Politecnico di Milano

School of Industrial and Information Engineering

Master of Science in Mathematical Engineering

Geostatistical analysis for Uncertainty Quantification in the SMART-SED model: a downscaling approach based on Digital Soil Mapping data

Supervisor
Prof. Alessandra MENAFOGLIO

Candidate
Niccolò TOGNI

Co-Supervisors
Prof. Luca BONAVENTURA
Dr. Davide BRAMBILLA

Academic Year 2018-2019

Abstract

SMART-SED is a project aimed at developing an innovative framework for the numerical simulation of sediment motion in river catchments, intended to be used by local territorial management institutions and professionals to design proper strategies for the mitigation of hydrogeological instability. Uncertainty analysis is an intrinsic feature of models simulating natural processes. In order to perform an effective uncertainty quantification, it is necessary to properly identify the variability of the input parameters and to design stochastic simulation methods able to provide realistic realisations, based on the available data. This thesis focuses on the use of digital soil maps for the prediction and stochastic simulation of terrain-related quantities used for the estimation of the input parameters of the SMART-SED model. The digital maps are obtained from SoilGrids, a system for automated soil mapping based on state-of-the-art spatial predictions methods. Innovative approaches are introduced to account for the limitations of SoilGrids data (low resolution, inaccuracy) and for the specificities of the variables in exam. Although the focus is on the SMART-SED project, the methods proposed can be generally used for geostatistical modelling at a local scale using auxiliary coarse information obtained through remote sensing or from previously fitted digital maps.

Contents

List of Figures

List of Tables

Acknowledgements

This work marks the conclusion of my university experience. As whiny and pathetic as it may sound, I faced some serious difficulties along the way, and I wouldn't have been able to overcome them if it weren't for the guidance and support received from the amazing people that became part of my life, or were there since the beginning.

I would first like to express my gratitude to my supervisor, Professor Alessandra Menafoglio, for the guidance and the trust she has been providing me during the development of this thesis, and to my co-supervisors, Professor Luca Bonaventura and Doctor Davide Brambilla, for the advices and the great assistance.

I wish to say *thank you* to all my university friends from École Centrale Paris: the Italians of promotions 2016,2017 and 2018, the people of CPI (Centrale Paris International), and all those who made the campus *soirées* unforgettable.

Thank you to my friends from Politecnico di Milano, too many to be listed.

Thank you to AIM (Associazione Ingegneri Matematici), and all the people who supported me as a member of the directive board.

Thank you to my travel buddies, for making Slovakia feel like a beautiful place.

Thank you to the Far Blue crew, for being like a family to me (and for the vacations in Costa Smeralda).

Thank you to my only high school friend (I wasn't very popular back then).

Thank you to my pals ”i ragazzi di Varazze”, I hope you'll never change.

Thank you to my amazing high school teachers, for transmitting your passion and making me eager to learn.

And finally a big *thank you* to all my family: my parents, my grandparents, my aunts, my brother and cousins.

Introduction

SMART-SED (Sustainable MAnagement of sediment transpoRT in responSE to climate change conDitions) is a project having for objective the development of a modelling framework allowing for the numerical simulation of soil erosion and sediment motion over a catchment. It will serve as a support tool for local authorities, allowing to design proper strategies for the prevention and control of damages from hydrogeological calamities.

The SMART-SED model relies on a more coherent approach with respect to preexisting methods and implements a more efficient and robust discretisation technique, a detailed explanation of the model is available in [6].

An important aspect of SMART-SED is **Uncertainty Quantification (UQ)**. In fact, it has become a standard practice for any numerical simulations of real-world phenomena to provide an extensive analysis of the uncertainty of the outputs related to the variability of the inputs. The classical methods to perform this task include Monte Carlo (MC) simulations and/or metamodelling techniques such as multi-fidelity Co-Kriging ([37], [27]). A crucial aspect of UQ is the analysis of the variability of the input parameters. A probabilistic analysis of the of the outcomes of the model cannot be made without first identifying the range of the possible inputs and their likelihood. The importance and the difficulty of this task increases with the number of parameters involved and their complexity.
The SMART-SED model aims at a more coherent approach for the simulation several hydrological phenomena occurring in a basin. For it to produce realistic and informative result, quantitative and qualitative information on the physical properties of the soil is required. In particular, there are two properties of the terrain that allow to properly characterise its hydrological response: *soil thickess*, meaning the depth of the permeable layer of soil from its surface to the underlying bedrock, and *soil texture*, which is determined by *particle-size fractions*, *i.e.* the relative percentages, in terms of soil composition, of *clay, silt and sand* (the three categories in which grains of fine earth are divided depending on their size, [1]). To obtain useful results, a proper estimation of these physical quantities that accounts for their variability over the whole domain is needed.

Obviously it would be impossible to directly measure the above-mentioned quantities in every location of the domain. When some sparse samples are available, geostatistical techniques such as *Kriging* can be used to interpolate the available observations and obtain estimates and their associated uncertainty.

Unfortunately, direct observations are not always available over the whole area. Measuring soil properties is expensive and time consuming since it requires costly devices and specialised professionals to operate them. One of the ambitions of SMART-SED is to reduce to a minimum the request of local surveys by using public databases of soil properties already available online.

Recently, the IRSIC (Word Soil Information) started a very ambitious project called **SoilGrids**. The latter is a public database containing worldwide estimations for numerous standardised soil properties with a resolution of 250 meters. These estimations were obtained using state-of-the-art machine learning algorithms trained with hundreds of thousands observations of hundreds of covariates [34]. SoilGrids is a public project still in progress: the precision of the estimations increases as new data are added to training sets. At present (2019) the level of accuracy is already largely satisfactory for many soil properties.

SoilGrids data have already been extensively used by many researchers, mainly for analysis at a large scale, some examples including [75] and [18]. Because of their coarse resolution and their limited accuracy, SoilGrids maps have not received the same attention when dealing with local scale analysis. The goal of this thesis is to show that SoilGrids data can prove useful even at local scales, and to provide a statistical framework for the prediction and stochastic simulation of soil-related quantities (in particular particle-size fractions and soil thickness) at a finer scale using coarse resolution information, a problem usually referred-to as **statistical downscaling**. The case study considered in this work involves the hydro-graphic basin of the river Caldone in the Northern Italy city of Lecco (Fig. 1).

Traditional downscaling procedures do not take into account the *compositional nature* of particle-size-fractions. An innovative approach is proposed to perform downscaling of particle-size fractions based on the application **Area-to-Point Kriging** [39] and derived techniques in the context of the **Aitchison geometry** [3] through the use of *isometric log-ratio transformations* [17], in order to fully account for the particular structure of the sample space in a coherent and efficient way.

As regards to soil thickness, downscaling is done with a different method called **dissever** [44, 68]. Soil thickness strongly depends on slope, elevation and other topographic variables whose measured values are generally available up to very fine resolutions (few meters): dissever uses this fine-scale secondary

Figure 1: Picture of the river Caldone in the city of Lecco, overflowing after heavy rainfalls

information to perform the downscaling.

Several proposals are made on how to integrate SoilGrids data and direct measurements to improve the accuracy of the estimations and better quantify their variability.

The statistical analysis has been done in R-3.6 [64] using libraries *gstat* [60, 30] for Area-to-Point Kriging and geostatistical simulation; *dissever* [44], for the homonymous algorithm; *compositions* [7] for the analysis of compositional data. In particular, the procedure of *variogram deconvolution* described in [23] was implemented.

