnames[[2]][which(stc[i,] == 1)]; stcl

d$USDATEXTCL <- stcl
table(d$USDATEXTCL) # --> 68 points inside ClLo = clay loam class

rm(st.cm,stp.m,mahal,n,stp,stp.mahal,d.alr,d.comp,i,stcl,stc)

# end of script, costara_ternaryd_vDiss.R, 14.12.2011
```

.4.13 costara_corFA_vDiss.R

```
#
################################################################################
####  Correlation and factor analysis + variable selection, Rio di Costara  ####
################################################################################
#
## last update: 30.11.2014

# scatterplot matrix of target variables and relief parameter
# (co-)variable selection for interpolation
# factor analysis of mixed data

# author: Michael Blaschek (blaschek@geographie.uni-kiel.de)

```

```
14 rm(list = ls())
15
16 library(caret); library(FactoMineR)
17 # R v3.0.2, caret_6.0-22, FactoMineR_1.26
18
19 load(paste(getwd(), "phd_calc_input/cos_basis20.RData", sep = "/"))
20 path.fig <- paste(getwd(), "phd_calc_figures", sep = "/")
21
22 n <- dim(d)[1]
23
24
25 #------------------------------------------------------------------------------
26 ## Correlation coefficients and scatterplot-matrix
27 #------------------------------------------------------------------------------
28
29 source("pearsons_cor_coeffs.R")
30
31 j <- c("ELEV", "SLOPE", "PROFC", "PLANC", "ASPECT",
32   "CONVG", "TWI", "SAGAWI", "INSOLAT")
33 z <- c("CLAY", "SILT", "SAND")
34
35 #cor.ll <- pearsons_corr(data = d@data, covars = j, targets = z)
36 cor.ll <- pearsons_corr(data = d@data, covars = j, targets = z,
37   scatterplot.matrix = TRUE, cex.lab = 1.5,
38   path.to = paste(path.fig, "costara_scatterpl_v1.pdf", sep = "/"))
39
40 # LS+SLOPE, ELEV+DIFFUS, DIRECT+INSOLAT are highly correlated (r = .97, 1, 1)
41 # --> LS, DIFFUS and DIRECT are removed from further analysis
42 # STREAMP: uncomfortable distribution (very strong outliers) --> removed
43
44
45 #------------------------------------------------------------------------------
46 ## Variable selection
47 #------------------------------------------------------------------------------
48
49 # repeat correlation analysis for alr-transformed target variables:
50 # logratio transform to account for the compositional character of the targets:
51
52 z2 <- c("CLAYalr", "SILTalr")
53 cor.l.alr <- pearsons_corr(data = d@data, covars = j, targets = z2)
54
55 j.signif <- NULL; signif.explan <- NULL; signif.explan.corr.matrix <- NULL
56 highly.corr.signif.explan <- NULL; nm.signif.uncorr.explan <- list(NULL)
57 for (i in 1:length(z2)) {
58   # select only the significant covariates:
59   j.signif <- as.character(
60     cor.l.alr[[i]]$COVAR[which(cor.l.alr[[i]]$SIGNIFICANCE != "-")])
61   # determine highly correlated variables
62   #   and suggests those for removal with the largest mean absolute correlation:
63   signif.explan <- d[,j.signif]
```

```
  signif.explan.corr.matrix <- cor(signif.explan@data)
  highly.corr.signif.explan <-
    findCorrelation(signif.explan.corr.matrix, cutoff = .65)
  # keep those that are not selected for removal:
  nm.signif.uncorr.explan[[i]] <-
    names(signif.explan@data[-highly.corr.signif.explan])
}
# --> critical corr-level: 0.65, critical significance level: p.value < 0.05

nm.signif.uncorr.explan
# CLAYalr --> ELEV, TWI, DIRECT
# SILTalr --> ASPECT, SAGAWI, DIFFUS, INSOLAT

# --> ELEV, SAGAWI, INSOLAT (selection based on correlations)

# + GEOLOGY (as dummy variables):
nm.signif.uncorr.explan2 <- c("ELEV", "SAGAWI", "INSOLAT")
prd.in <- c(nm.signif.uncorr.explan2,
  "GEOPPR12", "GEOPPR22", "GEOPPR470", "GEOPPR1465") # defined predictors

rm(cor.c,cor.test.p,i,j,k,q,n,z2,j.signif,signif.explan,
  highly.corr.signif.explan,signif.explan.corr.matrix,nm.signif.uncorr.explan)
# remaining objects: cor.l, nm.signif.uncorr.explan2, prd.in, path.fig, d, dem1


#-------------------------------------------------------------------------------
## Factor analysis for mixed data
#-------------------------------------------------------------------------------

t.dem1.geoppr <- table(dem1$GEOPPR)
# --> 7,9,12,18,19,22,52,470,472,824,833,1233,1465,2000
dem1$GEOPPRz <- dem1$GEOPPR
# summarize geological units with less than 500 pixels as -99:
under500 <- NULL; k <- 1
for (i in 1:length(t.dem1.geoppr)) {
  if(t.dem1.geoppr[[i]] <= 500) {
    under500[k] <- as.numeric(names(t.dem1.geoppr)[i]); k <- k + 1
  }
}
dem1$GEOPPRz[which(!is.na(match(dem1$GEOPPRz, under500)))] <- -99
t.dem1.geopprz <- table(dem1$GEOPPRz)

dem1$GEOPPRz <- as.factor(dem1$GEOPPRz) # categorical variable needed as factor
datd <- dem1@data[,c(1:3,5:8,13,26)] # no ASPECT (as it contains NA-values)
famd <- FAMD(datd, ncp = 15, graph = FALSE) # this takes some time

# selection of the right number of factors:
# after usual Kaiser-Guttman rule (cutoff = 1):
cv <- 1
nf <- which(famd$eig[,1] >= cv) # --> first six factors are significant
```

```
114
115 # critical value after Karlis, Saporta and Spinakis (2003):
116 p <- sum(famd$eig[,1]) # the first 14 factors explain 100%
117 ni <- dim(datd)[1] # number of instances
118 cv <- 1 + 1.65 * sqrt((p - 1)/(ni - 1))
119
120 nf <- which(famd$eig[,1] >= cv) # --> first five factors are significant
121
122 # variable coordinates:
123 # describes the influence of variables in the determination of the factors
124 famd$var$coord[,1:6] # square of correlation coefficient
125 # --> dim.1 influenced by SAGAWI, TWI, SLOPE, GEOPPRz, ELEV
126 # --> dim.2 influenced by PROFC, CONVG, PLANC
127 # --> dim.3 influenced by GEOPPRz, INSOLAT
128 # --> dim.4/dim.5/dim.6 influenced by GEOPPRz
129
130 # convert the derived factors into grids:
131 famd.comps <- as.data.frame(famd$ind$coord)
132 # insert grid index:
133 dem <- dem1
134 dem$nrs <- seq(1, length(dem@data[[1]]))
135 dem.pnt <- as(dem["nrs"], "SpatialPointsDataFrame")
136 # mask NA grid nodes:
137 maskpoints <- as.numeric(attr(famd$ind$coord, "dimnames")[[1]])
138 # attach coordinates:
139 famd.comps$X <- dem.pnt@coords[maskpoints,1]
140 famd.comps$Y <- dem.pnt@coords[maskpoints,2]
141 coordinates(famd.comps) <- ~ X + Y
142
143 # convert to a grid:
144 gridded(famd.comps) <- TRUE
145 famd.comps <- as(famd.comps, "SpatialGridDataFrame")
146 proj4string(famd.comps) <- dem@proj4string
147 names(famd.comps)
148
149 dem@data$FAMDa <- famd.comps@data$Dim.1; dem@data$FAMDb <- famd.comps@data$Dim.2
150 dem@data$FAMDc <- famd.comps@data$Dim.3; dem@data$FAMDd <- famd.comps@data$Dim.4
151 dem@data$FAMDe <- famd.comps@data$Dim.5; dem@data$FAMDf <- famd.comps@data$Dim.6
152
153 # overlay the points and factors:
154 dem.ov <- over(d, dem)
155 d@data$FAMDa <- dem.ov$FAMDa; d@data$FAMDb <- dem.ov$FAMDb
156 d@data$FAMDc <- dem.ov$FAMDc; d@data$FAMDd <- dem.ov$FAMDd
157 d@data$FAMDe <- dem.ov$FAMDe; d@data$FAMDf <- dem.ov$FAMDf
158
159 rm(dem.pnt,dem.ov,famd.comps,maskpoints)
160
161 # end of script, costara_corFA_vDiss.R, 30.08.2013
```

.4.14 costara_explor_sp_vDiss.R

```
#
################################################################################
####  Exploratory spatial data analysis of soil separates, Rio di Costara  #####
################################################################################
#
## last update: 09.12.2014

# trend detection and variography (of target variables)

# author: Michael Blaschek (blaschek@geographie.uni-kiel.de)

rm(list = ls())

library(rgdal); library(RColorBrewer); library(sp); library(gstat)
# R v3.0.2, rgdal_0.8-15, RColorBrewer_1.0-5, sp_1.0-14, gstat_1.0-16

load(paste(getwd(), "phd_calc_input/cos_basis20.RData", sep = "/"))
path.fig <- paste(getwd(), "phd_calc_figures", sep = "/")

n <- dim(d)[1]


#-------------------------------------------------------------------------------
## Spatial evaluation of stratification used in soil sampling
#-------------------------------------------------------------------------------

# build upon code published by
# Richard E. Plant: Spatial data analysis in ecology and agriculture using R
# Taylor & Francis Group, 2012

s <- readOGR(dsn = "phd_calc_input/costara_geology",
  layer = "geolmapUNDdtaintersect_1ha1_costara_ENDCLdissolve")

# relative size of each (geological) strata:
sg.1 <- s[s$endclass > 100 & s$endclass < 200,]
sg.2 <- s[s$endclass > 200 & s$endclass < 400,]
sg.4 <- s[s$endclass > 400 & s$endclass < 500,]
sg.5 <- s[s$endclass > 500 & s$endclass < 600,]

# summarize area of each (geological) strata:
quatern.area <- 0; oligomeso.area <- 0;
paleozintru.area <- 0; paleozmetam.area <- 0
for (i in 1:length(sg.1))
  quatern.area <- quatern.area + slot(slot(sg.1,"polygons")[[i]],"area")
for (i in 1:length(sg.2))
  oligomeso.area <- oligomeso.area + slot(slot(sg.2,"polygons")[[i]],"area")
for (i in 1:length(sg.4))
  paleozintru.area <- paleozintru.area + slot(slot(sg.4,"polygons")[[i]],"area")
```

```
for (i in 1:length(sg.5))
  paleozmetam.area <- paleozmetam.area + slot(slot(sg.5,"polygons")[[i]],"area")

# calculate total area:
total.area <- 0
for (i in 1:length(s))
  total.area <- total.area + slot(slot(s, "polygons")[[i]], "area")

frac.q <- quatern.area/total.area; frac.om <- oligomeso.area/total.area
frac.pi <- paleozintru.area/total.area; frac.pm <- paleozmetam.area/total.area

# double-check:
total.area - (quatern.area + oligomeso.area +
  paleozmetam.area + paleozintru.area) # should be zero
sum(frac.q, frac.om, frac.pi, frac.pm) # should be 1

sampled.size <- 197
des.size <- round(c(frac.q, frac.om, frac.pi, frac.pm) * sampled.size, 0)
# --> N(Quaternary) = 104, N(OligoMiocene) = 69, N(Paleozoic) = 24 (1 intrusive)

# actual sample points per (geological) strata:
n.geol <- by(d$CLAY, d$GEOLTYPE, length)
# --> N(Quaternary) = 92, N(Oligo-Miocene) = 77, N(Paleozoic) = 28 (3 intrusive)

rm(s,sg.1,sg.2,sg.4,sg.5,total.area,quatern.area,oligomeso.area,i,
  paleozintru.area,paleozmetam.area,sampled.size,frac.q,frac.om,frac.pi,frac.pm)
# remaining objects from this section: des.size, n.geol


#--------------------------------------------------------------------------------
## Trend detection - Postplots
#--------------------------------------------------------------------------------

# build upon code published by
# Richard E. Plant: Spatial data analysis in ecology and agriculture using R
# Taylor & Francis Group, 2012

catchm <- readOGR(dsn = "phd_calc_input", layer = "costara_catchm_v1")

d$color[which(d$GEOLTYPE == 1)] <- "#1B9E77"
d$color[which(d$GEOLTYPE == 2)] <- "#D95F02"
d$color[which(d$GEOLTYPE == 4)] <- "#E7298A"
d$color[which(d$GEOLTYPE == 5)] <- "#7570B3"

geol.units <- c("Quaternary deposits", "Oligo-Miocene sedimentary deposits",
  "Paleozoic metamorphic basement", "Paleozoic intrusive rocks")

pdf(paste(path.fig, "costara_postpl_soilsep_v99.pdf", sep = "/"),
  width = 15, height = 14, pointsize = 22)
layout(matrix(c(1,1,2,2,3,3,4,4), 2, 4, byrow = TRUE))
```

```
# layout.show(3)
par(mar = c(3,3,2,0)+.1, oma = c(.5,.5,.5,.5), mgp = c(2,1,0),
  cex.main = 1, font.main = 2)

# clay, n = 197:
plot(catchm, bg = NA, lwd = 1, lty = 2, axes = FALSE)
abline(h = seq(4359000,4365000,2000),
  v = seq(508000,514000,2000), lty = 1, col = "grey80")
plot(d, pch = 20, col = d$color, asp = 1, axes = FALSE, add = TRUE,
  cex = d$CLAY * 3/max(d$CLAY))
axis(side = 1, tck = -.02, at = seq(508000,514000,2000), labels = NA)
axis(side = 2, tck = -.02, at = seq(4361000,4363000,2000), labels = NA)
axis(side = 1, lwd = 0, line = -.35, at = seq(508000,514000,2000))
axis(side = 2, lwd = 0, line = -.35, at = seq(4361000,4363000,2000))
title(main = "a) Clay content in %", xlab = "UTM-E/m", ylab = "UTM-N/m")
box(which = "plot", lty = "solid", col = "black")
SpatialPolygonsRescale(layout.scale.bar(height = 0.065), plot.grid = FALSE,
  offset = c(507500,4359175), scale = 1000, fill = c("white", "black"))
text(507500,4359375, "0", cex = .75); text(508500,4359375, "1000 m", cex = .75)
text(511533, 4362047 + 135, "511", cex = .75)
text(512313 - 300, 4361110, "735", cex = .75)

# silt, n = 197:
plot(catchm, bg = NA, lwd = 1, lty = 2, axes = FALSE)
abline(h = seq(4359000,4365000,2000),
  v = seq(508000,514000,2000), lty = 1, col = "grey80")
plot(d, pch = 20, col = d$color, asp = 1, axes = FALSE, add = TRUE,
  cex = d$SILT * 3/max(d$SILT))
axis(side = 1, tck = -.02, at = seq(508000,514000,2000), labels = NA)
axis(side = 2, tck = -.02, at = seq(4361000,4363000,2000), labels = NA)
axis(side = 1, lwd = 0, line = -.35, at = seq(508000,514000,2000))
axis(side = 2, lwd = 0, line = -.35, at = seq(4361000,4363000,2000))
title(main = "b) Silt content in %", xlab = "UTM-E/m")
box(which = "plot", lty = "solid", col = "black")
SpatialPolygonsRescale(layout.scale.bar(height = 0.065), plot.grid = FALSE,
  offset = c(507500,4359175), scale = 1000, fill = c("white", "black"))
text(507500,4359375, "0", cex = .75); text(508500,4359375, "1000 m", cex = .75)

# sand, n = 197:
plot(catchm, bg = NA, lwd = 1, lty = 2, axes = FALSE, ylim = c(4360500,4363500))
abline(h = seq(4359000,4365000,2000),
  v = seq(508000,514000,2000), lty = 1, col = "grey80")
plot(d, pch = 20, col = d$color, asp = 1, axes = FALSE, add = TRUE,
  cex = d$SAND * 3/max(d$SAND))
axis(side = 1, tck = -.02, at = seq(508000,514000,2000), labels = NA)
axis(side = 2, tck = -.02, at = seq(4361000,4363000,2000), labels = NA)
axis(side = 1, lwd = 0, line = -.35, at = seq(508000,514000,2000))
axis(side = 2, lwd = 0, line = -.35, at = seq(4361000,4363000,2000))
title(main = "c) Sand content in %", xlab = "UTM-E/m", ylab = "UTM-N/m")
box(which = "plot", lty = "solid", col = "black")
```