This thesis is organised as follows. In Chapter 1, a list of the main parameters involved in the SMART-SED model is presented; it is also shown how to properly estimate them using soil properties. In chapter 2, SoilGrids is introduced and an exploratory analysis of the variables of interest is carried out. Chapter 3 contains the detail of the geostatistical techniques used for downscaling and simulation. In this chapter a novel approach for the downscaling of compositional data is proposed. The presented methods are then applied to SoilGrids maps of soil thickness and particle-size fractions, the results are shown in Chapter 4. The last chapter contains a discussion on the results and how to use them for UQ in the SMART-SED model, along with some proposals on how to integrate direct measurements from field surveys in the analysis.

Chapter 1

Soil properties and SMART-SED parameters

This chapter is organised as follows: first, the SMART-SED project is introduced and an overview of the case-study area is made. The SMART-SED model equations are then presented, the main parameters are listed and their link to soil properties is highlighted. The last two section are dedicated to the introduction of soil texture and soil thickness, the two variables which will be the object of the geostatistical analysis in the next chapters.

1.1 SMART-SED

Hydrogeological instability is and has always been a major concern for inhabited areas due to its negative consequences as a natural hazard for people and infrastructures. In recent years, all the processes typically included in this broad definition have been receiving increasing attention in light of the growing occurrence of calamitous events that is frequently explained with a climate-changing context. Nowadays, *"smart cities"* manage to prevent and/or control natural hazards by implementing sustainable strategies via advanced technology and innovation. Remarkably, many smart cities located in flood or landslide-prone areas still lack an appropriate consideration of hazards related to sediment instabilities and consequent problems.**SMART-SED (Sustainable MAnagement of sediment transpoRT in responSE to climate change conDitions)** is a project financed by *Cariplo foundation* that aims at filling this gap by providing local territorial management institutions with an advanced decision-making support tool regarding hydrogeological hazards.

In practical terms, the goal of the project is the development of a freeware containing a distributed model for sediment motion along a river catchment.

The SMART-SED model aims at improving the pre-existing tools thanks to the following characteristics:

1. ability to incorporate several physical processes in a coherent way;
2. a numerical scheme rigorously conserving mass;
3. automatic recognition of slope and drainage locations;
4. computational efficiency;
5. an extensive **uncertainty analysis.**

In the context of numerical simulation of real-world phenomena, the sources of uncertainty are numerous. The main ones can roughly be divided in three different categories ([69], [76]):

- *Structural uncertainty*: this includes the errors and approximations attributable to the model assumptions.
- *Numerical uncertainty*: the approximation error due to the discretisation for the numerical computation. This is the only error term that can be directly estimated with a certain accuracy, and the one with the least impact on the overall uncertainty of the outcomes.
- *Parametric uncertainty*: input data are seldom constant and known, in most cases they have an intrinsic variability, this is generally the main source of uncertainty of the simulations.

Uncertainty quantification (UQ) of computer codes is the process of identifying the possible outcomes and their likelihood depending of the variability of the parameters and on the other sources of error. UQ is a key aspect of simulation tools used in the decision-making process regarding matters of public safety. This thesis will focus on the estimation of the variability of soil-related input parameters of the model, which is a vital step in UQ.

1.1.1 Case study: the Caldone basin

For the case study the investigators of the project have chosen the hydrographic basin of the river Caldone near the town of Lecco, Northen Italy. Lecco is crossed by three streams (the Bione, Caldone, and Gerenzone) that have the typical characteristics of torrents in a pre-Alpine area. The Caldone has been already object of a field investigation in 2016. The hydro-graphic basin of this water course is 24 km^2 wide, with an altitude ranging from 197 m a.m.s.l. to 2118 m a.m.s.l. at the top of Grigna Meridionale mountain.

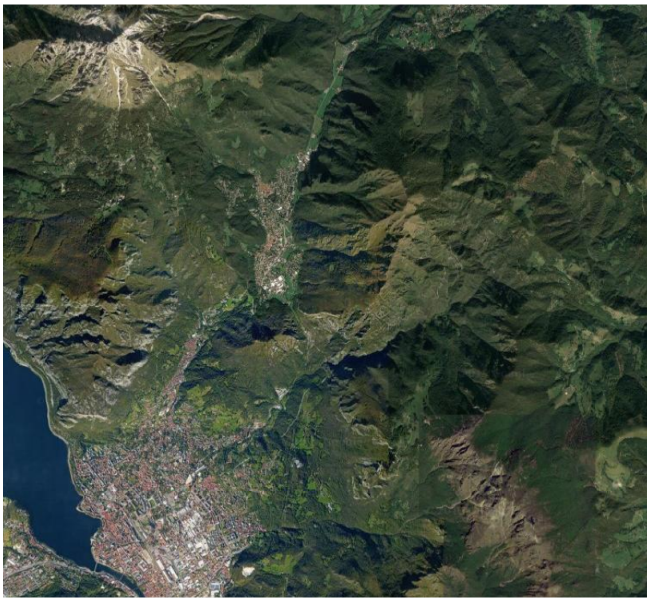


Figure 1.1: Aerial view of the study area

Geologically, the basin is characterised by rocky outcrops in the higher part (mainly limestone and clastic rock), while downstream towards the city the river flows through a floodplain. The average precipitation over the city of Lecco is about 1400 mm/y. The Caldone river flows from Mount Resegone and, just before entering the city, receives the water from the Grigna torrent. From that point on, the river flows through the town, mostly within artificial banks. Waters are withdrawn by industries in the surroundings (mostly for cooling machinery) and by residential buildings. On the other hand, the stream receives a significant amount of water from the sewer network that drains the (mostly impermeable) town area. In its last kilometer before the outlet into the Lario lake, the Caldone flows within a culvert that passes below the town centre. The combination between a short hydrologic time of response, high slope, intense sediment transport and flow within a densely urban area makes the Caldone river a suitable case study for hydro-geological instability and hazard. Two sediment retention basins (Fig. 1.2) are present in the last 5 km of the river, with volumes of around 8000 m^3 each. Quantitative data is available on the sediment-supplied volumes during short or large time spans, this information along with periodic surveys of the bathymetry of the retention basins allows to validate the modelling tools developed by SMART-SED.

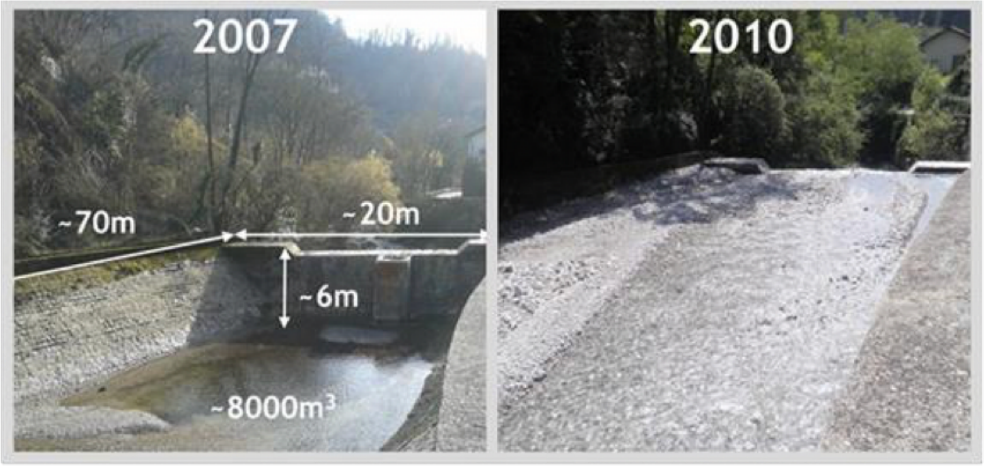
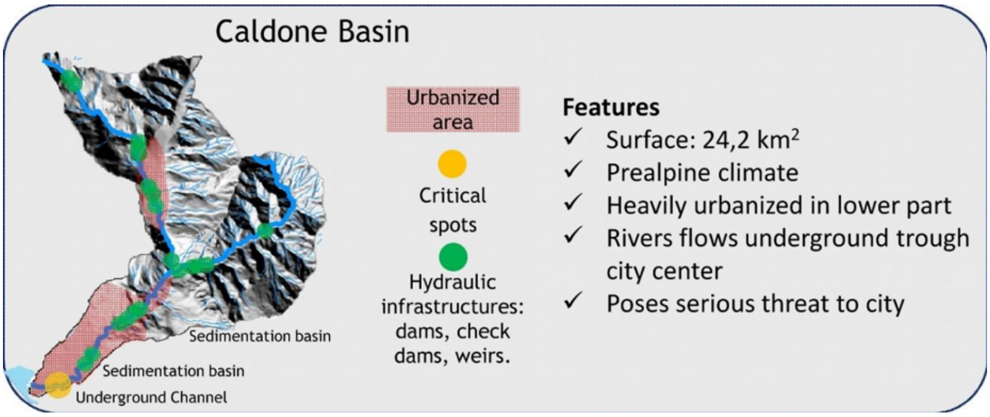


Figure 1.2: Retention Basin on Caldone River, topped up in three years



Caldone Basin

▨ Urbanized area

● Critical spots

● Hydraulic infrastructures: dams, check dams, weirs.

Features

- ✓ Surface: 24,2 km^2
- ✓ Prealpine climate
- ✓ Heavily urbanized in lower part
- ✓ Rivers flows underground trough city center
- ✓ Poses serious threat to city

Figure 1.3: Main features of the Caldone river basin

1.1.2 Model equations

SMART-SED equations model the variation in time of different layers of the terrain (including surface water) induced by precipitation phenomena, a reduced version of the model containing the most relevant equations for this thesis work is hereby presented. The complete model and all the references can be found in [2, 28].

Consider a domain $\Omega = [0, L_x] \times [0, L_y]$ which contains a basin subdomain $\Omega_b \subset \Omega$, defined by geometric considerations, and a drainage subdomain $\Omega_d \subset \Omega_b$, whose extension varies in time and which is only implicitly defined as the portion of Ω_b where the depth of the surface water layer H is above a minimum threshold. For $x \in \Omega_d$, we model the motion of the surface water layer by the Saint-Venant equations:

$$\begin{aligned} \frac{\partial H}{\partial t} &= -\nabla \cdot (H\mathbf{u}) + (1-\mu)P + E - I \\ \frac{\partial \mathbf{u}}{\partial t} &= -g\nabla\eta - \mathbf{u} \cdot \nabla\mathbf{u} - \gamma(\mathbf{u})\mathbf{u} \end{aligned} \tag{1.1}$$

Here b denotes the topographic profile, η is the height of water-free surface, so that $H = \eta - b \quad x \in \Omega_d$, $\mathbf{u}$ is the surface water velocity, $\gamma(\mathbf{u})$ is the friction coefficient. Concerning the source term, P is the precipitation intensity in $[m/s]$, μ is a non-dimensional parameter that takes the value of 1 if the ground temperature is lower or equal than the melting temperature T_m and 0 if it is higher. E and I are the ground **exfiltration** and **infiltration** terms, respectively. They represent the mass exchanges of water between the surface layer and the gravitational layer, to be defined in the following. It is assumed that the topographic profile is not changing in time, so that

$$\frac{\partial H}{\partial t} = \frac{\partial \eta}{\partial t}.$$

This simplification is justified in the limit of thin sediment layers.

The model is completed by a number of equations for the time evolution of the equivalent depths of other two-dimensional, vertically averaged water and sediment layers, all of which are defined for $x \in \Omega_b$, more specifically:

- a snow layer with equivalent depth h_{sn};
- a sediment layer with equivalent depth h_{sd};
- a gravitational layer with equivalent depth h_g.

For each of these layers, conservation of mass is assumed. For the the mass exchanges among layers and for the horizontal mass fluxes, relatively simple models are employed. Each of these could be replaced by more sophisticated approaches. The model equations read

$$\begin{aligned}\frac{\partial h_{sn}}{\partial t} &= \mu P - S \\ \frac{\partial h_{sd}}{\partial t} &= -\nabla \cdot \mathbf{f}_s(H\mathbf{u}) + W \\ \frac{\partial h_g}{\partial t} &= -\nabla \cdot \mathbf{f}_g(h_g) + S + I - E\end{aligned}$$

Each equation is now discussed in greater detail, starting from the topmost layer, the models employed to compute the exogenous source terms are also reviewed.
The atmospheric component is not modelled directly, but is instead assumed to be a reservoir of infinite capacity. Water leaves this reservoir through precipitation (snow or rain), which is characterised by intensity, duration and spatial distribution.

Precipitation can take the form of rain or snow, depending on the surface temperature. Rain occurs if the temperature is higher then the melting threshold of $T_m = 2°C$. In this case, water is assumed to end up in the surface run-off layer. In the opposite case, precipitation takes the form of snow and is being accumulated at the surface until temperature reaches values high enough to cause melting. The snow layer height in $[m]$ is denoted by h_n. S is the snow melting rate $[m/s]$, computed according to the Degree-Day approach:

$$S = \delta(T - T_m),$$

where δ is a parameter that determines the amount of snow that melts in one day at a given temperature T.

The sediment flux depends on the presence of run-off and on its discharge $\mathbf{f}_{sd}(H\mathbf{u})$. This correlation is expressed in the Grass formula [26] for x and y direction, respectively:

$$f_{sdx} = a_s u|u|^{b_s} \qquad f_{sdy} = a_s v|v|^{b_s},$$

where a_s is an empirical coefficient that depends on the grain diameter (usually a value between 0 and 1 is taken). The exponent b_s is also empirical and takes value between 0 and 3. In the Grass model, the critical shear stress is set to zero, so the sediment movement begins simultaneously with the water movement. The sediment source term W, expressed in $[m/s]$, is defined according to the **Gavrilovic approach** [13]. It corresponds to the rate of sediment production due to erosive processes as a result of precipitation, computed as:

$$W = \pi(1-\mu)P\tau_g Z^{3/2}.$$

The term τ_g is the temperature coefficient and is given by the following formula:

$$\tau_g = [(T/10) + 0.1]^{1/2},$$

where T is the mean annual temperature of the basin. The term Z is the *erosion coefficient* and can be computed as follows

$$Z = XY(\xi + S^{1/2}). \tag{1.2}$$

X, Y and ξ are respectively the **soil protection coefficient**, the **erodibility coefficient** and the kind and extent of erosion coefficient. These are empirical parameters that depend on the soil coverage and its composition, they are usually considered constant on the whole catchment. Although the Gavrilovic method gives results on yearly basis, it is assumed that it is also valid for shorter periods in which W will be seen as an intensity.

The gravitational layer is the soil portion in which water can move due to gravitational forces. This movement is governed mainly by the permeability of the soil and the horizontal fluxes are modelled in terms of the terrain slopes. h_g is the water content in gravitational zone $[m]$, which is limited by the maximum value $h_{g,max}$, a spatially variable quantity that depends on **soil thickness**. $\mathbf{f}_g$ are the horizontal fluxes that are formed inside the gravitational zone that govern the movement of water mass $[m^2/s]$. They are defined as

$$\mathbf{f}_g(h_g) = h_g\mathbf{u}_g,$$

where $\mathbf{u}_g$ represents the water velocity vertically averaged over the layer. This velocity is modeled as $\mathbf{u}_g = \beta_g(h_g, x, y)\mathbf{n}$, where β_g is a function of the soil characteristics and of the water level in the layer, while $\mathbf{n}$ is the unit vector determined by the terrain slope b direction $\mathbf{n} = \frac{\nabla b}{\|\nabla b\|}$.