```
149 SpatialPolygonsRescale(layout.scale.bar(height = 0.065), plot.grid = FALSE,
150   offset = c(507500,4359175), scale = 1000, fill = c("white", "black"))
151 text(507500,4359375, "0", cex = .75); text(508500,4359375, "1000 m", cex = .75)
152 plot(0, 0, type = "n", bty = "n", axes = FALSE, xlab = "", ylab = "")
153 legend("center", geol.units, cex = 1,
154   fill = brewer.pal(4, "Dark2"), bg = "white", ncol = 1)
155 dev.off()
156
157
158 #------------------------------------------------------------------------------
159 ## Variography
160 #------------------------------------------------------------------------------
161
162 h.max <- range(dist(coordinates(d)))[2] # max. distance between 2 pnts: 6559.2m
163
164 diag.bbox <-
165   sqrt((bbox(d)[1,1] - bbox(d)[1,2])^2 + (bbox(d)[2,1] - bbox(d)[2,2])^2)
166 diag.bbox/2; diag.bbox/3 # 2 after Journel and Huijbregts 78,3 = gstat's default
167 # --> distance of reliability between 2445 and 3670m
168
169 va.cutoff <- 3625; va.width <- 125
170 # classical "methods of moments" estimation using gstat-package:
171 va.cl <- variogram(CLAY ~ 1, loc = d, cutoff = va.cutoff, width = va.width)
172 vm.cl <- vgm(80, "Sph", 750, 15); vmf.cl <- fit.variogram(va.cl, vm.cl)
173 vm.cl <- vgm(40, "Sph", 2250, add.to = vmf.cl)
174 vmf2.cl <- fit.variogram(va.cl, vm.cl)
175
176 va.si <- variogram(SILT ~ 1, loc = d, cutoff = va.cutoff, width = va.width)
177 vm.si <- vgm(35, "Sph", 2500, 15); vmf.si <- fit.variogram(va.si, vm.si)
178
179 va.sa <- variogram(SAND ~ 1, loc = d, cutoff = va.cutoff, width = va.width)
180 vm.sa <- vgm(80, "Sph", 750, 15); vmf.sa <- fit.variogram(va.sa, vm.sa)
181 vm.sa <- vgm(40, "Lin", add.to = vmf.sa)
182 vmf2.sa <- fit.variogram(va.sa, vm.sa)
183
184 pdf(paste(path.fig, "costara_vario_soilsep_v2.pdf", sep = "/"),
185   width = 15, height = 14, pointsize = 31)
186 layout(matrix(c(1,1,2,2,4,3,3,5), 2, 4, byrow = TRUE))
187 par(mar = c(3,3,2,0)+.1, oma = c(.5,.5,.5,.5), mgp = c(2,1,0),
188   cex.main = 1, font.main = 2, las = 1)
189
190 plot(va.cl$gamma ~ va.cl$dist, pch = 20, cex = 1.25, col = "black",
191   xlim = c(0,max(va.cl$dist)*1.05), ylim = c(0,max(va.sa$gamma)*1.1),
192   axes = FALSE, ann = FALSE, xaxs = "i", yaxs = "i")
193 axis(side = 1, tck = -.02, labels = NA, at = seq(0,4000,1000))
194 axis(side = 2, tck = -.02, labels = NA)
195 axis(side = 1, lwd = 0, line = -.35, at = seq(0,4000,1000),
196   labels = c("0.1", seq(1000,4000,1000)))
197 axis(side = 2, lwd = 0, line = -.35, at = seq(0,200,50))
198 title(xlab = "Lag distance in m", ylab = "Variance", main = "a) Clay", sub = "")
```

```
lines(variogramLine(vmf2.cl, maxdist = max(va.cl$dist)),
  col = "black", lwd = 1.25)
legend("bottomright", "Nug + Sph + Sph", bty = "n")
box(which = "plot", lty = "solid", col = "black")

plot(va.sa$gamma ~ va.sa$dist, pch = 20, cex = 1.25, col = "black",
  xlim = c(0,max(va.sa$dist)*1.05), ylim = c(0,max(va.sa$gamma)*1.1),
  axes = FALSE, ann = FALSE, xaxs = "i", yaxs = "i")
axis(side = 1, tck = -.02, labels = NA, at = seq(0,4000,1000))
axis(side = 2, tck = -.02, labels = NA)
axis(side = 1, lwd = 0, line = -.35, at = seq(0,4000,1000),
  labels = c("0.1", seq(1000,4000,1000)))
axis(side = 2, lwd = 0, line = -.35, at = seq(0,200,50))
title(xlab = "Lag distance in m", ylab = "",  main = "b) Sand", sub = "")
lines(variogramLine(vmf2.sa, maxdist = max(va.sa$dist)),
  col = "black", lwd = 1.25)
legend("bottomright", "Nug + Sph + Lin", bty = "n")
box(which = "plot", lty = "solid", col = "black")

plot(va.si$gamma ~ va.si$dist, pch = 20, cex = 1.25, col = "black",
  xlim = c(0,max(va.si$dist)*1.05), ylim = c(0,max(va.si$gamma)*1.2),
  axes = FALSE, ann = FALSE, xaxs = "i", yaxs = "i")
axis(side = 1, tck = -.02, labels = NA, at = seq(0,4000,1000))
axis(side = 2, tck = -.02, labels = NA, at = seq(0,60,20))
axis(side = 1, lwd = 0, line = -.35, at = seq(0,3000,1000),
  labels = c("0.1", seq(1000,3000,1000)))
axis(side = 2, lwd = 0, line = -.35, at = seq(0,60,20))
title(xlab = "Lag distance in m", ylab = "Variance", main = "c) Silt", sub = "")
lines(variogramLine(vmf.si, maxdist = max(va.si$dist)),
  col = "black", lwd = 1.25)
legend("bottomright", "Nug + Sph", bty = "n")
box(which = "plot", lty = "solid", col = "black")
dev.off()

# variogram of alr-transformed target variables at calibration sites:
d.c <- d[-c(122:145,181:197),] # n = 156
d.v <- d[c(122:145,181:197),] # n = 41

va.cutoff <- 3250
# classical "methods of moments" estimation using gstat-package:
va.cl.alr <-
  variogram(CLAYalr ~ 1, loc = d.c, cutoff = va.cutoff, width = va.width)
vm.cl.alr <- vgm(.2, "Sph", 750, .05)
vmf.cl.alr <- fit.variogram(va.cl.alr, vm.cl.alr)
vm.cl.alr <- vgm(.2, "Lin", add.to = vmf.cl.alr)
vmf2.cl.alr <- fit.variogram(va.cl.alr, vm.cl.alr)

va.si.alr <-
  variogram(SILTalr ~ 1, loc = d.c, cutoff = va.cutoff, width = va.width)
vm.si.alr <- vgm(.075, "Sph", 250, .045)
```

```
vmf.si.alr <- fit.variogram(va.si.alr, vm.si.alr, fit.method = 0)
vm.si.alr <- vgm(.1, "Lin", add.to = vmf.si.alr)
vmf2.si.alr <- fit.variogram(va.si.alr, vm.si.alr)

pdf(paste(path.fig, "costara_vario_alr156_v2.pdf", sep = "/"),
  width = 15, height = 7, pointsize = 21)
par(mfrow = c(1,2), mar = c(3,3,2,0)+.1, oma = c(.5,.5,.5,.5), mgp = c(2,1,0),
  cex.main = 1, font.main = 2, las = 1)

plot(va.cl.alr$gamma ~ va.cl.alr$dist, pch = 20, cex = 1.25, col = "black",
  xlim = c(0,max(va.cl.alr$dist)*1.05), ylim = c(0,max(va.cl.alr$gamma)*1.2),
  axes = FALSE, ann = FALSE, xaxs = "i", yaxs = "i")
axis(side = 1, tck = -.02, labels = NA, at = seq(0,4000,1000))
axis(side = 2, tck = -.02, labels = NA)
axis(side = 1, lwd = 0, line = -.35, at = seq(0,4000,1000),
  labels = c("0.1", seq(1000,4000,1000)))
axis(side = 2, lwd = 0, line = -.35, at = seq(0,.5,.1))
title(xlab = "Lag distance in m",
  ylab = "Variance", main = "a) Clay, alr-transformed", sub = "")
lines(variogramLine(vmf2.cl.alr, maxdist = max(va.cl.alr$dist)),
  col = "black", lwd = 1.25)
legend("topleft", paste0("N = ", dim(d.c)[1]), bty = "n", inset = c(-.05,0))
legend("bottomright", "Nug + Sph + Lin", bty = "n")
box(which = "plot", lty = "solid", col = "black")

plot(va.si.alr$gamma ~ va.si.alr$dist, pch = 20, cex = 1.25, col = "black",
  xlim = c(0,max(va.si.alr$dist)*1.05), ylim = c(0,max(va.cl.alr$gamma)*1.2),
  axes = FALSE, ann = FALSE, xaxs = "i", yaxs = "i")
axis(side = 1, tck = -.02, labels = NA, at = seq(0,4000,1000))
axis(side = 2, tck = -.02, labels = NA, at = seq(0,.5,.1))
axis(side = 1, lwd = 0, line = -.35, at = seq(0,4000,1000),
  labels = c("0.1", seq(1000,4000,1000)))
axis(side = 2, lwd = 0, line = -.35, at = seq(0,.5,.1))
title(xlab = "Lag distance in m",
  ylab = "", main = "b) Silt, alr-transformed", sub = "")
lines(variogramLine(vmf2.si.alr, maxdist = max(va.si.alr$dist)),
  col = "black", lwd = 1.25)
legend("topleft", paste0("N = ", dim(d.c)[1]), bty = "n", inset = c(-.05,0))
legend("bottomright", "Nug + Sph + Lin", bty = "n")
box(which = "plot", lty = "solid", col = "black")
dev.off()

rm(h.max,va.cl,va.si,va.sa,vm.cl,vm.si,vm.sa,vmf.cl,vmf.si,vmf.sa,geol.units,
  vmf2.cl,vmf2.sa,va.cl.alr,va.si.alr,vm.cl.alr,vm.si.alr,vmf.cl.alr,vmf.si.alr,
  vmf2.cl.alr,vmf2.si.alr,diag.bbox,va.cutoff,va.width)

# end of script, costara_explor_sp_vDiss.R, 11.06.2014
```

.4.15 costara_nnrck20_vDiss.R

```
#
################################################################################
####  Neural Network Residual Cokriging of soil separates, Rio di Costara   ####
################################################################################
#
## last update: 27.03.2015

# neural network residual cokriging of alr-transformed soil separates

# author: Michael Blaschek (blaschek@geographie.uni-kiel.de)

rm(list = ls())

library(rgdal); library(gstat); library(RColorBrewer); library(compositions)
library(caret); library(RSNNS); library(RANN); library(spdep)
library(ggplot2); library(plyr)
# R v3.0.2, gstat_1.0-16, rgdal_0.8-15, RColorBrewer_1.0-5, compositions_1.40-1
# caret_6.0-22, RSNNS_0.4-6, RANN_2.3.0, spdep_0.5-74, ggplot2_0.9.3.1, plyr_1.8

load(paste(getwd(), "phd_calc_input/cos_nnBasis20a.RData", sep = "/"))
path.fig <- paste(getwd(), "phd_calc_figures", sep = "/")

n <- dim(d)[1]
# from costara_cor_vDiss.R:
nm.signif.uncorr.explan <- c("ELEV", "SAGAWI", "INSOLAT")
prd.in <- c(nm.signif.uncorr.explan,
  "GEOPPR12", "GEOPPR22", "GEOPPR470", "GEOPPR1465") # defined predictors


#-------------------------------------------------------------------------------
## Split data into training/validation/test sets:
#-------------------------------------------------------------------------------

nzv <- nearZeroVar(d@data, freqCut = 95/5, saveMetrics = TRUE) # requires caret
# --> GEOPPR18 + 1233 are near-zero-variance predictors

# logratio transform to account for the compositional character of the targets:
d.comp <- acomp(d@data, parts = c("CLAY", "SILT", "SAND"), total = 100)
d.alr <- alr(d.comp) # additive logratio transform!
d$CLAYalr <- d.alr[,1]; d$SILTalr <- d.alr[,2]

cor(d$CLAYalr, d$SILTalr) # 0.742

dv <- d[c(122:145,181:197),] # N = 41 = 500er und 800er
dc <- d[-c(122:145,181:197),] # N = 156

source("kennard_stone_func.R")

```

```
dcvi <- ks.galv(X = cbind(dc@data[,nm.signif.uncorr.explan[1]],
  dc@data[,nm.signif.uncorr.explan[2]], dc@data[,nm.signif.uncorr.explan[3]]),
  y = cbind(dc@data[,"CLAYalr"], dc@data["SILTalr"]),
  nc = length(dc)/5, dist.calc = "mahal")

dcc <- dc[-dcvi,]; dcv <- dc[dcvi,] # N = 125/31
# dcv = 1/5 of the dc-subset, as proposed in Saptoro12, p.9: 80% vs 20%)

rm(ks.galv,nzv,d.comp,d.alr)


#--------------------------------------------------------------------------
## Export training/validation/test sets and relevant predictors:
#--------------------------------------------------------------------------

dem1$GEOPPRe <- as.character(dem1$GEOPPRz)
dem1$GEOPPRe[dem1$GEOPPRe %in% c("18","1233")] <- -99
dem1$GEOPPRe[which(dem1$GEOPPRe == "-99")] <- NA
dem1$GEOPPRe <- as.integer(dem1$GEOPPRe)

proj4string(dem1) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")

exp.covar <- c(prd.in[1:3], "GEOPPRe")
for (i in exp.covar) {
  writeGDAL(dem1[i], drivername = "GTiff", type = "Float32",
  paste(getwd(), paste0("phd_calc_out/cos_", tolower(i), ".tif"), sep = "/"),
  mvFlag = -9999, options = c("TFW = YES", "DECIMAL_PRECISION = 3"))
}

proj4string(dcc) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
proj4string(dcv) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
proj4string(dv) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")

exp.d <- c("dv", "dcc", "dcv"); exp.dlist <- list(dv,dcc,dcv)
for (i in 1:length(exp.d)) {
  writeOGR(exp.dlist[[i]]["ID"], driver = "ESRI Shapefile",
    dsn = paste(getwd(), "phd_calc_out", sep = "/"),
    layer = paste0("cos_", exp.d[i]))
}
rm(exp.d, exp.dlist, i, exp.covar)


#--------------------------------------------------------------------------
## Define target grid size:
#--------------------------------------------------------------------------

```

```
# as proposed by T. Hengl 2006 - Finding the right pixel size:
# based on inspection density/working scale:
ezg.area <- 16.44 # km^2
obs <- 2.5 # observations per 1cm^2 of the map = recommended compromise

sn <- sqrt(obs * ezg.area * 1e6/n) * 100 # working scale = appr. 1:50.000

# the scale number can be used to estimate the grid resolution:
p <- sqrt(obs * ezg.area * 1e6/n) * 100 * .0005; p # grid resolution = appr. 20m

rm(ezg.area,obs,sn,p)


#-------------------------------------------------------------------------------
## Multi-layer perceptron:
#-------------------------------------------------------------------------------

# prepare data for mlp in RSNNS:
dcc.nn <- as.data.frame(dcc); dcv.nn <- as.data.frame(dcv)
dv.nn <- as.data.frame(dv)
# training datasets:
dcc.nn.in <- dcc.nn[,prd.in]; dcc.nn.out <- dcc.nn[,c("CLAYalr", "SILTalr")]
# test datasets:
dcv.nn.in <- dcv.nn[,prd.in]; dcv.nn.out <- dcv.nn[,c("CLAYalr", "SILTalr")]
# validation dataset:
dv.nn.in <- dv.nn[,prd.in]