The gravitational layer exchanges mass directly with the surface layer. This process is represented by the ground exfiltration and infiltration terms E and I, respectively. Exfiltration occurs when the maximum water storage capacity in gravitational zone is reached and the excess becomes run-off:

$$\begin{aligned} E &= 0 \quad \text{if} \quad h_g \leq h_{g,max} \\ E &= \frac{\partial(h_g - h_{g,max})}{\partial t} \quad \text{if} \quad h_g > h_{g,max}. \end{aligned} \tag{1.3}$$

The infiltration rate I $[m/s]$ from the surface run-off to the gravitational layer describes how fast water can enter terrain from surface. It mainly depends on the first layer of terrain features, in particular **soil texture**, vegetation cover and actual saturation degree. The infiltration rate in dry soil is denoted f_0, infiltration in saturated soil is denoted f_c. Transition from f_0 to f_c is managed by SMART-SED model using the *SCS-CN (Soil Conservation Service - Curve Number) modified method* [49]. As suggested by the name, the SCS-CN method relies on the **Curve Number** (CN), a spatially varying empirical value that depends on soil coverage and soil texture. As the CN, the parameters f_0 and f_c depend on soil texture, in particular the f_c value is equal to the **Darcy permeability**, used to model underground flow.

To sum up, many of the "free" parameters depend on the soil characteristics, especially soil texture (Y, ξ, β_g, f_0, f_c, CN), some of them depend on soil coverage (X, CN) and the (crucial) parameter $h_{g,max}$ is linked to soil thickness. The following three sections are devoted to the presentation of these three very important soil characteristics.

1.2 Soil coverage

Soil coverage generally refers to the land use and/or the type and the quantity of vegetation covering the topmost layer of the terrain. There is no universal codification for the possible types of soil coverage, since the relevant aspects depend on the context, so that different categories are defined in different applications.

Table 1.1 contains the reference values for the soil protection coefficient of the Gavrilovic method, used in the formula (1.2) (cf. [14]); these values are linked to the type and density of vegetation cover of the soil. The *Curve Number* also depends on soil coverage.

Soil protection coefficient X	
Mixed and dense forest	0.05-0.2
Low density forest with grove	0.05-0.2
Coniferous forest with little grove, scarce bushes, bush prairie	0.2-0.4
Damaged forest and bushes, pasture	0.4-0.6
Damaged pasture and cultivated land	0.6-0.8
Areas w/o vegetation cover	0.8-1.0

Table 1.1: Table of soil protection coefficient X

For the Italian region of Lombardy, a geographical database which classifies the territory based on the principal types of use and coverage is available. The database is kept up-to-date and is usually referred to by the name *DUSAF (Destinazione d'Uso dei Suoli Agricoli e Forestali).* DUSAF soil coverage classification is based on five levels, from more general to more specific.

LIVELLI DUSAF					
I	II	III	IV	V	CODE
Antropizzato	Urbanizzato	Tessuto Continuo	Denso		1111
Antropizzato	Urbanizzato	Tessuto Discontinuo	Residenziale		1121
Antropizzato	Urbanizzato	Tessuto Discontinuo	Sparso	Cascine	11231
Antropizzato	Produttivo	Grandi impianti servizi	Insediamenti e annessi	Agricoli	12112
Antropizzato	Produttivo	Grandi impianti servizi	Grandi impianti	cimitero	12124
Antropizzato	Verde non agricolo	Urbano	Parchi /giardini		1411
Agricolo	Seminativo	Seminativo	Orto familiare		2115
Agricolo	Coltura permanente	Frutteti			222
Agricolo	Prati permanenti	Prati permanenti	Alberi e arbusti e sparsi		2312
Seminaturale	Boscato	Latifoglie	Media e alta densità		31111
Seminaturale	Boscato	Latifoglie	Media e alta densità		31112
Seminaturale	Vegetazione arbustiva/erbacea	Prateria d'alta quota	No alberi e arbusti		3211

Table 1.2: Example of DUSAF classes of soil use

In 2004 Rosso [67] produced a reference table which assigns to any codified type of soil coverage of the DUSAF database *four* possible values of the CN parameter, depending on the hydrogeological properties of the underlying topsoil. The DUSAF classes can also be easily matched to the classes of the Gavrilovic reference table for the soil protection coefficient, allowing to identify the tabulated value for the area.

Raster maps with resolution of $5m$ containing the value of the DUSAF code at each point are available for the whole Lombardy region, and hence for the case study area. Figure 1.4 shows the different areas corresponding to different DUSAF codes thus having a different soil coverage. These data are used for the determination of the Gavrilovic parameter X, although the information is not sufficient for the determination of CN maps, since additional information on soil texture is required. The following section explains how the texture of the soil is determined.

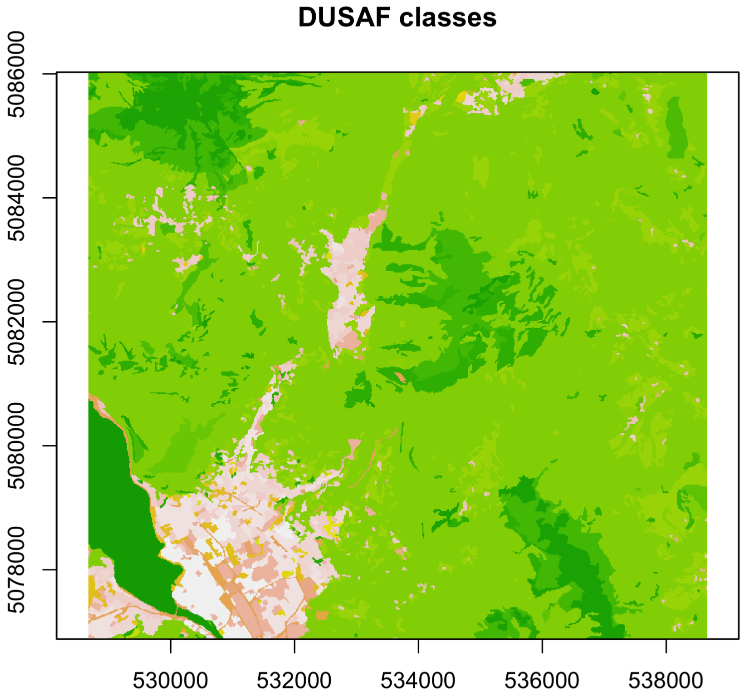


Figure 1.4: Map of DUSAF classes on Lecco area
The map shows the whole catchment in exam, different colours correspond to different DUSAF classes. Green colours are associated to wooded areas, whereas grey/brownish colours are associated to urbanised zones.

Figure 1.5: Example of soil texture

1.3 Soil texture

Soil texture is a classification instrument used to determine soil classes based on their physical texture [1]. More specifically, soil texture is quantitatively determined on the basis of the relative fractions of the fine particles of different sizes that compose the terrain. Soil particles under 2 *mm* are divided in three groups:

- clay: particles with a diameter less than 2 μm;
- silt: particles with a diameter comprised between 2 μm and 50 μm;
- sand: particles with a diameter comprised between 50 μm and 2 *mm*.

Fractions of clay, silt and sand are usually indicated with the acronym **psf (particle-size fractions).** Soil texture classes are determined by the relative percentages of clay/silt/sand, according to a standard that may vary depending on the country.

The most common classification is the one used by the *United States Department of Agriculture (USDA)*, which distinguishes twelve major soil texture classes shown in Figure 1.6. The classes are typically named after the primary constituent particle-size or a combination of the most abundant particles sizes, e.g. *sandy clay* or *silty clay.* A fourth term, loam, is used to describe equal proportions of sand, silt, and clay in a soil sample, and leads to the naming of even more classes, e.g. *clay loam* or *silt loam.*

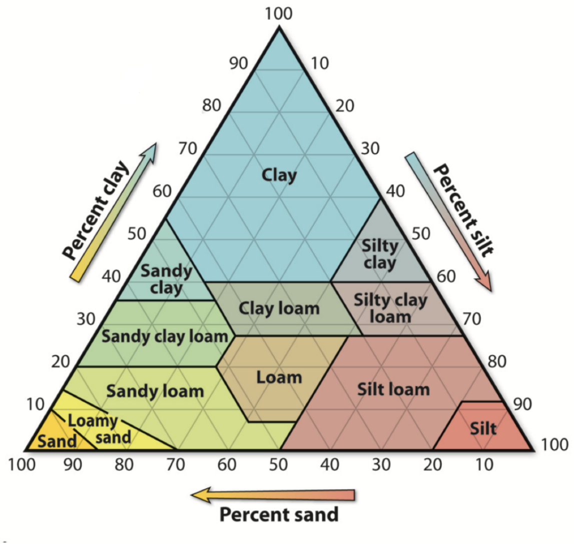


Figure 1.6: Soil texture triangle
Soil texture classification according to the USDA classification system, based on relative fractions of clay, silt and sand.

Soil texture is of paramount importance for the characterisation of the hydrological properties of the soil.

As mentioned in the previous section, Rosso [67] created a reference table for the estimation of the *Curve Number*. For each type of soil coverage, the table provides four possible values associated to the four classes of soil shown in Table 1.3.

A	Soils with high infiltration rates in moist conditions and low runoff potential, includes deep *sand*, *loamy sand*, *sandy loam* (with very low proportion of clay and silt) and gravel. Highly permeable soils with transmission rates over 7.6 mm/h. Very high infiltration capacity at saturation
B	Soils with moderate infiltration rates in moist conditions and moderate runoff potential, includes most sandy soils sufficiently deep and drained (less deep than group A), with moderately fine and moderately coarse texture. Transmission rates between 3,8 and 7,6 mm/h. High infiltration capacity at saturation.
C	Soils with low infiltration rates in moist conditions and moderately high runoff potential, includes *sandy clay loam* with elevate proportion of clay and silt and generally a fine texture. Transmission rate between 1,3 and 3,8 mm/h. Low infiltration capacity at saturation.
D	Soils with very low infiltration rates in moist conditions and high runoff potential, includes shallow *clay*, *clay loam*, *sandy clay* and *silty clay*. Very low transmission rate (0 - 1,3 mm/h). Very low infiltration capacity at saturation.

Table 1.3: Table of the four hydrological categories defined by Rosso

As a consequence, information on soil texture at each point of the area in exam would allow to better estimate the CN number.

In a similar way, the Gavrilovic parameters Y and ξ have reference tabulated values (Table 1.4); information on the texture of the topsoil would prove useful for a better identification of the relevant values for the basin.

Soil erodibility coefficient Y	
Hard, erosion-resistant rock	0.2-0.6
Rock with moderate erosion resistance	0.6-1
Weak rock, stabilised	1-1.3
Sediments, moraines, clay and other rock with little resistance	1.3-1.8
fine sediments and soils without erosion resistance	1.8-2
Coefficient of type and extent of erosion ξ	
Little erosion on watershed	0.1-0.2
Erosion in waterways on 20-50% of the catchment area	0.3-0.5
Erosion in rivers, gullies and alluvial deposits, karstic erosion	0.6-0.7
50-80% of catchment area affected by surface erosion and landslides	0.8-0.9
Whole watershed affected by erosion	1

Table 1.4: Table of Gavrilovic parameters

Particle-size fractions are used to estimate soil permeability, for instance using the Kozeny-Carman equations ([47], [80]), permeability is necessary to estimate the SMART-SED parameters β_g, f_0, and f_c.

In order to measure particle-size fractions it is necessary to collect samples from the ground and to perform laboratory tests on the soil particles. The collection can be made using the equipment shown in Figure 1.7 (double ring infiltrometer). Soil texture measurement is a costly, time-consuming operation which requires on-field surveys with specialised devices and laboratory analysis; for this reason it is only possible to provide information on few locations of the basin area. Usually, only a limited number of observations are collected and subsequently used to infer soil texture on the whole domain through the use of geostatistical estimation techniques such as **Kriging**, which will be presented in chapter 3.

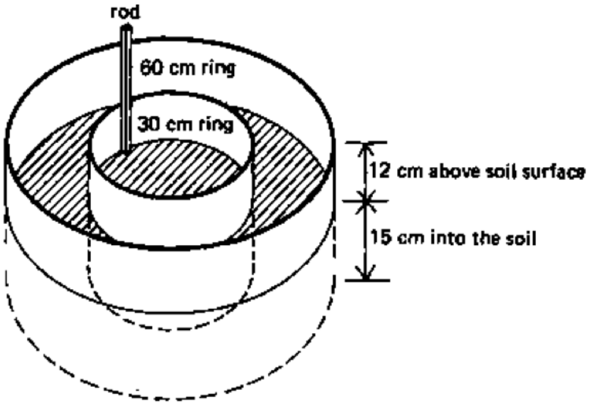


Figure 1.7: Double ring infiltrometer

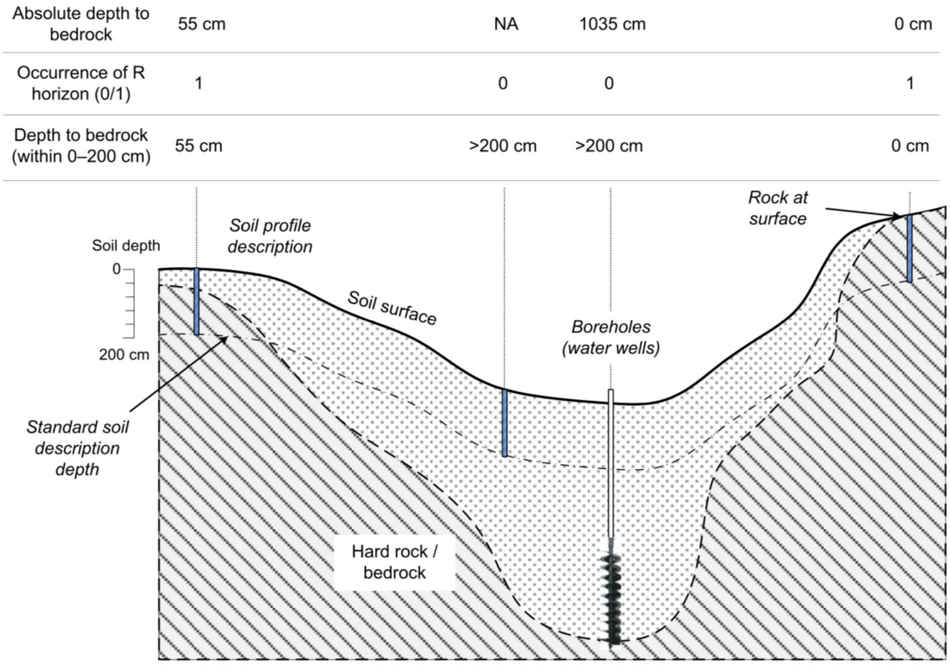
1.4 Soil thickness

One of the critical parameters of the SMART-SED model is $h_{g,max}$, namely the maximum quantity of water storable in the gravitational soil. This parameter depends on soil porosity and on its thickness, meaning the depth of the permeable layer of fine earth (also called *active layer*) over the underlying **bedrock** (also called **R horizon**).

Measuring soil thickness is a difficult process that requires the use of drills or other expensive equipment; for this reason in some cases a constant value is considered for the whole catchment [73]. This might be justified on particularly homogeneous planar areas, but catchments with a complex topography (like the one we are examining) are characterized by greatly varying soil depths.