# standardize input variables (zero mean, unit variance (1)):
dcc.nn.in.nrm <- normalizeData(dcc.nn.in[1:3], type = "norm")
dcv.nn.in.nrm <- normalizeData(dcv.nn.in[1:3], type = attr(dcc.nn.in.nrm,
  "normParams")) # the same normalization parameters as training set required
dcc.nn.in[,1:3] <- dcc.nn.in.nrm[,1:3]; dcv.nn.in[,1:3] <- dcv.nn.in.nrm[,1:3]

dv.nn.in.nrm <-
  normalizeData(dv.nn.in[1:3], type = attr(dcc.nn.in.nrm, "normParams"))
dv.nn.in[,1:3] <- dv.nn.in.nrm[,1:3]

# mlp training:
# test for different size (number of hidden neurons) and maxit (number of runs):
n.size <- c(5, 7, 9, 11, 13, 15); n.maxit <- c(20, 40, 60); n.rep <- 10

te <- data.frame(iterr = rep(1,length(n.size) * length(n.maxit) * n.rep),
  group = rep("A",length(n.size) * length(n.maxit) * n.rep),
    mlp = rep("A",length(n.size) * length(n.maxit) * n.rep)); j <- 1
te$group <- as.character(te$group); te$mlp <- as.character(te$mlp)
for (i in 1:length(n.size)) {
  n.size.name <- as.character(n.size[i])
  for (q in n.maxit) {
    n.maxit.name <- q
    for (z in 1:n.rep) {
```

```
      mlp <- mlp(dcc.nn.in, dcc.nn.out, size = c(n.size[i]),
        initFunc = "Randomize_Weights", initFuncParams = c(-2,2),
        maxit = q, learnFunc = "Rprop", learnFuncParams = c(.1,30,5),
        updateFunc = "Topological_Order", updateFuncParams = c(0),
        hiddenActFunc = "Act_TanH", shufflePatterns = FALSE,
        linOut = TRUE, inputsTest = dcv.nn.in, targetsTest = dcv.nn.out)
      te$iterr[j] <- mlp$IterativeTestError[q]
      te$group[j] <- paste("mlp", n.size.name, n.maxit.name, sep = ".")
      te$mlp[j] <- paste("mlp", n.size.name, n.maxit.name, z, sep = ".")
      j <- j + 1
      assign(paste("mlp", n.size.name, n.maxit.name, z, sep = "."), mlp)
    }
  }
}

te$group <- as.factor(te$group)
ddply(te, .(group), summarize, mean = round(mean(iterr), 2))
# mlp.9.40 --> group with lowest average iterative test error
# mlp.9.40.4 --> best run in mlp.9.40 --> used for prediction!
te[which(te$iterr == min(te$iterr[which(te$group == "mlp.9.40")])),3]
mlp.9.40.4$IterativeTestError[40] # 26.0893

save(list = ls(pattern = "mlp."),
  file = paste(getwd(), "phd_calc_out/cos_nnModels.RData", sep = "/"))

mlp <- mlp.9.40.4

rm(famd,n.size,n.maxit,n.rep,i,q,z,j)
rm(list = ls(pattern = "mlp."))

sum((mlp$fittedTestValues - dcv.nn.out)^2); mlp$IterativeTestError[40] # check!!

#--- check how mlp propagates information:
# activation of single particular hidden unit:
# note how bias is included..
extractNetInfo(mlp) # --> weights and biases..
extr.m <- extractNetInfo(mlp)$fullWeightMatrix
tanh(c(extractNetInfo(mlp)$unitDefinitions[8,4],as.vector(extr.m[1:7,8])) %*%
  rbind(1,t(as.vector(dcv.nn.in[31,])))) # unit h1 = 0.9998203
extractNetInfo(mlp)$unitDefinitions[8,3] # 0.99982

tanh(c(extractNetInfo(mlp)$unitDefinitions[9,4],as.vector(extr.m[1:7,9])) %*%
  rbind(1,t(as.vector(dcv.nn.in[31,])))) # unit h1 = -0.9689528
extractNetInfo(mlp)$unitDefinitions[9,3] # -0.96895
#---

dcc$CLAY_NNFIT <- mlp$fitted.values[,1]; dcc$SILT_NNFIT <- mlp$fitted.values[,2]
dcv$CLAY_NNFIT <- mlp$fittedTestValues[,1]
dcv$SILT_NNFIT <- mlp$fittedTestValues[,2]
dcn <- rbind(dcc, dcv)
```

```
dcn$CLAY_NNRES <- dcn$CLAYalr - dcn$CLAY_NNFIT
dcn$SILT_NNRES <- dcn$SILTalr - dcn$SILT_NNFIT


#-------------------------------------------------------------------------
## Ordinary Cokriging of MLP-residuals:
#-------------------------------------------------------------------------

cor(dcn$CLAY_NNRES, dcn$SILT_NNRES) # 0.62 --> cokriging strongly justified

# Moran's I to check remaining autocorrelation in MLP-residuals:
nlistk4 <- knn2nb(knearneigh(dcn, k = 8)); w <- nb2listw(nlistk4, style = "B")
clay.res.mi <-
  moran.test(dcn$CLAY_NNRES, w, randomisation = F, alternative = "two.sided")
silt.res.mi <-
  moran.test(dcn$SILT_NNRES, w, randomisation = F, alternative = "two.sided")
clay.res.mi # 0.067/1.966/0.049
silt.res.mi # 0.073/2.138/0.032
rm(nlistk4,w) # remaining objects: clay.res.mi, silt.res.mi

# fitting linear model of coregionalization to (cross-)variograms:
g <- gstat(NULL, id = "CLAY_NNRES", form = CLAY_NNRES ~ 1, data = dcn)
g <- gstat(g, id = "SILT_NNRES", form = SILT_NNRES ~ 1, data = dcn)
va.cross <- variogram(g, cutoff = 1000, width = 125)
va <- variogram(CLAY_NNRES ~ 1, loc = dcn, cutoff = 1000, width = 125)
vm <- vgm(.15, "Sph", 500, .05); vmf <- fit.variogram(va, vm, fit.method = 7)
g <- gstat(g, id = "CLAY_NNRES", model = vmf, fill.all = T)
g <- fit.lmc(va.cross, g)
# nugget, partial sill and (constant) range parameter:
g$model$CLAY_NNRES[1,2]; g$model$SILT_NNRES[1,2]
g$model$CLAY_NNRES.SILT_NNRES[1,2]
g$model$CLAY_NNRES[2,2]; g$model$SILT_NNRES[2,2]
g$model$CLAY_NNRES.SILT_NNRES[2,2]
g$model$CLAY_NNRES.SILT_NNRES[2,3]
# Nugget-to-sill-ratios (NSR) after Cambardella94
g$model$CLAY_NNRES[1,2] /
  (g$model$CLAY_NNRES[1,2] + g$model$CLAY_NNRES[2,2]) * 100
g$model$SILT_NNRES[1,2] /
  (g$model$SILT_NNRES[1,2] + g$model$SILT_NNRES[2,2]) * 100

pdf(paste(path.fig, "costara_crossvario_MLPres_v2.pdf", sep = "/"),
  width = 15, height = 14, pointsize = 31)
par(mar = c(3,3,2,0)+.1, oma = c(.5,.5,.5,.5), mgp = c(2,1,0),
  cex.main = 1, font.main = 2, las = 1)
plot(va.cross$gamma[17:24] ~ va.cross$dist[17:24], pch = 16, cex = .75,
  xlim = c(0, max(va.cross$dist) * 1.05), col = "black", xaxs = "i", yaxs = "i",
  ylim = c(0, max(va.cross$gamma) * 1.2), axes = FALSE,
  xlab = "lag distance in m", ylab = "(cross-)variance", main = "")
axis(side = 1, tck = -.02, cex.axis = 1, labels = NA)
axis(side = 2, tck = -.02, cex.axis = 1, labels = NA)
```

```
axis(side = 1, lwd = 0, line = -.35, at = seq(0,800,200),
  labels = c("0.1", seq(200,800,200)))
axis(side = 2, lwd = 0, line = -.35, at = c("0.00", "0.05", ".10", ".15",
  "0.20", "0.25"), labels = c("0.00", "0.05", ".10", ".15", "0.20", "0.25"))
box(which = "plot", lty = "solid", col = "black")
lines(variogramLine(g$model$CLAY_NNRES, maxdist = max(va.cross$dist[17:24])),
  col = "black", lwd = 1.35)
lines(variogramLine(g$model$SILT_NNRES, maxdist = max(va.cross$dist[9:16])),
  col = "black", lwd = 1.35, lty = 2)
lines(variogramLine(g$model$CLAY_NNRES.SILT_NNRES,
  maxdist = max(va.cross$dist[1:8])), col = "black", lwd = 1.35, lty = 4)
points(va.cross$dist[9:16], va.cross$gamma[9:16],
  col = "black", pch = 22, cex = .75)
points(va.cross$dist[1:8], va.cross$gamma[1:8],
  col = "black", pch = 17, cex = .75)
legend("bottomright", c("alr CLAY","alr SILT","Cross"),
  col = c("black"), pch = c(16,22,17), cex = .75)
legend("topleft", c("alr CLAY","alr SILT","Cross"),
  col = c("black"), lty = c(1,2,4), cex = .75)
dev.off()
rm(va,vm,vmf) # remaining objects: g, va.cross


#-------------------------------------------------------------------------------
## NNRCK-predictions at unknown locations:
#-------------------------------------------------------------------------------

dem1.nn <- as.data.frame(dem1); dem1.nn.in <- dem1.nn[,prd.in]

dem1.nn.in.nrm <-
  normalizeData(dem1.nn.in[1:3], type = attr(dcc.nn.in.nrm, "normParams"))
dem1.nn.in[,1:3] <- dem1.nn.in.nrm[,1:3]

# ensure unique CRS:
proj4string(dcn) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
proj4string(dem1) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
proj4string(g$data$CLAY_NNRES$data) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
proj4string(g$data$SILT_NNRES$data) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")

prd.dem1 <- predict(mlp, dem1.nn.in); nnrck.dem1 <- predict.gstat(g, dem1)
dem1$CLAY_NNRCOKALR <- prd.dem1[,1] + nnrck.dem1$CLAY_NNRES.pred
dem1$SILT_NNRCOKALR <- prd.dem1[,2] + nnrck.dem1$SILT_NNRES.pred

dem1.alr <- matrix(c(dem1$CLAY_NNRCOKALR, dem1$SILT_NNRCOKALR),
  nrow = length(dem1$ELEV), ncol = 2, dimnames = list(NULL, c("CLAY", "SILT")))
dem1.comp.back <- alrInv(dem1.alr)
```

```
dem1$CLAY_NNRCOKALRBACK <- dem1.comp.back[,1] * 100
dem1$SILT_NNRCOKALRBACK <- dem1.comp.back[,2] * 100
dem1$SAND_NNRCOKALRBACK <- dem1.comp.back[,3] * 100

# set all pixels outside the catchment boundaries to NA:
catchm <- readOGR(dsn = "phd_calc_input", layer = "costara_catchm_v1")
proj4string(catchm) <-
  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
dem2 <- dem1; asdf <- which(!is.na(over(dem2, catchm))); dem3 <- dem2[asdf,]

# calculate max/min-values for comparable prediction maps:
nnrck.min <- min(apply(dem3@data[,c("CLAY_NNRCOKALRBACK", "SILT_NNRCOKALRBACK",
  "SAND_NNRCOKALRBACK")], 2, min, na.rm = TRUE)); nnrck.min # 8.4
nnrck.max <- max(apply(dem3@data[,c("CLAY_NNRCOKALRBACK", "SILT_NNRCOKALRBACK",
  "SAND_NNRCOKALRBACK")], 2, max, na.rm = TRUE)); nnrck.max # 79.4

pts.c <- list("sp.points", dc, pch = 19, col = "black", cex = .6, alpha = 1)
pts.v <- list("sp.points", dv, pch = 21, col = "black", alpha = 1, cex = .6)

cat <- fortify(catchm, region = "NAME")
ptc <- as.data.frame(pts.c); ptv <- as.data.frame(pts.v)
dem <- as.data.frame(dem3);
v.interest <-
  c("CLAY_NNRCOKALRBACK","SILT_NNRCOKALRBACK", "SAND_NNRCOKALRBACK", "x", "y")
dem.clay <- dem[,v.interest]
dem.clay$VARS <- "Clay"; dem.clay$VALUE <- dem.clay$CLAY_NNRCOKALRBACK
dem.silt <- dem[,v.interest]
dem.silt$VARS <- "Silt"; dem.silt$VALUE <- dem.silt$SILT_NNRCOKALRBACK
dem.sand <- dem[,v.interest]
dem.sand$VARS <- "Sand"; dem.sand$VALUE <- dem.sand $SAND_NNRCOKALRBACK
dem.clay <- dem.clay[,-c(1:3)]; dem.silt <- dem.silt[,-c(1:3)]
dem.sand <- dem.sand[,-c(1:3)]
demm <- rbind(dem.clay, dem.silt, dem.sand)

demm$CUTV <- cut(demm$VALUE, breaks = seq(0,81,9)) # 81 from nnrck.max
levels(demm$CUTV) <- c("0 - 9", "9 - 18", "18 - 27", "27 - 36", "36 - 45",
  "45 - 54", "54 - 63", "63 - 72", "72 - 81")

# change plot order of facet grid by changing the order of levels with factor():
demm$VARS <- factor(demm$VARS, levels = c("Clay", "Silt", "Sand"))

map <- ggplot(cat, aes(long, lat)) +
  geom_raster(aes(x, y, fill = CUTV), data = demm) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1,
    data = ptc, shape = 19) +
  geom_point(mapping = aes(x = EAST, y = NORTH), size = 1,
    data = ptv, shape = 21) +
  geom_polygon(size = .5, linetype = "dashed", color = "black",
    fill = "grey40", alpha = 0) +
  coord_equal() +
```

```
  theme_bw(base_size = 9, base_family = "Helvetica") +
  scale_fill_brewer(name = "Content in %", palette = "YlOrBr") +
  labs(x = "\nUTM-E/m", y = "UTM-N/m\n") +
  scale_y_continuous(breaks = c(4361000,4363000)) +
  scale_x_continuous(breaks = c(508000,510000,512000,514000)) +
  facet_wrap(~ VARS, ncol = 2) +
  theme(axis.text.y = element_text(angle = 90, hjust = .5),
    legend.position = c(.625,.25), legend.text = element_text(size = 10),
    legend.title = element_text(size = 10), legend.key.size = unit(14, "pt"),
    axis.title.x = element_text(hjust = .25),
    strip.text.x = element_text(size = 12),
    axis.text = element_text(size = 10), axis.title = element_text(size = 10))

ggsave(paste(path.fig, "costara_nnrck_predmaps_v1.pdf", sep = "/"), map,
  width = 8.27, height = 5.19)
print(map)
dev.off()

rm(dv.nn.in.nrm,dcc.nn.in.nrm,dcv.nn.in.nrm)
rm(pts.c,pts.v,nnrck.dem1,dem1.comp.back,dem1.alr,dem,dem1,dem2,asdf,prd.dem1,
  dem1.nn.in.nrm,dem.clay,dem.sand,dem.silt,v.interest,map,nnrck.max,nnrck.min)
# remaining objects: cat, demm, ptc, ptv


#-------------------------------------------------------------------------------
## Validation preparation (holdout method):
#-------------------------------------------------------------------------------

# final (NNRCK) predictions at validation points:
prd <- predict(mlp, dv.nn.in); cok.out <- predict.gstat(g, dv)
dv$CLAY_NNRCOKALR <- prd[,1] + cok.out$CLAY_NNRES.pred
dv$SILT_NNRCOKALR <- prd[,2] + cok.out$SILT_NNRES.pred

# backtranform (biased) = additive generalized logistic (agl) transform:
dv.alr <- matrix(c(dv$CLAY_NNRCOKALR, dv$SILT_NNRCOKALR),
  nrow = length(dv$CLAYalr), ncol = 2, dimnames = list(NULL, c("CLAY", "SILT")))
d.comp.back <- alrInv(dv.alr)
dv$CLAY_NNRCOKALRBACK <- d.comp.back[,1] * 100
dv$SILT_NNRCOKALRBACK <- d.comp.back[,2] * 100
dv$SAND_NNRCOKALRBACK <- d.comp.back[,3] * 100

rm(d.comp.back,dv.alr,cok.out,prd,
  dcc.nn,dcc.nn.in,dcc.nn.out,dcv.nn,dcv.nn.in,dcv.nn.out,dv.nn,dv.nn.in)