When direct measurements are possible there are two types of approaches that can be chosen to infer soil thickness on the whole basin:

- *A model-based approach:* soil is mobile and undergoes transportation phenomena driven by topographical variables. Physical models of these geo-morphological processes properly calibrated by the means of dedicated measurements can be used to estimate the thickness of the active layer, an example is given in [61].
- *A statistical approach*: regression techniques could be used instead, the covariates considered are generally topographical variables such as elevation, slope, aspect and curvature. A relevant example is the work [56], who identified a significant linear relationship between soil thickness and curvature. Other authors such as [41] compared several machine learning methods and used seven terrain variables, all related to the topography.



Absolute depth to bedrock	55 cm	NA	1035 cm	0 cm
Occurrence of R horizon (0/1)	1	0	0	1
Depth to bedrock (within 0–200 cm)	55 cm	>200 cm	>200 cm	0 cm

Figure 1.8: Absolute depth to bedrock, censored depth to bedrock, probability of occurrence of R horizon. Source: [71]
Schematic explanation of the depth to bedrock. The R-horizon is the dashed line at the interface with the hard rock or bedrock.

Soil thickess is a critical parameter in many applications, so the problem of its estimation has received a lot of attention in the literature. Accurate soil thickness maps are difficult to obtain, especially in mountain areas with steep slopes, peaks and valleys, since in these cases soil thickness presents a great spatial variability with very high values in the valleys (even hundreds of meters) and values close to 0 m in steep locations and on the hilltops. This leads to bimodal distributions with an high, narrow peak around zero and the second broad peak (or more) at high values. This makes soil thickness hard to estimate using traditional statistical regression techniques or spatial interpolation methods.

For this reason, instead of just considering the **absolute depth to bedrock**, which is another name of soil thickness, other two variables are considered: a **censored depth to bedrock** up to a threshold value, typically 2 meters *(standard soil description depth)*, and the **probability of occurrence of the R horizon** within that threshold (cf. Figure 1.8).

Chapter 2

Explorative analysis of SoilGrids data

This chapter contains a presentation of SoilGrids and its limitations. Following, an overview of SoilGrids predictions on the study area for six variables of interest (clay/silt/sand percentages, absolute and censored depth to bedrock and probability of occurrence of R horizon).

2.1 Digital soil mapping

Having reliable quantitative geographical information on the physical and chemical properties of the terrain is fundamental in agriculture and in civil and environmental engineering. Before the computer era, the only possible way to determine the soil properties over an area consisted in (i) dividing it in separate zones sufficiently homogeneous, (ii) collecting samples and making measurements for each of these sectors, and (iii) inferring the properties on each zone by taking averages or by the means of more sophisticated statistical methods like spatial interpolation.

With the advent of the computer, Geographic Information Systems (GIS) where created to process digital geographic information. These systems made it easier to store, process and share geographical information in the form of **raster,** a data format that allows to store maps of values associated to geographical coordinates. Databases of soil information collected in different corners of the world started circulating on the web and became available to public and private institutions. At the same time, satellite technologies combined to GIS allowed to collect a great amount of additional remote-sensed information relative to the Earth surface.

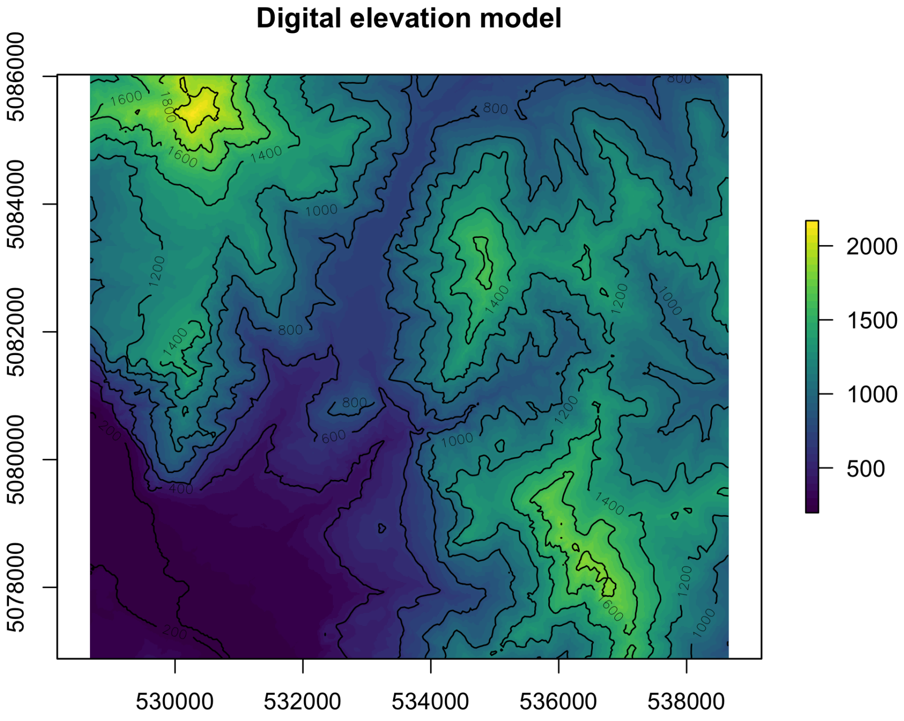


Figure 2.1: Digital Elevation Model of the Caldone basin (raster data format)

For example, **Digital Elevation Models (DEMs)** are now available for the whole Earth surface up to a resolution of few meters, and consequently it is possible to obtain a complete characterisation of the topography of an area. The DEM of the Lecco area is shown in Figure 2.1.

All these factors paved the way to the birth of **Digital soil mapping**, the process of generating rasters containing information on the properties of the terrain by applying statistics and machine learning to the data collected by direct measurements and remote sensing.

The first review of digital soil maps techniques was compiled in 2003 by McBratney [46]; since then, the ever increasing volume of public data issued from on field surveys and remote imaging, combined with the development of statistical methods and machine learning has led to the creation of increasingly accurate and detailed maps.

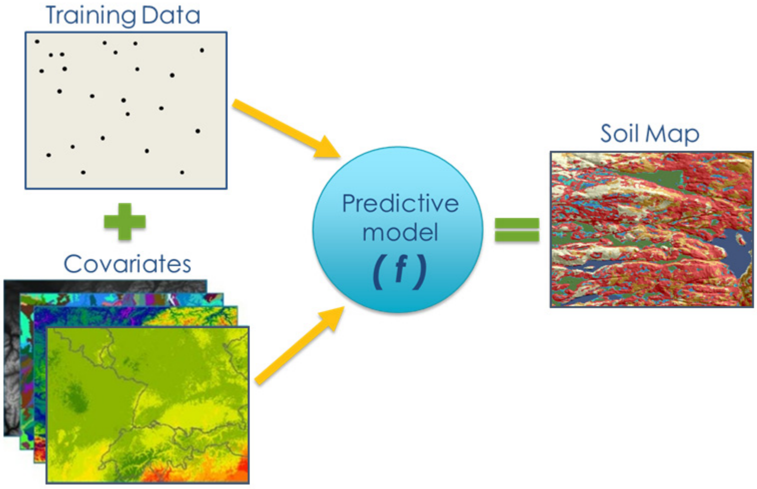


Figure 2.2: Digital soil mapping
This image is taken from the site of the United States Department of Agriculture (USDA) and shows a simple conceptual map of the process of digital soil mapping

2.2 SoilGrids: global gridded soil information

In 2014 *ISRIC (International Soil Reference and Information Centre) - World Soil Information,* a non-profit organisation funded by the Dutch government released SoilGrids, a system for automated digital soil mapping based on state-of-the-art spatial predictions methods. SoilGrids predictions are based on globally fitted models using soil profile and environmental covariate data [35]. When first released, SoilGrids.org served a collection of updatable soil properties and class maps of the world at 1 km spatial resolutions produced using automated soil mapping based on statistical regression models. In 2017, the resolution has been increased to 250 m and the accuracy of the predictions has been greatly improved by using machine learning algorithms instead of the previously employed linear regression [34]. SoilGrids.org aims at becoming the analogue of OpenStreetMap and/or OpenWeatherMap for soil data. SoilGrids data are available publicly under the Open DataBase License.