#-------------------------------------------------------------------------------
## Export:
#-------------------------------------------------------------------------------

# remove interim results from final objects:
```

```
drop.c <- c("CLAY_NNRCOKALR","SILT_NNRCOKALR")
dv@data <- dv@data[,!(names(dv) %in% drop.c)]
dem3@data <- dem3@data[,!(names(dem3) %in% drop.c)]
rm(drop.c)

rm(silt.res.mi, clay.res.mi,dem1.nn,dem1.nn.in)
save.image(paste(getwd(), "phd_calc_out/cos_nnrckResults.RData", sep = "/"))

writeGDAL(dem3["CLAY_NNRCOKALRBACK"], drivername = "GTiff", type = "Float32",
  paste(getwd(), "phd_calc_out/nnrckclay.tif", sep = "/"),
  mvFlag = -9999, options = c("TFW = YES", "DECIMAL_PRECISION = 3"))

writeGDAL(dem3["SILT_NNRCOKALRBACK"], drivername = "GTiff", type = "Float32",
  paste(getwd(), "phd_calc_out/nnrcksilt.tif", sep = "/"),
  mvFlag = -9999, options = c("TFW = YES", "DECIMAL_PRECISION = 3"))

writeGDAL(dem3["SAND_NNRCOKALRBACK"], drivername = "GTiff", type = "Float32",
  paste(getwd(), "phd_calc_out/nnrcksand.tif", sep = "/"),
  mvFlag = -9999, options = c("TFW = YES", "DECIMAL_PRECISION = 3"))

# end of script, costara_nnrck20_vDiss.R, 04.07.2014
```

.4.16 costara_nnEval_vDiss.R

```
#
##############################################################################
#### Validation of various digital soil mapping techniques, Rio di Costara  ####
##############################################################################
#
## last update: 22.03.2015

# evaluate neural network modelling of alr-transformed soil separates
#   neural interpretation diagram, relative importance of inputs

# author: Michael Blaschek (blaschek@geographie.uni-kiel.de)

rm(list = ls())

library(NeuralNetTools); library(ggplot2)
# R v3.1.3, NeuralNetTools_1.0.1, ggplot2_1.0.0

load(paste(getwd(), "phd_calc_out/cos_nnResults.RData", sep = "/"))

# neural interpretation diagram:
# https://
#   beckmw.wordpress.com/2013/11/14/visualizing-neural-networks-in-r-update/

z <- c("CLAYalr", "SILTalr")
zn <- c("alr Clay", "alr Silt")
```

```
26
27 pdf("costara_nid_v5.pdf", width = 8.27, height = 5.69)
28 #par(mar = numeric(4), family = "serif")
29 par(mar = numeric(4))
30 plotnet(mlp, rel_rsc = 4, cex_val = 1, alpha_val = .65, bord_col = "black",
31   x_lab = prd.in, y_lab = zn, circle_col = as.list(c("black", "gray60")))
32 dev.off()
33
34
35 # relative importance of input variables using Garson's algorithm:
36 gs <- NULL
37 gs[[1]] <- garson(mlp, "Output_1", bar_plot = FALSE)
38 gs[[2]] <- garson(mlp, "Output_2", bar_plot = FALSE)
39 # output-var must be named as in mlp$snnsObject$getUnitDefinitions()
40
41 n.garson <- length(prd.in) * length(z); n.garson
42 j <- 1
43
44 # prepare data.frame for ggplot2:
45 garson <- data.frame(
46   RI = rep(1,n.garson), METHOD = rep(1,n.garson), VARS = rep(1,n.garson))
47 for (t in 1:length(z)) {
48   for (i in 1:length(prd.in)) {
49     garson[j,"RI"] <- gs[[t]][i,"rel_imp"]
50     garson[j,"METHOD"] <- prd.in[i]
51     garson[j,"VARS"] <- z[t]
52     j <- j + 1
53   }
54 }
55
56 garson[which(garson$VARS == "CLAYalr"),"VARS"] <- "alr Clay"
57 garson[which(garson$VARS == "SILTalr"),"VARS"] <- "alr Silt"
58
59 # change plot order of facet grid by changing the order of levels with factor():
60 garson$METHOD <- factor(garson$METHOD, levels = prd.in)
61 garson$VARS <- factor(garson$VARS, levels = zn)
62
63 garson[,1] <- garson[,1] * 100 # --> in percent
64
65 garson
66
67 p.garson <- ggplot(data = garson,
68     aes(x = METHOD, y = RI, width = .75)) +
69   geom_bar(stat = "identity", fill = "black") +
70   facet_wrap(~ VARS, ncol = 2) +
71   labs(x = "\nNN/MLP input variables", y = "Relative importance in %\n") +
72   theme_bw(base_size = 12) +
73   theme(strip.text.x = element_text(size = 12),
74     axis.text.y = element_text(size = 12),
75     axis.title = element_text(size = 12),
```

```
76     axis.text.x = element_text(angle = 45, size = 12, hjust = 1),
77     panel.grid.major.x = element_blank())
78 print(p.garson)
79
80 ggsave("costara_garson_v1.pdf", p.garson, width = 8.27, height = 6.19)
81 print(p.garson)
82 dev.off()
83
84 # end of script, costara_nnEval_vDiss.R, 21.03.2015
```

.4.17 costara_modelcomp20_vDiss.R

```
1 #
2 ##############################################################################
3 #### Comparison of various digital soil mapping techniques, Rio di Costara  ####
4 ##############################################################################
5 #
6 ## last update: 09.04.2015
7
8 # compare neural network residual cokriging of alr-transformed soil separates
9 #   with five (common) DSM techniques (IDW, OK, RK, COK, UCOK)
10
11 # author: Michael Blaschek (blaschek@geographie.uni-kiel.de)
12
13 rm(list = ls())
14
15 library(rgdal); library(gstat); library(RColorBrewer); library(compositions)
16 library(caret); library(ggplot2); library(spdep); library(soiltexture)
17 library(irr); library(directlabels) # irr = inter-rater reliability
18 # R v3.0.2, gstat_1.0-16, rgdal_0.8-15, RColorBrewer_1.0-5, compositions_1.40-1
19 # caret_6.0-22, ggplot2_0.9.3.1, spdep_0.5-74, soiltexture_1.2.13, irr_0.84
20 # directlabels_2013.6.15
21
22 load(paste(getwd(), "phd_calc_out/cos_nnrckResults.RData", sep = "/"))
23 dem1 <- dem3
24 proj4string(dc) <-
25   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
26 proj4string(dv) <-
27   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
28 proj4string(dem1) <-
29   CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")
30
31
32 #------------------------------------------------------------------------------
33 ## Inverse Distance Weighting (IDW):
34 #------------------------------------------------------------------------------
35
36 # with fixed idp-value = 2.0:
37 idp.opt <- 2.0; idp.opt2 <- 2.0
```

```
38
39 idw.out <- idw(CLAY ~ 1, dc, dem1, idp = idp.opt)
40 idw.out2 <- idw(SILT ~ 1, dc, dem1, idp = idp.opt2)
41 idw.dv <- idw(CLAY ~ 1, dc, dv, idp = idp.opt)
42 idw.dv2 <- idw(SILT ~ 1, dc, dv, idp = idp.opt2)
43
44 dem3$CLAY_IDW <- idw.out$var1.pred; dem3$SILT_IDW <- idw.out2$var1.pred
45 dv$CLAY_IDW <- idw.dv$var1.pred; dv$SILT_IDW <- idw.dv2$var1.pred
46
47 # sand, calculated as difference from predicted clay + silt:
48 dem3$SAND_IDWd <- 100 - dem3$SILT_IDW - dem3$CLAY_IDW
49 dv$SAND_IDWd <- 100 - dv$SILT_IDW - dv$CLAY_IDW
50
51 dem3$CLAY_IDW[which(is.na(dem3$ELEV))] <- NA
52 dem3$SILT_IDW[which(is.na(dem3$ELEV))] <- NA
53 dem3$SAND_IDWd[which(is.na(dem3$ELEV))] <- NA
54
55 rm(idp.opt,idp.opt2,idw.out,idw.out2,idw.dv,idw.dv2)
56
57
58 #------------------------------------------------------------------------------
59 ## Ordinary Kriging (OK):
60 #------------------------------------------------------------------------------
61
62 va.cl <- variogram(CLAY ~ 1, loc = dc, cutoff = 4250, width = 125)
63 vm.cl <- vgm(80, "Sph", 750, 15); vmf.cl <- fit.variogram(va.cl, vm.cl)
64 vm.cl <- vgm(40, "Sph", 2250, add.to = vmf.cl)
65 vmf2.cl <- fit.variogram(va.cl, vm.cl)
66
67 va.si <- variogram(SILT ~ 1, loc = dc, cutoff = 4250, width = 125)
68 vm.si <- vgm(35, "Sph", 2500, 15)
69 vmf.si <- fit.variogram(va.si, vm.si, fit.method = 7)
70
71 ok.out <- krige(CLAY ~ 1, dc, dem1, model = vmf2.cl)
72 ok.out2 <- krige(SILT ~ 1, dc, dem1, model = vmf.si)
73 ok.dv <- krige(CLAY ~ 1, dc, dv, model = vmf2.cl)
74 ok.dv2 <- krige(SILT ~ 1, dc, dv, model = vmf.si)
75
76 dem3$CLAY_OK <- ok.out$var1.pred; dem3$SILT_OK <- ok.out2$var1.pred
77 dv$CLAY_OK <- ok.dv$var1.pred; dv$SILT_OK <- ok.dv2$var1.pred
78
79 # sand, calculated as difference from predicted clay + silt:
80 dem3$SAND_OKd <- 100 - dem3$SILT_OK - dem3$CLAY_OK
81 dv$SAND_OKd <- 100 - dv$SILT_OK - dv$CLAY_OK
82
83 dem3$CLAY_OK[which(is.na(dem3$ELEV))] <- NA
84 dem3$SILT_OK[which(is.na(dem3$ELEV))] <- NA
85 dem3$SAND_OKd[which(is.na(dem3$ELEV))] <- NA
86
87 rm(ok.out,ok.out2,ok.dv,ok.dv2,va.cl,va.si,vm.cl,vmf.cl,vmf2.cl,vm.si,vmf.si)
```

```
88
89
90 #-----------------------------------------------------------------------------
91 ## Regression Kriging (RK) = Kriging with external drift (KED):
92 #-----------------------------------------------------------------------------
93
94 #prd.in <- c(nm.signif.uncorr.explan,
95 #  "GEOPPR12", "GEOPPR22", "GEOPPR470", "GEOPPR1465") # defined predictors
96 # stepwise regression:
97 regr.start <- lm(CLAY ~ ELEV + SAGAWI + INSOLAT + GEOPPR12 + GEOPPR22 +
98   GEOPPR470 + GEOPPR1465, data = dc) # adj-r.squared = 0.14
99 regr <- step(regr.start, direction = "backward") # adj-r.squared = 0.16
100 regr.fwd <- step(lm(CLAY ~ 1, dc), scope = list(lower = CLAY ~ 1,
101   upper = CLAY ~ ELEV + SAGAWI + INSOLAT + GEOPPR12 + GEOPPR22 +
102     GEOPPR470 + GEOPPR1465), direction = "forward")
103 # forward and backward lead to the same result: CLAY ~ ELEV + GEOPPR1465
104
105 regr.start <- lm(CLAY ~ GEOPPR22 + SAGAWI + GEOPPR1465 + GEOPPR12 + ELEV +
106   GEOPPR470 + INSOLAT, data = dc); summary(regr.start)
107 regr <- step(regr.start, direction = "backward"); summary(regr)
108 # another order yields the same result as above --> very good
109
110 regr.start2 <- lm(SILT ~ ELEV + SAGAWI + INSOLAT + GEOPPR12 + GEOPPR22 +
111   GEOPPR470 + GEOPPR1465, data = dc); summary(regr.start2) # adj-r.squared =
        0.22
112 regr2 <- step(regr.start2, direction = "backward"); summary(regr2) # 0.22
113 regr.fwd2 <- step(lm(SILT ~ 1, dc), scope = list(lower = SILT ~ 1,
114   upper = SILT ~ ELEV + SAGAWI + INSOLAT + GEOPPR12 + GEOPPR22 +
115     GEOPPR470 + GEOPPR1465), direction = "forward"); summary(regr.fwd2)
116 # final regression formula: SILT ~ ELEV + INSOLAT + GEOPPR470 + GEOPPR1465
117
118 regr.start2 <- lm(SILT ~ GEOPPR22 + SAGAWI + GEOPPR1465 + GEOPPR12 + ELEV +
119   GEOPPR470 + INSOLAT, data = dc); summary(regr.start2)
120 regr2 <- step(regr.start2, direction = "backward"); summary(regr2)
121 # another order yields the same result as above
122
123 regr <- lm(CLAY ~ ELEV + GEOPPR1465, data = dc) # r.squared = 0.17
124 regr2 <- lm(SILT ~ ELEV + INSOLAT + GEOPPR470 + GEOPPR1465, data = dc) # r2: .24
125
126 rva.cl <- variogram(formula(regr), loc = dc, cutoff = 4250, width = 125)
127 rvm.cl <- vgm(60,"Exp",750,15); rvmf.cl <- fit.variogram(rva.cl, rvm.cl)
128
129 rva.si<- variogram(formula(regr2), loc = dc, cutoff = 4250, width = 125)
130 rvm.si <- vgm(30, "Exp", 750, vmf.si[1,2])
131 rvmf.si <-
132   fit.variogram(rva.si, rvm.si, fit.sills = c(F,T), fit.method = 6) # 6 = OLS
133
134 ked.out <- krige(formula(regr), dc, dem1, model = rvmf.cl)
135 ked.out2 <- krige(formula(regr2), dc, dem1, model = rvmf.si)
136 ked.dv <- krige(formula(regr), dc, dv, model = rvmf.cl)
```

```
ked.dv2 <- krige(formula(regr2), dc, dv, model = rvmf.si)

dem3$CLAY_KED <- ked.out$var1.pred; dem3$SILT_KED <- ked.out2$var1.pred
dv$CLAY_KED <- ked.dv$var1.pred; dv$SILT_KED <- ked.dv2$var1.pred

# sand, calculated as difference from predicted clay + silt:
dem3$SAND_KEDd <- 100 - dem3$SILT_KED - dem3$CLAY_KED
dv$SAND_KEDd <- 100 - dv$SILT_KED - dv$CLAY_KED

dem3$CLAY_KED[which(is.na(dem3$ELEV))] <- NA
dem3$SILT_KED[which(is.na(dem3$ELEV))] <- NA
dem3$SAND_KEDd[which(is.na(dem3$ELEV))] <- NA

rm(regr,regr2,rva.cl,rvm.cl,rvmf.cl,rva.si,rvm.si,rvmf.si,
  ked.dv,ked.dv2,ked.out,ked.out2,regr.fwd,regr.fwd2,regr.start,regr.start2)


#-------------------------------------------------------------------------------
## Ordinary Cokriging (COK):
#-------------------------------------------------------------------------------

# fitting linear model of coregionalization to (cross-)variograms:
g <- gstat(NULL, id = "CLAYalr", form = CLAYalr ~ 1, data = dc)
g <- gstat(g, id = "SILTalr", form = SILTalr ~ 1, data = dc)

va.cross <- variogram(g, cutoff = 4250, width = 125)
va <- variogram(CLAY ~ 1, loc = dc, cutoff = 4250, width = 125)
vm <- vgm(80, "Sph", 750, 15); vmf <- fit.variogram(va, vm)
vm <- vgm(40, "Sph", 2250, add.to = vmf)
vmf2 <- fit.variogram(va, vm)

va.cross <- variogram(g, cutoff = 3000, width = 125)
va <- variogram(CLAY ~ 1, loc = dc, cutoff = 3000, width = 125)
vm <- vgm(80, "Sph", 750, 15); vmf <- fit.variogram(va, vm)
vm <- vgm(40, "Lin", 3000, add.to = vmf)
vmf2 <- fit.variogram(va, vm)

g <- gstat(g, id = "CLAYalr", model = vmf2, fill.all = T)
g <- fit.lmc(va.cross, g)

g.cok <- g; va.cross.cok <- va.cross

cok.out <- predict.gstat(g, dem1)
dem1$CLAY_COKALR <- cok.out$CLAYalr.pred
dem1$SILT_COKALR <- cok.out$SILTalr.pred

dem1.alr <- matrix(c(dem1$CLAY_COKALR, dem1$SILT_COKALR),
  nrow = length(dem1$ELEV), ncol = 2, dimnames = list(NULL, c("CLAY", "SILT")))
dem1.comp.back <- alrInv(dem1.alr)
dem3$CLAY_COKALRBACK <- dem1.comp.back[,1] * 100
```