The numbers and figures reported in this section are taken from [35] and [34], where a detailed presentation of SoilGrids can be found.

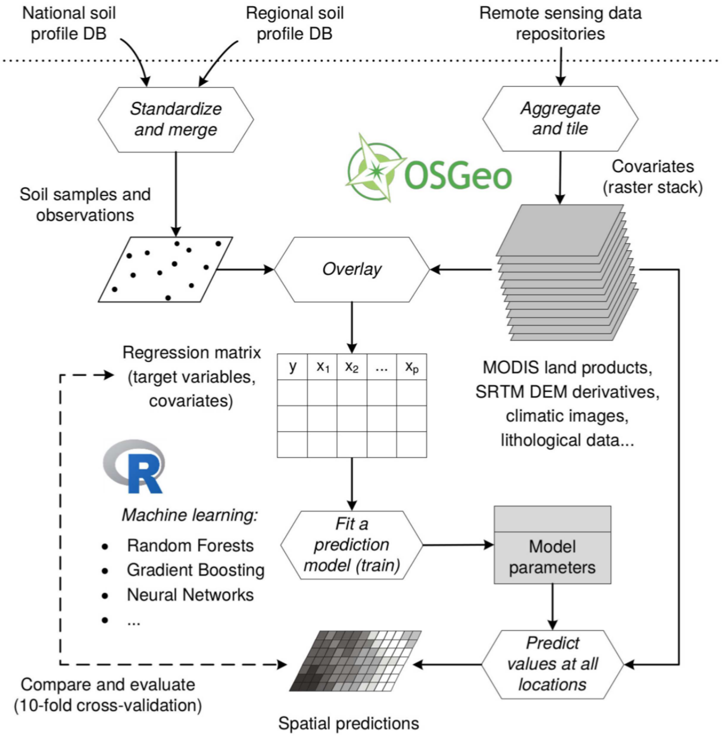

National soil profile DB
Regional soil profile DB
Remote sensing data repositories
Standardize and merge
Aggregate and tile
OSGeo
Covariates (raster stack)
Soil samples and observations
Overlay
Regression matrix (target variables, covariates)
y
x_1
x_2
...
x_p
MODIS land products, SRTM DEM derivatives, climatic images, lithological data...
R
Machine learning:
Random Forests
Gradient Boosting
Neural Networks
...
Fit a prediction model (train)
Model parameters
Predict values at all locations
Compare and evaluate (10-fold cross-validation)
Spatial predictions

Figure 2.3: Statistical framework used for generating SoilGrids

SoilGrids predictions are based on *ca.* 150,000 soil profiles used for training and a stack of 158 remote sensing-based soil covariates (primarily derived from MODIS (Moderate Resolution Imaging Spectroradiometer) land products, SRTM (Shuttle Radar Topography Mission) DEM derivatives, climatic images and global landform and lithology maps), which were used to fit an ensemble of machine learning methods — random forest, gradient boosting and/or multinomial logistic regression — as implemented in the **R** packages *ranger, xgboost, nnet* and *caret*. The data-driven statistical framework used for generating SoilGrids maps is shown in 2.3.

Among the predicted variables there are clay, silt and sand percentages at different soil depths, absolute and censored depth to bedrock and probability of occurrence of R horizon. Table 2.1 shows the prediction accuracy for particle-size fractions and absolute depth to bedrock, based on 10–fold cross-validation. For particle-size fractions the amount of variation explained exceeds 70% and the RMSE doesn't exceed 13%. The good level of accuracy of SoilGrids predictions for clay/silt/sand content is testified by the plots in Figure 2.4.

Variable name	**N**	**MAE**	**RMSE**	**R-squared**
Sand content (%)	616,762	9.0	13.1	76.6%
Silt content (%)	613,750	6.7	9.8	79.4%
Clay content (%)	625,159	6.6	9.5	72.6%
Depth to bedrock (cm)	1,580,798	678	835	54.0%

Table 2.1: SoilGrids average prediction error for key soil properties based on 10–fold cross-validation. **N** = number of samples used for training, **MAE** = mean absolute error, **RMSE** = root mean square error. Source: [34]

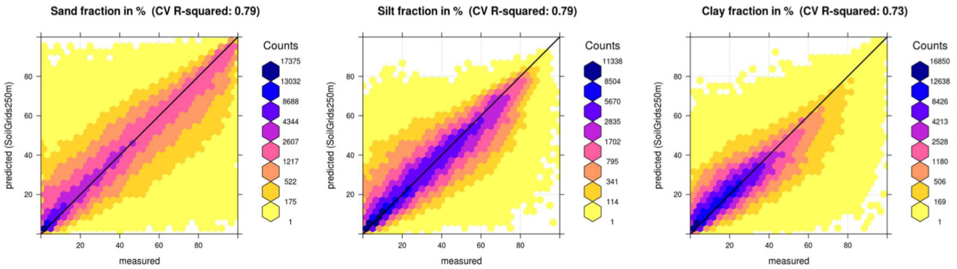


Figure 2.4: Correlation/density plots of particle-size fractions (10–fold cross-validation). Source: [34]

The problem of estimation of soil thickness is addressed in a dedicated article by some of the authors of SoilGrids [71].

As explained in the last section of chapter 1, estimation of absolute depth to bedrock (soil thickness) is a complex issue since the range of values goes from 0 m (**outcrops**) to thousands of meters. In many applications (in particular for the SMART-SED model) sometimes what matters is not a precise estimation of the depth to bedrock when this exceeds a few meters, but rather identifying the locations in which soil thickness is thin (< 2 m) and estimating the depth to bedrock in these points.

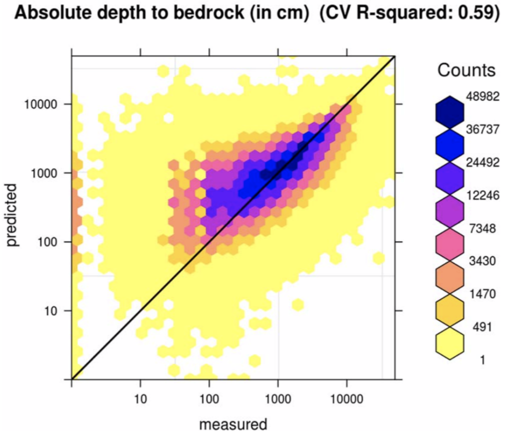


Figure 2.5: Correlation plot of absolute depth to bedrock result of 10–fold cross-validation. Source: [34]

Figure 2.5 shows that SoilGrids predictions of absolute depth to bedrock tend to significantly overestimate soil thickness for values $< 2\ m$, this tells us that considering only absolute depth is not advisable, especially when analysing mountain catchments where one would expect several outcrops and steep areas covered by thin soil. Fortunately, SoilGrids provides accurate estimations of the probability of occurrence of R horizon: the area under the ROC curve is 0.87 [71]. Censored observations (up to 2 m) of soil depth are also provided, for this variable the fitted model explains 35% of the total variability, with an RMSE of 50 cm [71].