```
dem3$SILT_COKALRBACK <- dem1.comp.back[,2] * 100
dem3$SAND_COKALRBACK <- dem1.comp.back[,3] * 100

dem3$CLAY_COKALRBACK[which(is.na(dem3$ELEV))] <- NA
dem3$SILT_COKALRBACK[which(is.na(dem3$ELEV))] <- NA
dem3$SAND_COKALRBACK[which(is.na(dem3$ELEV))] <- NA


cok.dv <- predict.gstat(g, dv)
dv$CLAY_COKALR <- cok.dv$CLAYalr.pred
dv$SILT_COKALR <- cok.dv$SILTalr.pred

# backtranform (biased) = additive generalized logistic (agl) transform:
dv.alr <- matrix(c(dv$CLAY_COKALR, dv$SILT_COKALR),
  nrow = length(dv$CLAYalr), ncol = 2, dimnames = list(NULL, c("CLAY", "SILT")))
d.comp.back <- alrInv(dv.alr)
dv$CLAY_COKALRBACK <- d.comp.back[,1] * 100
dv$SILT_COKALRBACK <- d.comp.back[,2] * 100
dv$SAND_COKALRBACK <- d.comp.back[,3] * 100

rm(va,vm,vmf,vmf2,va.cross,cok.dv,cok.out,d.comp.back,dem1.comp.back,g)


#-------------------------------------------------------------------------------
## Regression cokriging (RCOK):
#-------------------------------------------------------------------------------

#prd.in <- c(nm.signif.uncorr.explan,
#  "GEOPPR12", "GEOPPR22", "GEOPPR470", "GEOPPR1465") # defined predictors
# stepwise regression:
regr.start <- lm(CLAYalr ~ ELEV + SAGAWI + INSOLAT + GEOPPR12 + GEOPPR22 +
  GEOPPR470 + GEOPPR1465, data = dc)
regr <- step(regr.start, direction = "backward"); summary(regr) # 0.20
# CLAYalr ~ ELEV + GEOPPR12 + GEOPPR22 + GEOPPR470 + GEOPPR1465

regr.start2 <- lm(SILTalr ~ ELEV + SAGAWI + INSOLAT + GEOPPR12 + GEOPPR22 +
 GEOPPR470 + GEOPPR1465, data = dc)
regr2 <- step(regr.start2, direction = "backward"); summary(regr2) # 0.26
# SILTalr ~ ELEV + INSOLAT + GEOPPR12 + GEOPPR22 + GEOPPR470

# fitting linear model of coregionalization to (cross-)variograms:
g <- gstat(NULL, id = "CLAYalr", form = formula(regr), data = dc)
g <- gstat(g, id = "SILTalr", form = formula(regr2), data = dc)
va.cross <- variogram(g, cutoff = 1000, width = 125)
va <- variogram(CLAYalr ~ ELEV + SAGAWI + INSOLAT + GEOPPR12 + GEOPPR22 +
  GEOPPR470 + GEOPPR1465, loc = dc, cutoff = 1000, width = 125)
vm <- vgm(.15, "Sph", 500, .05); vmf <- fit.variogram(va, vm, fit.method = 7)
g <- gstat(g, id = "CLAYalr", model = vmf, fill.all = T)
g <- fit.lmc(va.cross, g)

```

```
ucok.out <- predict.gstat(g, dem1)
dem1$CLAY_UCOKALR <- ucok.out$CLAYalr.pred
dem1$SILT_UCOKALR <- ucok.out$SILTalr.pred

dem1.alr <- matrix(c(dem1$CLAY_UCOKALR, dem1$SILT_UCOKALR),
  nrow = length(dem1$ELEV), ncol = 2, dimnames = list(NULL, c("CLAY", "SILT")))
dem1.comp.back <- alrInv(dem1.alr)
dem3$CLAY_UCOKALRBACK <- dem1.comp.back[,1] * 100
dem3$SILT_UCOKALRBACK <- dem1.comp.back[,2] * 100
dem3$SAND_UCOKALRBACK <- dem1.comp.back[,3] * 100

dem3$CLAY_UCOKALRBACK[which(is.na(dem3$ELEV))] <- NA
dem3$SILT_UCOKALRBACK[which(is.na(dem3$ELEV))] <- NA
dem3$SAND_UCOKALRBACK[which(is.na(dem3$ELEV))] <- NA


ucok.dv <- predict.gstat(g, dv)
dv$CLAY_UCOKALR <- ucok.dv$CLAYalr.pred
dv$SILT_UCOKALR <- ucok.dv$SILTalr.pred

# backtranform (biased) = additive generalized logistic (agl) transform:
dv.alr <- matrix(c(dv$CLAY_UCOKALR, dv$SILT_UCOKALR),
  nrow = length(dv$CLAYalr), ncol = 2, dimnames = list(NULL, c("CLAY", "SILT")))
d.comp.back <- alrInv(dv.alr)
dv$CLAY_UCOKALRBACK <- d.comp.back[,1] * 100
dv$SILT_UCOKALRBACK <- d.comp.back[,2] * 100
dv$SAND_UCOKALRBACK <- d.comp.back[,3] * 100

rm(va,vm,vmf,va.cross,ucok.dv,ucok.out,d.comp.back,dem1.comp.back,g,
  va.cross.cok,g.cok,regr,regr2,regr.start,regr.start2)


#------------------------------------------------------------------------------
## Prediction maps:
#------------------------------------------------------------------------------

#catchm <- readOGR(dsn = "phd_calc_input", layer = "costara_catchm_v1")
#proj4string(catchm) <-
#  CRS("+proj=utm +zone=32 +ellps=WGS84 +datum=WGS84 +units=m +no_defs")

#pts.c <- list("sp.points", dc, pch = 19, col = "black", cex = .6, alpha = 1)
#pts.v <- list("sp.points", dv, pch = 21, col = "black", alpha = 1, cex = .6)

#cat <- fortify(catchm, region = "NAME")
#ptc <- as.data.frame(pts.c); ptv <- as.data.frame(pts.v)
z <- c("CLAY", "SILT", "SAND")
methods <- c("NNRCK", "IDW", "OK", "RK", "COK", "RCOK")
methods2 <- c("NNRCOKALRBACK", "IDW", "OK", "KED", "COKALRBACK", "UCOKALRBACK")
methods3 <-
  c("NNRCOKALRBACK", "IDWd", "OKd", "KEDd", "COKALRBACK", "UCOKALRBACK")
```

```
dem <- as.data.frame(dem3)

cmap <- list(NULL)
for (i in 1:length(z)) {
  #i <- 2
  if(z[i] == "SAND") methods2 <- methods3
  cdemm <- dem[,c(paste(z[i], methods2[1], sep = "_"),"x","y")]
  cdemm$METHOD <- methods[1]
  colnames(cdemm) <- c("VALUE", "x", "y", "METHOD")
  for (q in 2:length(methods)) {
    x <- dem[,c(paste(z[i], methods2[q], sep = "_"),"x","y")]
    x$METHOD <- methods[q]; colnames(x) <- c("VALUE", "x", "y", "METHOD")
    cdemm <- rbind(cdemm, x)
  }
  # grap the overall maximum value from all model predictions involved:
  max.max <- max(apply(dem[,c(paste(z[i], methods2, sep = "_"))], 2, max))
  # create the legend for the desired map collection (classify automatically):
  brks <- seq(0,ceiling(max.max/9) * 9, ceiling(max.max/9))
  brks1 <- brks[1:length(brks) - 1]; brks2 <- brks[2:length(brks)]
  cdemm$CUTV <- cut(cdemm$VALUE, breaks = brks)
  levels(cdemm$CUTV) <- c(paste0(brks1, " - ", brks2))
  # change plot order of facets by changing the order of levels with factor():
  cdemm$METHOD <- factor(cdemm$METHOD, levels = methods)

  cmap[[i]] <- ggplot(cat, aes(long, lat)) +
    geom_raster(aes(x, y, fill = CUTV), data = cdemm) +
    geom_point(mapping = aes(x = EAST, y = NORTH), size = 1,
    data = ptv, shape = 21) +
    geom_polygon(size = .5, linetype = "dashed", color = "black",
      fill = "grey40", alpha = 0) +
    coord_equal() +
    theme_bw(base_size = 12, base_family = "Helvetica") +
    scale_fill_brewer(name = "Content in %", palette = "YlOrBr") +
    labs(x = "\nUTM-E/m", y = "UTM-N/m\n") +
    scale_y_continuous(breaks = c(4361000,4363000)) +
    scale_x_continuous(breaks = c(508000,510000,512000,514000)) +
    facet_wrap(~ METHOD, ncol = 2) +
    theme(axis.text.y = element_text(angle = 90, hjust = .5),
      strip.text.x = element_text(size = 12))

    ggsave(paste(path.fig,
      paste("costara_modelcompMaps_", tolower(z[i]), "_v4.pdf", sep = ""),
        sep = "/"), cmap[[i]], width = 8.27, height = 6.19)
    print(cmap[[i]])
    dev.off()
}
rm(x,brks,brks1,brks2,cmap,i,max.max,methods2,q)
# remaining objects: z, methods, methods3, cdemm


```

```
#------------------------------------------------------------------------------
## Count of best predictions for different methods in percentage terms:
#------------------------------------------------------------------------------

# very similar to what is proposed by Vasat et al. 2013:
# Mapping the Topsoil pH and Humus Quality of Forest Soils in the North Bohemian
# Jizersk  hory Mts. Region with Ordinary, Universal, and Regression Kriging:
# Cross-Validation Comparison

meth.r <- 1:6
methods2 <- c("NNRCOKALRBACK", "IDW", "OK", "KED", "COKALRBACK", "UCOKALRBACK")

best.model <- rep(list(NULL), 3) # initialize a list
for (i in 1:length(z)) {
  for (q in 1:length(dv)) {
    res.p <- NULL
    for (t in 1:length(methods)) {
      if(z[i] == "SAND") methods2 <- methods3
      res.p[t] <- abs(
        dv@data[q,z[i]] - dv@data[q,paste(z[i], methods2[t], sep = "_")])
    }
    best.model[[i]][q] <- meth.r[which(res.p == min(res.p))]
  }
}

round(table(best.model[[1]])*100/length(dv), 0)
round(table(best.model[[2]])*100/length(dv), 0)
round(table(best.model[[3]])*100/length(dv), 0)

rm(i,q,res.p,meth.r,methods2,t) # keep: best.model


#------------------------------------------------------------------------------
## Histograms of model residuals:
#------------------------------------------------------------------------------

methods2 <- c("NNRCOKALRBACK", "IDW", "OK", "KED", "COKALRBACK", "UCOKALRBACK")
xh <- rep(list(data.frame(NNRCK = rep(1,length(dv)), IDW = rep(1,length(dv)),
  OK = rep(1,length(dv)), RK = rep(1,length(dv)),
  COK = rep(1,length(dv)), RCOK = rep(1,length(dv)))), 3)
for (i in 1:length(z)) {
  for (q in 1:length(methods2)) {
    if(z[i] == "SAND") methods2 <- methods3
    xh[[i]][q] <- dv@data[,paste(z[i], methods2[q], sep = "_")] - dv@data[,z[i]]
  }
}

hk <- 3.49 * sd(xh[[1]][,1]) * n[1]^(-1/3) # hk = 4.965 --> appr. 5

n.hist <- length(dv) * length(methods) * length(z)
```

```
j <- 1

# prepare data.frame for ggplot2:
hist <- data.frame(VALUE = rep(1,n.hist), METHOD = rep(1,n.hist),
  VARS = rep(1,n.hist))
for (t in 1:length(z)) {
  for (i in 1:length(methods)) {
    for (q in 1:length(dv)) {
      hist[j,"VALUE"] <- xh[[t]][q,i]
      hist[j,"METHOD"] <- methods[i]
      hist[j,"VARS"] <- z[t]
      j <- j + 1
    }
  }
}

hist[which(hist$VAR == "CLAY"),"VARS"] <- "Clay"
hist[which(hist$VAR == "SILT"),"VARS"] <- "Silt"
hist[which(hist$VAR == "SAND"),"VARS"] <- "Sand"

# change plot order of facet grid by changing the order of levels with factor():
hist$METHOD <- factor(hist$METHOD,
  levels = c("NNRCK", "IDW", "OK", "RK", "COK", "RCOK"))
hist$VARS <- factor(hist$VARS, levels = c("Clay", "Silt", "Sand"))

hist.r <- ggplot(data = hist, aes(x = VALUE)) +
  geom_histogram(aes(y = ..count..), binwidth = 5) +
  facet_grid(METHOD ~ VARS) +
  labs(x = "\nResidual content in %", y = "Counts\n") +
  scale_y_continuous(breaks = seq(0,15,5)) +
  scale_x_continuous(limits = c(-40,40), breaks = seq(-20,40,20)) +
  theme_bw(base_size = 12, base_family = "Helvetica") +
  theme(strip.text = element_text(size = 14),
    axis.text.y = element_text(size = 14),
    axis.title = element_text(size = 14),
    axis.text.x = element_text(size = 14),
    panel.grid.minor = element_blank())

# add mean and standard deviation as text elements into the histogram plots:
n.num <- length(methods) * length(z)

res.num <- data.frame(MEAN = rep(1,n.num), SD = rep(1,n.num),
  METHOD = rep(1,n.num), VARS = rep(1,n.num))
res.num$MEAN <- c(round(colMeans(xh[[1]]),3), round(colMeans(xh[[2]]),3),
  round(colMeans(xh[[3]]),3))
res.num$SD <- c(round(sapply(xh[[1]],sd),3), round(sapply(xh[[2]],sd),3),
  round(sapply(xh[[3]],sd),3))
res.num$METHOD <- rep(names(sapply(xh[[1]],sd)),3)
res.num$VARS <- c(rep("Clay",6), rep("Silt",6), rep("Sand",6))

```

```
hist.r2 <- hist.r +
  geom_text(data = res.num, aes(x = -38.5, y = 16.25,
    label = paste("mu == ", round(MEAN,2), sep = "")),
    vjust = 0.5, hjust = 0, parse = TRUE, size = 4) +
  geom_text(data = res.num, aes(x = -38.5, y = 13.25,
    label = paste("sigma == ", round(SD,2), sep = "")),
    vjust = 0.5, hjust = 0, parse = TRUE, size = 4)
print(hist.r2)

ggsave(paste(path.fig, "costara_resHists_v3.pdf", sep = "/"),
  hist.r2, width = 8.27, height = 11.69)
print(hist.r2)
dev.off()
rm(i,q,t,methods2,xh,hist,n.num,n.hist,j,hk,res.num,hist.r,hist.r2)


#------------------------------------------------------------------------------
## Evaluation of spatial agreement:
#------------------------------------------------------------------------------

methods2 <- c("NNRCOKALRBACK", "IDW", "OK", "KED", "COKALRBACK", "UCOKALRBACK")
# using kappa statistics (Cohen 1960):
# classify according to the USDA classification:
tri <- list()
tri.nm <- as.data.frame(matrix(ncol = 6,
  nrow = length(which(!is.na(dem3$CLAY_IDW)))))
colnames(tri.nm) = methods
for (i in 1:length(methods)) {
  df <- dem3@data[which(!is.na(dem3$CLAY_IDW)),
    c(paste(z[1], methods2[i], sep = "_"), paste(z[2], methods2[i], sep = "_"),
    paste(z[3], methods3[i], sep = "_"))]
  colnames(df) <- z
  tri[[i]] <- TT.points.in.classes(tri.data = df,class.sys = "USDA.TT",
    base.css.ps.lim = c(0,2,63,2000))
  for (q in 1:dim(tri[[1]])[1]) {
    tri.nm[q,methods[i]] <- names(which(tri[[i]][q,] == 1))
  }
} # this takes some time

# calculate Cohen's kappa statistic for each method-combination:
kappa <- matrix(ncol = length(methods), nrow = length(methods))
colnames(kappa) <- methods; rownames(kappa) <- methods; kappa

for (i in 1:length(methods)) {
  for (q in 2:length(methods)) {
    kappa[i,q] <- round(kappa2(tri.nm[,c(i,q)],"unweighted")$value,3)
  }
}
diag(kappa) <- 1.0; kappa[lower.tri(kappa)] <- NA; kappa
rm(i,q,tri,tri.nm,lang.par2,df) # keep: kappa
```

```
487
488
489 #----------------------------------------------------------------------------
490 ## Bubble plots of validation residuals:
491 #----------------------------------------------------------------------------
492
493 methods2 <- c("NNRCOKALRBACK", "IDW", "OK", "KED", "COKALRBACK", "UCOKALRBACK")
494 bmx <- dv; dv.id <- as.data.frame(dv[,1]); msc <- list(NULL)
495 for (i in 1:length(z)){
496   bmx@data[paste(z[i], "RES", sep = "_")] <-
497     dv@data[,z[i]] - dv@data[,paste(z[i], methods2[1], sep = "_")]
498   msc[[i]] <- as.data.frame(
499     cbind(bmx$ID, bmx@data[,paste(z[i], "RES", sep = "_")]))
500   colnames(msc[[i]]) <- c("PNTID", "RES")
501   msc[[i]]$EAST <- dv.id[,2]; msc[[i]]$NORTH <- dv.id[,3]
502   # define whether the model over- or under-estimates:
503   msc[[i]]$SIGN <- ifelse(msc[[i]]$RES > 0,
504     msc[[i]]$SIGN <- "Underestimated", msc[[i]]$SIGN <- "Overestimated")
505   msc[[i]]$VARS <- z[i] # needed for facets in ggplot2
506 }
507
508 msc3 <- rbind(msc[[1]], msc[[2]], msc[[3]])
509 msc3[which(msc3$VARS == "CLAY"),"VARS"] <- "Clay"
510 msc3[which(msc3$VARS == "SILT"),"VARS"] <- "Silt"
511 msc3[which(msc3$VARS == "SAND"),"VARS"] <- "Sand"
512 msc3$VARS <- factor(msc3$VARS, levels = c("Clay", "Silt", "Sand"))
513
514 pr.br <- pretty(c(min(abs(msc3$RES)), max(abs(msc3$RES))))
515
516 val.r <- ggplot(cat, aes(long, lat)) +
517   geom_point(data = msc3, aes(x = EAST, y = NORTH, size = abs(RES),
518     fill = SIGN), shape = 21, colour = "black") +
519   scale_size_area(breaks = pr.br, max_size = 9) +
520   geom_polygon(size = .5, linetype = "dashed", color = "black",
521     fill = "grey40", alpha = 0) +
522   facet_wrap(~ VARS, ncol = 2) +
523   coord_equal() +
524   theme_bw(base_size = 15, base_family = "Helvetica") +
525   guides(fill = guide_legend(order = 2, override.aes = list(size = 4,
526     shape = 21)), size = guide_legend(order = 1, override.aes = list(
527       colour = "black", shape = 21))) +
528   labs(x = "\nUTM-E/m", y = "UTM-N/m\n", fill = "Misprediction",
529     size = "Absolute residual\ncontent in %") +
530   scale_y_continuous(breaks = c(4361000,4363000)) +
531   scale_x_continuous(breaks = c(508000,510000,512000,514000)) +
532   theme(axis.text.y = element_text(angle = 90, hjust = .5, size = 14),
533     axis.title = element_text(size = 14),
534     axis.text.x = element_text(size = 14),
535     axis.title.x = element_text(hjust = .25),
536     strip.text.x = element_text(size = 15),
```

```
537     legend.position = c(.75,.25), legend.box = "horizontal",
538     panel.grid.minor = element_blank())
539
540 pnt511.lab <- data.frame(EAST = 511258, NORTH = 4362297,
541   VARS = factor("Clay",levels = c("Clay","Silt","Sand"))); pnt511.lab
542
543 val.r2 <- val.r +
544   geom_text(data = pnt511.lab, aes(x = EAST, y = NORTH, label = "511"), size =
        4)
545
546 ggsave(paste(path.fig, "costara_resBubble_v5.pdf", sep = "/"),
547   val.r2, width = 11.69, height = 7.07)
548 print(val.r2)
549 dev.off()
550 rm(msc3,bmx,i,methods2,msc,pr.br,val.r,val.r2)
551
552 rm(dem1.alr,dv.alr,z,methods,methods3)
553 save.image(paste(getwd(), "phd_calc_out/cos_modelcompResults.RData",sep = "/"))
554
555
556 #------------------------------------------------------------------------------
557 ## Examine OK/COK differences:
558 #------------------------------------------------------------------------------
559
560 ind <- which(dv$SAND >= 70); ind
561 dv[ind, c("ID","SAND","SAND_OKd","SAND_COKALRBACK")] # 513, 816
562 mean(abs(dv$SAND_OKd - dv$SAND_COKALRBACK)) # 1.53
563 mean(abs(dv$SAND_OKd[-ind] - dv$SAND_COKALRBACK[-ind])) # 1.41
564 dv@data[ind,"SAND_OKd"] - dv@data[ind,"SAND_COKALRBACK"]
565 # --> -4.43/-3.62; 88.64/80.70 (true) vs. 67.71/63.23 (OK) vs. 72.15/66.85 (COK)
566
567 plot(as.factor(dv$ID), dv$SAND)
568 points(dv$SAND_OKd, col = "red"); points(dv$SAND_COKALRBACK, col = "blue")
569
570 # end of script, costara_modelcomp20_vDiss.R, 25.07.2014
```

.4.18 costara_valid_vDiss.R

```
 1 #
 2 ##############################################################################
 3 #### Validation of various digital soil mapping techniques, Rio di Costara  ####
 4 ##############################################################################
 5 #
 6 ## last update: 07.05.2015
 7
 8 # validate neural network residual cokriging of alr-transformed soil separates
 9 #   compare with five common DSM techniques (IDW, OK, RK, COK, UCOK)
10
11 # author: Michael Blaschek (blaschek@geographie.uni-kiel.de)
```

```
12
13 rm(list = ls())
14
15 library(rgdal); library(gstat); library(RColorBrewer); library(compositions)
16 library(caret); library(ggplot2); library(spdep)
17 # R v3.0.2, gstat_1.0-16, rgdal_0.8-15, RColorBrewer_1.0-5, compositions_1.40-1
18 # caret_6.0-22, ggplot2_0.9.3.1, spdep_0.5-74
19
20 load(paste(getwd(), "phd_calc_out/cos_modelcompResults.RData", sep = "/"))
21
22
23 #--------------------------------------------------------------------------
24 ## Univariate validation (residual-based + association-based):
25 #--------------------------------------------------------------------------
26
27 source("valid_measures.R")
28
29 z <- c("CLAY", "SILT", "SAND")
30 methods <- c("NNRCK", "IDW", "OK", "RK", "COK", "RCOK")
31 methods2 <- c("NNRCOKALRBACK", "IDW", "OK", "KED", "COKALRBACK", "UCOKALRBACK")
32 methods3 <-
33   c("NNRCOKALRBACK", "IDWd", "OKd", "KEDd", "COKALRBACK", "UCOKALRBACK")
34
35 val.msr <- univar_error_metrics(data = dv@data, targets = z,
36   methods.nm = methods, methods.attr1 = methods2, methods.attr2 = methods3,
37   sp.obj = dv)
38
39
40 #--------------------------------------------------------------------------
41 ## Combined goodness of estimation measure:
42 #--------------------------------------------------------------------------
43
44 methods2 <- c("NNRCOKALRBACK", "IDW", "OK", "KED", "COKALRBACK", "UCOKALRBACK")
45 nm <- length(methods2)
46 stress.msr <- data.frame(STRESS = rep(1,nm)); rownames(stress.msr) <- methods
47
48 delta.ij <- NULL; delta.z0.ij <- NULL; j <- 2
49 for (q in 1:length(methods2)) {
50   x <- acomp(dv@data, parts = c("CLAY", "SILT", "SAND"), total = 100)
51   y <- acomp(dv@data, parts = c(paste(z[1], methods2[q], sep = "_"),
52     paste(z[2], methods2[q], sep = "_"),
53       paste(z[3], methods3[q], sep = "_")), total = 100)
54   delta.ij <- NULL; delta.z0.ij <- NULL; j <- 2
55   for (i in 1:(length(x[,1])-1)) {
56     delta.ij[i]  <- sum((clr(x[i,]) - clr(x[j,]))^2)
57     delta.z0.ij[i]  <- sum((clr(y[i,]) - clr(y[j,]))^2)
58     j <- j + 1
59   }
60   stress.msr[q,1] <- sqrt(sum((delta.ij - delta.z0.ij)^2)/sum(delta.ij^2))
61 }; stress.msr
```

```
rm(i,j,q,x,y,delta.ij,delta.z0.ij,methods,methods2)


#------------------------------------------------------------------------------
## Scatterplots of obs vs. pred regression:
#------------------------------------------------------------------------------

# NNRCK only:
lr1 <- lm((dv$CLAY ~ dv$CLAY_NNRCOKALRBACK))
lr2 <- lm((dv$SILT ~ dv$SILT_NNRCOKALRBACK))
lr3 <- lm((dv$SAND ~ dv$SAND_NNRCOKALRBACK))

methods <- c("NNRCK"); methods2 <- c("NNRCOKALRBACK")

yx <- NULL; yx.lm <- list(NULL); lines = list(NULL); y <- NULL; x <- NULL
for (i in 1:length(z)){
  y <- dv@data[,z[i]]
  x <- dv@data[,paste(z[i], methods2[1], sep = "_")]
  yx[[i]] <- data.frame(Y = y, X = x, VARS = z[i])
  yx[[i]]$VARS <- as.character(yx[[i]]$VARS)
  yx.lm[[i]] <- lm(y ~ x)
  # preparing elements for 1:1 and regression lines in scatterplots:
  lines[[i]] <- as.data.frame(cbind(as.numeric(yx.lm[[i]]$coefficients[2]),
    as.numeric(yx.lm[[i]]$coefficients[1]),
    as.numeric(summary(yx.lm[[i]])$r.squared)))
  colnames(lines[[i]]) <- c("SLOPE_LM", "INTERC_LM", "RSQUARED_LM")
}

yx3 <- rbind(yx[[1]], yx[[2]], yx[[3]])
line.e <- rbind(lines[[1]], lines[[2]], lines[[3]])

yx3[which(yx3$VARS == "CLAY"),"VARS"] <- "Clay"
yx3[which(yx3$VARS == "SILT"),"VARS"] <- "Silt"
yx3[which(yx3$VARS == "SAND"),"VARS"] <- "Sand"

line.e$VARS <- c("Clay", "Silt", "Sand")

yx3$VARS <- factor(yx3$VARS, levels = c("Clay", "Silt", "Sand"))
line.e$VARS <- factor(line.e$VARS, levels = c("Clay", "Silt", "Sand"))

regr.sc <- ggplot(data = yx3, aes(y = Y, x = X)) +
  geom_point(shape = 1, size = 2) +
  geom_abline(data = line.e,
    mapping = aes(slope = SLOPE_LM, intercept = INTERC_LM)) +
  geom_abline(
    yintercept = 0, slope = 1, linetype = "dashed", colour = "gray50") +
  scale_x_continuous(limits = c(0, 90)) +
  scale_y_continuous(limits = c(0, 90)) +
  labs(x = "\nPredicted value in %", y = "Observed value in %\n") +
  facet_wrap(~ VARS, ncol = 3) +
```

```
  coord_equal() +
  theme_bw(base_size = 12, base_family = "Helvetica") +
  theme(axis.text.y = element_text(size = 12),
    axis.title = element_text(size = 13),
    axis.text.x = element_text(size = 12),
    strip.text.x = element_text(size = 12),
    panel.grid.minor = element_blank())
print(regr.sc)

regr.sc2 <- regr.sc +
  geom_text(data = line.e, aes(x = 2.5, y = 85,
    label = paste0("y = ", round(SLOPE_LM, 2), "x", round(INTERC_LM, 2))),
    vjust = .5, hjust = 0, parse = FALSE, size = 4) +
  geom_text(data = line.e, aes(x = 87.5, y = 5,
    label = paste0("R^2 == ", round(RSQUARED_LM, 2))),
    vjust = .5, hjust = 1, parse = TRUE, size = 4)
print(regr.sc2)

ggsave(paste(path.fig, "costara_valScatter_v2.pdf", sep = "/"),
  regr.sc2, width = 8.27, height = 3.64)
print(regr.sc2)
dev.off()

rm(line.e,yx3,i,lines,lr1,lr2,lr3,
  regr.sc,regr.sc2,x,y,yx,yx.lm,z,methods,methods2,methods3,nm)
save.image(paste(getwd(), "phd_calc_out/cos_validResults.RData", sep = "/"))

# end of script, costara_valid_vDiss.R, 27.07.2014
```

.4.19 descr_statistics.R

```
#
################################################################################
####  Calculate summary statistics as part of an exploratory data analysis  ####
################################################################################
#
## last update: 26.09.2014

# author: Michael Blaschek (blaschek@geographie.uni-kiel.de)

summary_stats <- function(data, vars, export.latex = FALSE, path.to) {
 j <- c("Min", "Max", "1st Qu.", "Median", "3rd Qu.", "Mean", "Std",
   "Skewness", "Kurtosis", "Octile skew")
 r <- 1; descr.d <- NULL
 for (q in 1:length(data)) {
   dd <- data[[q]]; k <- 1; a <- 1
   descr <- data.frame(
     z, 1:length(z), 1:length(z), 1:length(z), 1:length(z), 1:length(z),
       1:length(z), 1:length(z), 1:length(z), 1:length(z))
```

```
19    names(descr) <- c("Target variables", j) # variable names
20    for (i in z) {
21      descr[k,"Min"] <- round(min(dd[,i]), digits = 2)
22      descr[k,"Max"] <- round(max(dd[,i]), 2)
23      descr[k,"1st Qu."] <- round(fivenum(dd[,i])[2], 2)
24      descr[k,"Median"] <- round(fivenum(dd[,i])[3], 2)
25      descr[k,"3rd Qu."] <- round(fivenum(dd[,i])[4], 2)
26      # fivenum --> hinges according to Tukey77, same as used by bwplot function
27      descr[k,"Mean"] <- round(mean(dd[,i]), 2)
28      descr[k,"Std"] <- round(sd(dd[,i]), 2)
29      descr[k,"Skewness"] <-
30       round((sum((dd[,i] - mean(dd[,i]))^3) / length(dd[,i])) / sd(dd[,i])^3, 2)
31      descr[k,"Kurtosis"] <- round(
32       ((sum((dd[,i] - mean(dd[,i]))^4) / length(dd[,i])) / sd(dd[,i])^4) - 3, 2)
33      a <- quantile(
34       dd[,i], probs = c(.125,.5,.875), na.rm = FALSE, names = TRUE, type = 7)
35      descr[k,"Octile skew"] <- round(((a[3]-a[2]) - (a[2]-a[1]))/(a[3]-a[1]), 2)
36      k <- k + 1
37    }
38    descr.d[[q]] <- descr
39  }
40  if (export.latex == TRUE) {
41    library(xtable) # xtable_1.7-3
42    descr.out <- NULL
43    for (i in 1:length(descr.d)) {
44      descr.out <- xtable(descr.d[[i]], caption = "Summary statistics",
45        label = "tab:descr")
46      print.xtable(descr.out, file = path.to[i])
47    }
48  }
49  return(descr.d)
50 }
51 # end of function: summary_stats, 26.09.2014
```

.4.20 pearsons_cor_coeffs.R

```
1  #
2  ##############################################################################
3  ####  Calculate Pearsons correlation coefficients and scatterplot matrices  ####
4  ##############################################################################
5  #
6  ## last update: 26.09.2014
7
8  # author: Michael Blaschek (blaschek@geographie.uni-kiel.de)
9
10 pearsons_corr <- function(data, covars, targets, scatterplot.matrix = FALSE,
11   cex.lab, path.to) {
12  cor.l <- NULL # initialize (cor)relation.(l)ist
13  for (q in 1:length(targets)) {
```

```
14    k <- 1; cor.c <- data.frame(covars, 1:length(covars), 1:length(covars), NA)
15    names(cor.c) <- c("COVAR", "PEARSON", "PVALUE", "SIGNIFICANCE")
16    for (i in covars) {
17      cor.test.p <- cor.test(data[,targets[q]],
18        data[,i], method = "pearson", alternative = "two.sided")
19      cor.c[k,"PEARSON"] <- round(cor.test.p$estimate, 4)
20      cor.c[k,"PVALUE"] <- round(cor.test.p$p.value, 4)
21      ifelse(cor.test.p$p.value <= .001, cor.c[k,"SIGNIFICANCE"] <- "***",
22       ifelse(cor.test.p$p.value > .001 && cor.test.p$p.value <= .01,
23         cor.c[k,"SIGNIFICANCE"] <- "**",
24        ifelse(cor.test.p$p.value > .01 && cor.test.p$p.value < .05,
25          cor.c[k,"SIGNIFICANCE"] <- "*", cor.c[k,"SIGNIFICANCE"] <- "-")))
26      k <- k + 1
27    }
28    cor.l[[q]] <- cor.c
29  }
30  if (scatterplot.matrix == TRUE) {
31    # plot optimized for 12 variables
32    all.vars <- c(targets, covars)
33    pdf(path.to, width = 15, height = 15, pointsize = 15)
34    pairs(formula(paste0(" ~ ", paste(all.vars, collapse = "+"))),
35      data = data, main = "",
36      diag.panel = panel.hist, cex.labels = cex.lab, xaxt = "n", yaxt = "n",
37      font.labels = 2, lower.panel = panel.cor, upper.panel = panel.smooth)
38    dev.off()
39  }
40  return(cor.l)
41 }
42 # end of function: pearsons_corr, 26.09.2014
43
44 # histogram on the diagonal:
45 panel.hist <- function(x, ...) {
46  usr <- par("usr"); on.exit(par(usr)); par(usr = c(usr[1:2], 0, 1.5))
47  hd <- hist(x, plot = FALSE)
48  breaks <- hd$breaks
49  nB <- length(breaks)
50  y <- hd$counts
51  y <- y/max(y)
52  rect(breaks[-nB], 0, breaks[-1], y, col = "gray80", ...)
53  box()
54 }
55
56 # correlation coefficient in the lower panel:
57 panel.cor <- function(x, y, digits = 2, prefix = "", cex.cor, ...) {
58  usr <- par("usr"); on.exit(par(usr)); par(usr = c(0, 1, 0, 1))
59  r <- round(cor(x, y, method = "pearson"), 2)
60  c.test <- cor.test(x, y, method = "pearson", alternative = "two.sided")
61  ifelse(c.test$p.value <= .001, c.signif <- "***",
62   ifelse(c.test$p.value > .001 && c.test$p.value <= .01, c.signif <- "**",
63    ifelse(c.test$p.value > .01 && c.test$p.value < .05, c.signif <- "*",
```

```
64      c.signif <- ""))
65  txt <- format(c(r, 0.123456789), digits = digits)[1]
66  txt <- paste(prefix, txt, sep = "")
67  if(missing(cex.cor)) cex <- 0.8/strwidth(txt)
68  text(0.5, 0.5, txt, cex = 1.9, font = 2)
69  text(1, .75, c.signif, cex = 1.9, pos = 2, col = "black")
70 }
71
72 # definition of points in the upper panel:
73 panel.smooth <- function (x, y,
74   col = "gray10", bg = NA, pch = 20, cex = 1, ...) {
75    points(x, y, pch = pch, col = col, bg = bg, cex = cex)
76 }
```

.4.21 kennard_stone_func.R

```
 1 #
 2 ##############################################################################
 3 #### Kennard-Stone algorithm to optimally split data into training/val. set ####
 4 ##############################################################################
 5 #
 6 ## last update: 06.09.2014
 7
 8 # replacing Euclidean with Mahalanobis distance in SPXY-method from Galvao
 9 #   et al. 05 - A method for calibration and validation subset partitioning
10
11 # author: Michael Blaschek (blaschek@geographie.uni-kiel.de)
12 # reference: Saptoro et al. 2012 - A Modified Kennard-Stone Algorithm for
13 #   Optimal Division of Data for Developing Artificial Neural Network Models
14
15 # R v3.0.2
16
17 ks.galv <- function(X, y, nc, dist.calc = "euclidean") {
18 dminmax <- rep(0, times = nc) # initializes the vector of minimum distance
19 M <- nrow(X) # collects the number of objects in X (= nr of given samples)
20 samples <- 1:M
21 Dx <- mat.or.vec(nr = M, nc = M) # initializes the matrix of X-distances
22 Dy <- mat.or.vec(nr = M, nc = M) # initializes the matrix of y-distances
23 D <- mat.or.vec(nr = M, nc = M)
24 poss.dist.calc <- c("euclidean", "mahalanobis")
25 if (!is.na(pmatch(dist.calc, poss.dist.calc))) {
26   if (pmatch(dist.calc, poss.dist.calc) == 1) {
27     for (i in 1:(M - 1)) {
28       xa <- X[i,]; ya <- y[i,]
29       for (j in (i + 1):M) {
30         xb <- X[j,]; yb <- y[j,]
31         Dx[i,j] <- sqrt(sum((xa - xb)^2))
32         Dy[i,j] <- sqrt(sum((ya - yb)^2))
33       }
```

```
    }
    Dxmax <- max(Dx) # returns the maximum value in Dx
    Dymax <- max(Dy)

    D <- Dx/Dxmax + Dy/Dymax # combines the X and y distances
  } else if (pmatch(dist.calc, poss.dist.calc) == 2) {
    # variance-covariance matrix:
    X <- cbind(X,y)
    Xm <- mat.or.vec(nr = (M * (M - 1))/2, nc = ncol(X))
    k <- 1
    for (i in 1:(M - 1)) {
      xa <- X[i,]
      for (j in (i + 1):M) {
        xb <- X[j,]
        Xm[k,] <- as.numeric(xa - xb)
        k <- k + 1
      }
    } # calculates an n*(n-1)/2 x ncol(X) matrix
    # each row represents the difference of a variable between two points
    n.covar <- ncol(X)
    vr <- apply(Xm, 2, var)
    vcm <- mat.or.vec(nr = n.covar, nc = n.covar)
    for (i in 1:(n.covar - 1)) {
      xu <- Xm[,i]
      for (j in (i + 1):n.covar) {
        xv <- Xm[,j]
        vcm[i,j] <- cov(xu, xv)
      }
    } # covariance written into upper-triangle of the variance-covariance matrix
    diag(vcm) <- vr
    vcm[lower.tri(vcm, diag = F)] <- t(vcm)[lower.tri(vcm, diag = F)]
    for (i in 1:(M - 1)) {
      xa <- X[i,]
      for (j in (i + 1):M) {
        xb <- X[j,]
        D[i,j] <-
         sqrt(rowSums((as.numeric(xa - xb) %*% solve(vcm))*as.numeric(xa - xb)))
      }
    }
  }
} else {
  stop("invalid distance method")
}

# stepwise selection procedure similar to the KS algorithm:
maxD <- apply(D, 2, max)
index_row <- apply(D, 2, function(x) which(x == max(x))[1])
index_column <- which(maxD == max(maxD))

m <- c(index_row[index_column], index_column)
```

```
for (i in 3:nc) {
  pool <- setdiff(samples, m)
  dmin <- rep(0, times = M - i + 1)
  for (j in 1:(M - i + 1)) {
    indexa <- pool[j]
    ddd <- rep(0, times = i - 1)
    for (k in 1:(i - 1)) {
      indexb <- m[k]
      if (indexa < indexb) {
      ddd[k] <- D[indexa, indexb]
      } else {
      ddd[k] <- D[indexb, indexa]
      }
    }
    dmin[j] <- min(ddd)
  }
  index <- which(dmin == max(dmin))
  m[i] <- pool[index]
}; return(m)
}
# end of script, kennard_stone_func.R, 14.07.2013
```

.4.22 regr_diagnostics.R

```
#
################################################################################
#### Calculate selected numerical regression diagnostics + test statistics  ####
################################################################################
#
## last update: 22.02.2015

# author: Michael Blaschek (blaschek@geographie.uni-kiel.de)

library(car); library(spdep); library(MASS)
# restricted to one SpatialPointsDataFrame per run

regr_diagn <- function(lm.obj, sp.obj, vars, nk) {
  j <- c("W", "p<W"); k <- 1
  diagn.f <- data.frame(vars, 1:length(vars), 1:length(vars))
  names(diagn.f) <- c("Target variables", j) # variable names
  for (q in 1:length(lm.obj)) {
    dr <- lm.obj[[q]]; ds <- summary(dr)
    # Shapiro-Wilk normality test:
    diagn.f[k,"W"] <- round(shapiro.test(ds$residuals)$statistic, 4)
    diagn.f[k,"p<W"] <- round(shapiro.test(ds$residuals)$p.value, 4)
    # Breusch-Pagan test for homoscedasticity:
    diagn.f[k,"ChiSquare"] <- round(ncvTest(dr)$ChiSquare, 4)
    diagn.f[k,"p-value"] <- round(ncvTest(dr)$p, 4)
    # Moran's I to check remaining spatial autocorrelation in lm-residuals:
```

```
26      sp.obj$OLSRES <- dr$residuals
27      nlistk <- knn2nb(knearneigh(sp.obj, k = nk))
28      w <- nb2listw(nlistk, style = "B")
29      mi <- moran.test(
30        sp.obj$OLSRES, w, randomisation = FALSE, alternative = "two.sided")
31      diagn.f[k,"MORAN I"] <- round(mi$estimate[1], 4)
32      diagn.f[k,"STDEV"] <- round(mi$statistic, 4)
33      diagn.f[k,"p>STDEV"] <- round(mi$p.value, 4)
34      # residual summary statistics:
35      diagn.f[k,"R_MEAN"] <- round(mean(dr$residuals), 2)
36      diagn.f[k,"R_SD"] <- round(sd(dr$residuals), 2)
37      diagn.f[k,"R_MIN"] <- round(min(dr$residuals), 2)
38      diagn.f[k,"R_MAX"] <- round(max(dr$residuals), 2)
39      diagn.f[k,"STUD_MEAN"] <-round(mean(lmwork(dr)$studres), 2)
40      diagn.f[k,"STUD_SD"] <- round(sd(lmwork(dr)$studres), 2)
41      diagn.f[k,"STUD_MIN"] <- round(min(lmwork(dr)$studres), 2)
42      diagn.f[k,"STUD_MAX"] <- round(max(lmwork(dr)$studres), 2)
43      # Bonferroni p-values to evaluate studentized residuals for being outliers:
44      oT <- outlierTest(dr, cutoff = .05)
45      if (oT$signif == FALSE) print("No outliers detected!")
46      # maximum observed hat leverage value:
47      diagn.f[k,"HAT_MAX"] <- round(max(hatvalues(dr)), 4)
48      hat.val <- hatvalues(dr)
49      t.3 <- 3 * ((length(dr$coefficients) - 1) + 1)/dim(sp.obj)[1] # Belsley80
50      leverage.id <- as.vector(which(hat.val > t.3))
51      leverage.ID <- sp.obj$ID[leverage.id]
52      print(paste0("IDs of noteworthy hat values/leverage points: ", ifelse(
53        length(leverage.ID) == 0, "None", paste(leverage.ID, collapse = ", "))))
54      # effects of outliers + leverage points, detecting influential:
55      r.cooksD <- cooks.distance(dr)
56      t <- 4/(length(r.cooksD) - ((length(dr$coefficients) - 1) + 1))
57      influential.id <- as.vector(which(r.cooksD > t))
58      influential.ID <- sp.obj$ID[influential.id]
59      print(paste0("IDs of influential observations: ", ifelse(length(
60        influential.ID) == 0, "None", paste(influential.ID, collapse = ", "))))
61      k <- k + 1
62    }
63    return(diagn.f)
64  }
65  # end of function: regr_diagn, 17.01.2015
```

.4.23 valid_measures.R

```
1  #
2  ##############################################################################
3  #### Compute univariate error measures based on separate validation sample  ####
4  ##############################################################################
5  #
6  ## last update: 06.04.2015
```

```
7
8 # author: Michael Blaschek (blaschek@geographie.uni-kiel.de)
9
10 library(spdep)
11
12 univar_error_metrics <- function(data, targets, methods.nm,
13   methods.attr1 = methods.nm, methods.attr2 = methods.attr1, sp.obj) {
14
15   nm <- length(methods.nm)
16   val.msr <- rep(list(data.frame(RMSE = rep(1,nm), BIAS = rep(1,nm),
17     MIN = rep(1,nm), MAX = rep(1,nm), MEAN = rep(1,nm), STD = rep(1,nm),
18     SKEW = rep(1,nm), r2 = rep(1,nm), EF = rep(1,nm), W = rep(1,nm),
19     pW = rep(1,nm), I = rep(1,nm), STDEV = rep(1,nm), pSTDEV = rep(1,nm))),
20     length(z))
21   rownames(val.msr[[1]]) <- methods.nm; rownames(val.msr[[2]]) <- methods.nm
22   rownames(val.msr[[3]]) <- methods.nm
23   nlistk <- knn2nb(knearneigh(sp.obj, k = 4))
24
25   # calculate validation measures:
26   # RMSE = root mean squared error,
27   # r2 = coefficient of determination, EF = Nash-Sutcliffe coefficient
28   sum.lm <- NULL; x <- NULL; y <- NULL
29   for (i in 1:length(targets)) {
30     for (q in 1:length(methods.attr1)) {
31       if(targets[i] == "SAND") methods.attr1 <- methods.attr2
32       x <- data[,paste(targets[i], methods.attr1[q], sep = "_")]
33       y <- data[,targets[i]]
34       val.msr[[i]][q,"MIN"] <- round(min(x - y), digits = 2)
35       val.msr[[i]][q,"MAX"] <- round(max(x - y), digits = 2)
36       val.msr[[i]][q,"MEAN"] <- round(mean(x - y), digits = 2)
37       val.msr[[i]][q,"STD"] <- round(sd(x - y), digits = 2)
38       val.msr[[i]][q,"SKEW"] <- round(
39         (sum(((x - y) - mean((x - y)))^3) / length((x - y))) / sd((x - y))^3, 2)
40       val.msr[[i]][q,"RMSE"] <- round(sqrt(mean((x - y)^2)), digits = 3)
41       val.msr[[i]][q,"BIAS"] <- round((mean(x) - mean(y)), 3)
42       obs.vs.pred <- lm((y ~ x)); sum.lm <- summary(obs.vs.pred)
43       val.msr[[i]][q,"r2"] <- round(sum.lm$r.squared, digits = 2)
44       val.msr[[i]][q,"EF"] <-
45         round((1 - (sum((y - x)^2) / sum((y - mean(y))^2))), 2)
46       val.msr[[i]][q,"W"] <- round(shapiro.test(x - y)$statistic, 2)
47       val.msr[[i]][q,"pW"] <- round(shapiro.test(x - y)$p.value, 3)
48       sp.obj.2 <- sp.obj
49       sp.obj.2$MODRES <- x - y
50       w <- nb2listw(nlistk, style = "B")
51       mi <- moran.test(
52         sp.obj.2$MODRES, w, randomisation = FALSE, alternative = "two.sided")
53       val.msr[[i]][q,"I"] <- round(mi$estimate[1], 2)
54       val.msr[[i]][q,"STDEV"] <- round(mi$statistic, 2)
55       val.msr[[i]][q,"pSTDEV"] <- round(mi$p.value, 3)
56     }
```

```
57   }
58   return(val.msr)
59 }
60 # end of function: univar_error_metrics, 06.04.2015
```

Appendix C

.5 Soil data

Appendix C provides all measurement values with regards to soil textural fractions of the top 30 cm soil layer at both the field and the landscape scale. Covariates from digital elevation models and geological maps can be reproduced using the R-scripts given in the previous section. Meteorological time series as well as geophysical data can be made available on demand.

Table 3: Measured soil data at field 21

ID	X	Y	CLAY	SILT	SAND
1	508826	4362469	23.14	23.18	53.69
2	508833	4362538	30.46	24.28	45.26
3	508828	4362591	28.97	26.05	44.99
4	508828	4362643	34.95	31.47	33.58
5	508828	4362696	37.77	28.57	33.66
6	508823	4362745	38.02	26.90	35.07
14	508950	4362596	27.55	24.12	48.33
15	508908	4362648	33.67	26.75	39.59
16	508936	4362591	24.51	23.82	51.65
17	508886	4362591	41.22	22.41	36.37
18	508908	4362543	21.43	27.83	50.74
19	508859	4362541	27.28	24.77	47.95
20	508834	4362493	26.06	23.32	50.61
21	508809	4362542	32.93	22.80	44.28
22	508758	4362543	32.75	24.24	43.01
23	508784	4362595	34.88	28.67	36.44
24	508733	4362593	35.28	30.98	33.76
25	508758	4362646	34.00	34.51	31.49
26	508807	4362644	27.17	34.87	37.96
27	508857	4362644	30.45	33.58	35.97
28	508783	4362694	38.39	29.03	32.57
29	508759	4362742	33.67	29.74	36.59

Table 3 – (Continued)

ID	X	Y	CLAY	SILT	SAND
30	508808	4362742	35.91	28.07	36.03
51	508908	4362571	24.99	19.70	55.31
52	508927	4362603	27.32	23.13	49.55
53	508859	4362497	25.82	21.27	52.91
54	508871	4362525	26.81	19.31	53.88
55	508812	4362509	32.07	19.08	48.85
56	508776	4362758	35.22	28.15	36.63
57	508856	4362705	43.78	19.98	36.24
58	508771	4362768	34.69	28.98	36.33
59	508746	4362627	42.23	25.06	32.72
60	508779	4362667	44.52	24.46	31.02
61	508907	4362628	38.32	22.24	39.44
62	508778	4362727	38.61	25.42	35.97
63	508715	4362604	40.88	27.11	32.01
64	508713	4362623	35.63	25.88	38.49
66	508809	4362594	37.27	26.03	36.70
67	508848	4362636	37.02	22.87	40.11
68	508844	4362609	34.43	31.23	34.34
69	508773	4362618	39.12	25.22	35.66
71	508850	4362662	38.35	27.19	34.47
72	508884	4362664	32.79	25.55	41.66

Table 4: Measured soil data at field 33

ID	X	Y	CLAY	SILT	SAND
1	508970	4363005	20.94	27.82	51.24
2	509035	4363005	27.56	23.02	49.42
3	508903	4362936	39.43	12.94	47.63
4	508970	4362939	33.09	18.28	48.63
5	509036	4362936	38.71	14.88	46.41
6	509098	4362939	43.42	15.77	40.81
7	508842	4362873	46.49	20.45	33.06
8	508904	4362872	51.63	13.47	34.90
9	508968	4362873	43.61	22.21	34.19
10	509035	4362873	29.41	19.64	50.94
11	509165	4362873	40.32	24.15	35.54
12	508839	4362808	41.19	22.86	35.95
13	508906	4362807	41.25	21.83	36.93
14	508966	4362807	32.09	23.66	44.26
15	509036	4362807	43.20	13.53	43.26
16	509101	4362808	32.88	23.61	43.51
17	509165	4362809	34.64	14.04	51.32
18	508905	4362746	39.93	21.56	38.51
19	509039	4362745	43.75	19.89	36.36
20	509099	4362743	41.90	14.47	43.63
21	508969	4362679	38.72	24.74	36.54
22	509036	4362677	48.55	22.67	28.77
51	508849	4362816	37.80	22.06	40.13

Table 4 – (Continued)

ID	X	Y	CLAY	SILT	SAND
52	508898	4362779	27.56	35.86	36.58
53	508808	4362815	42.34	24.35	33.31
54	508978	4362824	34.33	22.32	43.35
55	508889	4362813	42.30	20.36	37.34
56	508858	4362891	49.65	17.24	33.12
57	508913	4362836	43.58	20.10	36.32
58	508833	4362888	45.11	23.65	31.25
59	508942	4362762	40.34	22.52	37.14
60	509118	4362759	42.21	16.97	40.82
61	508992	4362739	42.48	25.33	32.19
62	508995	4362657	41.45	24.35	34.20
63	509032	4362790	44.09	10.79	45.12
64	509038	4362836	32.60	18.36	49.05
65	509058	4362675	44.54	23.28	32.17
66	508935	4362733	40.89	23.73	35.37
67	509080	4362695	40.93	20.11	38.95
68	509134	4362870	45.32	28.33	26.35
69	509008	4362899	25.32	19.38	55.31
70	509018	4362958	39.38	20.03	40.59
71	508890	4362941	48.03	20.11	31.85
72	508934	4362912	47.89	18.12	33.99
73	508993	4362926	48.89	11.75	39.36
74	509139	4362893	38.74	18.95	42.30
75	508942	4362988	32.28	19.80	47.92
76	508933	4362694	46.65	16.29	37.06
77	508901	4362700	39.75	17.07	43.18
91	508855	4362763	45.20	20.29	34.51
92	508902	4362718	36.59	18.33	45.08
93	508930	4362789	46.08	16.51	37.41
94	509011	4362696	48.91	22.36	28.73
95	509156	4362825	45.27	24.30	30.42
96	509085	4362897	27.83	23.71	48.46
97	508953	4362645	29.79	23.21	46.99
98	508821	4362850	42.34	23.71	33.94
37	508818	4362796	39.74	25.25	35.01
38	508813	4362835	36.46	32.07	31.46
39	508803	4362869	37.69	33.20	29.11
40	508870	4362787	39.94	22.21	37.85
41	508934	4362717	38.81	24.83	36.36
42	509005	4362643	35.71	29.57	34.71
43	508973	4362611	32.71	25.08	42.21

Table 5: Measured soil data at the Rio di Costara test site

ID	X	Y	CLAY	SILT	SAND
101	508891	4363407	36.79	25.31	37.91
102	508430	4363418	11.26	30.70	58.04
103	508837	4363576	17.55	30.27	52.17

Table 5 – (Continued)

ID	X	Y	CLAY	SILT	SAND
104	509037	4363142	17.23	26.22	56.55
106	508695	4363256	33.04	28.71	38.25
107	510628	4362992	44.87	19.34	35.78
108	510685	4362948	40.27	22.76	36.97
109	507794	4362948	21.02	28.85	50.13
110	508180	4362760	24.53	21.53	53.94
111	510695	4362704	38.91	18.21	42.88
112	509541	4362777	15.93	33.45	50.61
113	510620	4362795	39.72	27.65	32.63
114	509018	4362470	38.32	21.24	40.44
115	509175	4362558	38.50	29.19	32.31
116	511882	4362507	43.45	23.89	32.66
117	511258	4362597	44.29	26.56	29.15
118	509210	4362510	43.44	26.22	30.33
119	509233	4362329	28.68	30.05	41.28
120	510565	4362182	23.77	38.89	37.34
121	511169	4362052	32.95	24.62	42.43
122	507859	4362036	22.79	27.59	49.63
123	510923	4361955	58.44	15.98	25.57
124	508272	4362963	27.25	29.69	43.07
125	507673	4361875	33.84	32.15	34.01
126	510498	4361774	26.71	35.39	37.90
127	508463	4362037	39.72	25.86	34.43
128	511913	4361795	44.41	15.93	39.66
129	510794	4361699	41.26	25.14	33.60
131	508017	4361559	49.87	14.99	35.15
132	508322	4361448	32.98	22.10	44.92
133	508474	4361345	37.38	25.96	36.66
134	511584	4361460	43.65	21.65	34.70
135	509094	4361338	27.39	32.08	40.53
136	509948	4361158	34.49	27.34	38.17
137	508570	4361412	28.70	32.88	38.41
138	508560	4361317	25.49	26.57	47.93
139	511550	4361092	41.33	28.38	30.30
140	512740	4361117	32.88	23.86	43.25
141	510563	4360890	36.19	40.76	23.05
142	513094	4361055	37.66	24.60	37.73
143	509927	4360871	36.84	30.84	32.32
144	512578	4360796	26.38	33.31	40.31
145	509468	4361555	38.86	31.25	29.88
146	511084	4360722	39.62	16.77	43.61
147	509440	4360871	45.82	30.64	23.54
149	509899	4360448	20.53	49.81	29.65
150	510013	4360776	39.90	32.22	27.89
151	510043	4360416	45.05	27.19	27.76
152	510197	4360457	21.37	50.23	28.40
154	508807	4363043	46.50	17.62	35.88
155	513605	4362828	18.16	12.69	69.14
156	508819	4362826	37.58	25.11	37.31
157	508460	4362688	37.95	30.15	31.90

Table 5 – (Continued)

ID	X	Y	CLAY	SILT	SAND
158	514010	4362862	12.10	11.31	76.59
159	513485	4362623	7.03	8.51	84.46
160	508344	4362527	25.62	19.67	54.72
161	512081	4362497	27.70	19.89	52.41
162	511079	4362363	50.45	23.79	25.75
163	509102	4362360	31.59	28.62	39.78
164	513080	4362399	16.43	11.33	72.25
165	513806	4362575	15.93	16.45	67.62
166	511894	4362300	32.69	20.12	47.19
167	510844	4362198	30.99	23.00	46.00
168	512020	4362155	32.91	20.40	46.69
169	512518	4362151	28.11	14.09	57.79
170	513572	4361983	30.43	19.26	50.31
171	512565	4361985	39.98	17.34	42.67
172	514082	4361919	22.38	18.90	58.72
173	513845	4362030	39.59	27.05	33.35
174	507853	4361816	36.03	30.94	33.03
175	513590	4361816	23.79	23.14	53.07
176	507605	4361779	30.30	31.55	38.15
177	513160	4361859	28.78	19.00	52.22
178	513143	4361786	39.78	20.22	40.00
179	511747	4361781	41.70	29.99	28.31
180	513449	4361756	29.78	22.81	47.41
181	511119	4361865	35.82	30.04	34.14
182	512941	4361614	23.48	24.41	52.11
183	509933	4361635	40.35	27.41	32.25
184	511440	4361605	37.89	25.26	36.85
185	513367	4361631	22.76	34.58	42.66
186	510742	4361725	39.04	32.82	28.13
187	513103	4361482	39.05	24.38	36.57
188	512285	4361808	41.70	23.97	34.33
189	513066	4361377	35.15	22.64	42.21
190	512622	4361446	34.90	21.93	43.18
191	510839	4361479	45.27	24.00	30.73
192	508658	4361460	39.28	26.07	34.65
193	511699	4360882	35.27	31.37	33.36
194	512797	4361473	34.33	21.77	43.90
195	510512	4360677	40.80	32.19	27.01
196	510945	4360565	32.62	33.96	33.42
197	512631	4360714	31.89	20.30	47.80
198	510471	4360495	30.79	22.38	46.83
199	510747	4360562	27.75	32.64	39.60
200	509142	4361106	29.89	25.09	45.02
201	509995	4360639	21.37	21.02	57.61
202	510339	4360719	24.72	25.39	49.89
203	509481	4360908	39.59	29.75	30.65
204	510171	4360569	30.90	37.20	31.90
205	511681	4360385	41.25	27.95	30.80
207	509282	4363335	15.34	29.47	55.20
208	509198	4363106	10.61	25.63	63.76

Table 5 – (Continued)

ID	X	Y	CLAY	SILT	SAND
216	511526	4363379	17.79	36.25	45.96
217	509394	4363318	20.26	25.04	54.70
219	512010	4363677	19.38	32.89	47.73
220	510615	4363205	21.74	31.93	46.32
222	511617	4363180	20.61	39.26	40.13
224	511378	4363056	19.87	27.45	52.68
226	510859	4363164	16.10	22.00	61.90
229	511473	4362994	25.92	32.36	41.72
231	513276	4362957	16.41	35.76	47.82
232	509600	4363345	19.98	29.73	50.29
233	510489	4363019	18.38	31.07	50.55
234	512122	4363051	19.31	26.62	54.08
236	510933	4363253	31.80	38.47	29.73
238	509781	4362815	18.86	27.98	53.16
240	510379	4362688	19.93	29.27	50.79
241	509742	4362817	21.77	31.91	46.32
242	509359	4362657	33.22	29.02	37.76
243	509960	4362278	37.35	29.88	32.76
501	507971	4362839	27.57	22.12	50.31
502	508069	4362626	30.32	17.82	51.86
503	510503	4362182	28.75	36.42	34.83
504	509438	4362247	25.40	36.26	38.34
505	509235	4361713	39.68	25.70	34.62
506	512233	4361533	31.16	23.41	45.42
507	511808	4361681	42.73	22.21	35.06
508	509777	4361485	33.12	27.75	39.13
509	512526	4360777	36.28	19.22	44.50
510	512483	4362396	25.23	14.45	60.32
511	511533	4362047	17.02	14.29	68.68
512	510383	4362029	25.36	36.41	38.23
513	513620	4362675	5.20	6.16	88.64
514	509277	4361034	37.10	42.59	20.31
515	511726	4360768	37.21	27.41	35.38
516	509180	4363554	21.82	25.24	52.94
517	509158	4363233	21.44	18.00	60.56
518	512168	4363579	17.39	31.12	51.48
520	511697	4363238	19.11	36.89	44.00
521	512234	4363208	15.31	32.21	52.48
523	511492	4363074	14.91	27.59	57.51
525	512417	4362877	19.65	28.57	51.78
526	512720	4362861	18.01	30.36	51.63
527	512753	4362760	25.07	20.63	54.30
700	511166	4362101	37.41	28.53	34.06
701	511230	4360502	42.14	22.78	35.08
702	511136	4360551	45.10	22.16	32.74
703	511476	4360461	34.61	24.10	41.29
704	510411	4360778	36.23	29.79	33.98
705	509497	4361037	40.49	32.14	27.37
706	509572	4361070	49.18	31.68	19.14
707	508009	4362483	27.84	27.63	44.53

Table 5 – (Continued)

ID	X	Y	CLAY	SILT	SAND
708	507525	4362089	29.11	25.67	45.22
709	511156	4362161	41.44	23.54	35.02
710	511159	4362089	36.48	23.46	40.06
711	511308	4362649	40.71	26.78	32.51
712	511397	4362640	44.13	29.84	26.03
713	510849	4362979	32.27	27.98	39.75
714	509642	4362077	43.59	27.09	29.31
715	510590	4362015	24.60	32.94	42.45
716	509872	4361935	23.78	32.86	43.36
717	511253	4360405	38.56	34.71	26.73
718	512996	4361056	38.44	24.86	36.70
719	509846	4360607	37.98	25.24	36.78
720	510635	4361659	55.69	27.98	16.33
722	510555	4361200	34.88	29.59	35.53
723	513205	4362359	21.85	17.44	60.72
724	513364	4361846	33.83	27.23	38.94
725	512180	4361918	31.88	31.56	36.56
726	512252	4361859	50.72	18.06	31.22
727	508950	4362695	40.71	24.22	35.07
728	512714	4361820	21.84	26.66	51.50
729	508318	4362158	26.46	32.87	40.67
730	508512	4362405	26.24	20.71	53.05
731	509845	4361749	41.76	28.23	30.01
732	508761	4361786	36.19	20.79	43.02
733	511346	4360923	26.36	33.39	40.25
734	511458	4360982	42.29	22.82	34.89
735	512313	4361110	63.19	28.08	8.73
801	512744	4361226	48.66	21.98	29.35
802	510426	4361459	36.87	31.27	31.86
803	510406	4360889	45.13	28.35	26.51
804	510927	4360875	38.03	27.35	34.61
805	511038	4361353	57.33	27.95	14.72
806	511176	4360586	43.08	22.20	34.72
807	511598	4360811	48.97	25.78	25.25
808	511215	4361710	42.18	25.89	31.94
809	511934	4361430	23.02	27.77	49.21
811	512873	4360909	42.49	22.34	35.17
812	511999	4361879	44.86	22.19	32.95
813	511530	4361470	52.93	20.81	26.25
814	513455	4361849	40.52	31.16	28.32
815	513519	4362245	42.01	17.40	40.59
816	513679	4362815	9.88	9.42	80.70
817	508107	4362121	35.61	20.87	43.52
818	508370	4362297	39.15	23.06	37.78