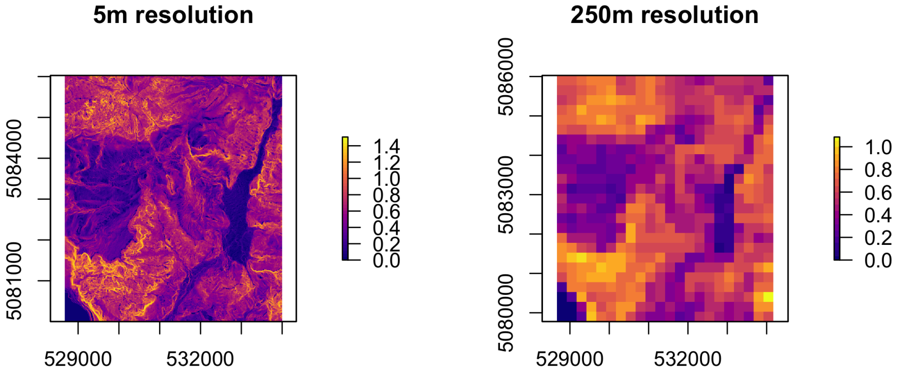


Figure 2.6: Fine resolution Vs coarse resolution
The two images show the slope (in radians) of a sector of the study area. The one on the left has a resolution of 5 m, the one on the right 250 m. This figure shows the inability of a low resolution raster to capture the local variability and the effect of reduction of the range of values

2.3 The issue of coarse resolution

SoilGrids data are predictions obtained from machine learning algorithms, therefore, they inevitably contain an error term which is hard to characterise and quantify in the absence of *hard data* from direct measurements. Apart from this, SoilGrids data present another issue which makes them not particularly suited - as they are - for analyses at local scales: their resolution is relatively low (about 250 m), and the variability within the pixels is unknown. When modelling a phenomena at a much finer scale (for instance SMART-SED should be able to model the basin up to a 5 m resolution), SoilGrids data could be interpreted as averages over square areas (comprised of several geographical units). The actual values at each location might greatly vary and go out of the range of the predictions (this is particularly true for soil thickness and will be discussed in the last section of this chapter). The problem of passing from a particular geographical resolution to another is called **change of support** [24], in particular, passing from a lower (or more coarse) resolution to a higher one is called **downscaling,** whereas **upscaling** is the inverse process. The most appropriate methods to deal with this issue will be presented in the following chapters.

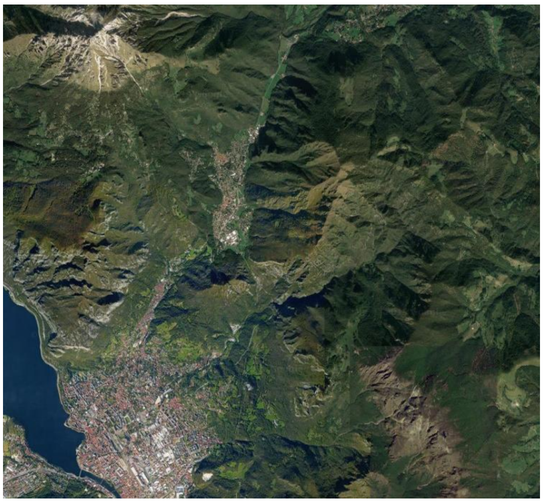


Figure 2.7: Aerial view of the region

2.4 Clay, silt and sand fraction predictions in the study area

In this section we explore the SoilGrids predictions for the percentages of clay, silt and sand in the region of interest for the SMART-SED case study. Instead of just considering the 24.2 km^2 of the Caldone basin (whose geometry is shown in Figure 1.3) we consider a broader square region containing the v-shaped catchment area, since convexity is preferred in a geo-statistical context.

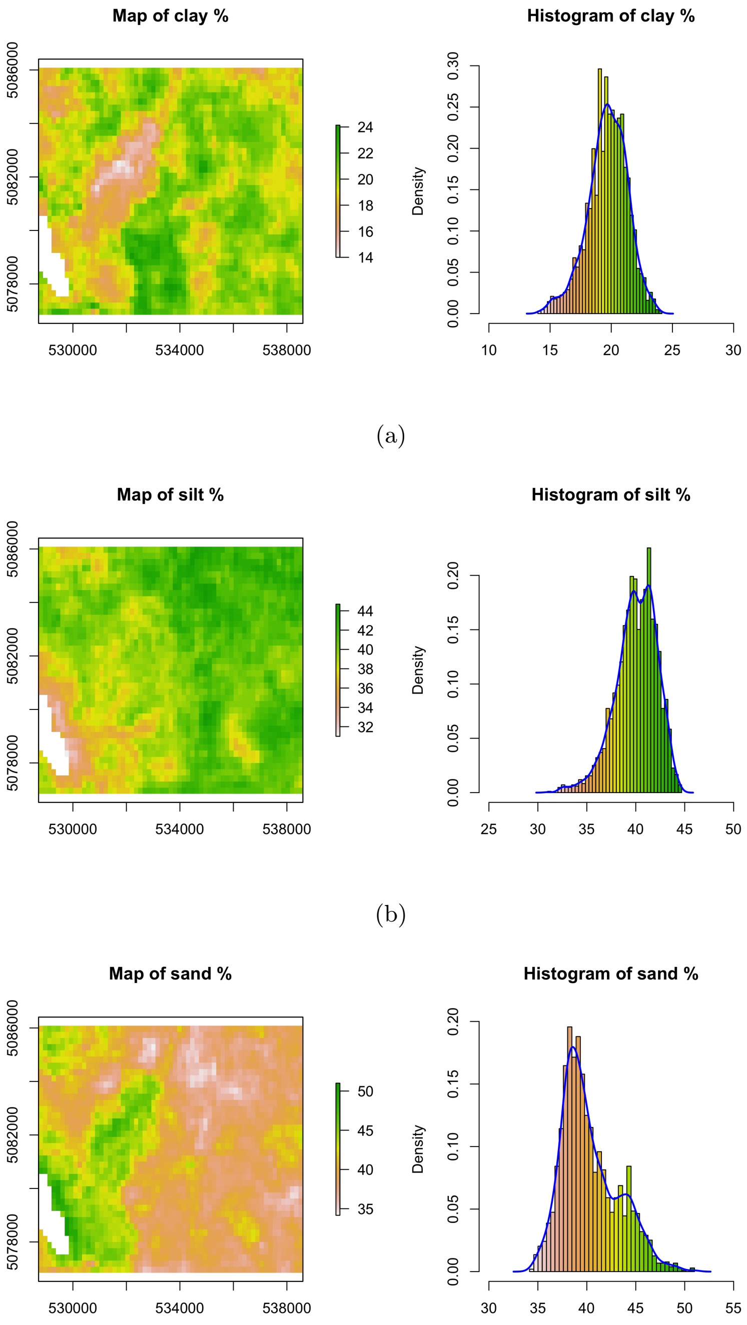
We report in Figure 2.7 the aerial view of the region for a better visual interpretation of the maps.

SoilGrids provides fractions of clay/silt/sand at seven standard soil depths: 0 cm, 5 cm, 15 cm, 30 cm, 60 cm, 100 cm and 200 cm.

In this thesis the values considered are those referred to the topsoil (depth of 0 cm). As suggested in [34], averages over standard depth intervals, e.g. 0-5 cm or 0-30 cm can easily be derived by taking a weighted average of the predictions within the depth interval using numerical integration, such as the trapezoidal rule [62]:

$$\frac{1}{b-a}\int_a^b f(x)dx \approx \frac{1}{2(b-a)}\sum_{k=1}^{N-1}(x_{k+1}-x_k)(f(x_k)+f(x_{k+1})),$$

where the x_k are the different depths.



(a)

(b)

(c)

Figure 2.8: SoilGrids prediction of the fraction of clay (a), silt (b) and sand (c) in the topsoil of the study area. There is no available data on the lake.

	min	**max**	**average**	**sd**
Clay fraction (%)	14%	24.14%	19.68%	1.64%
Silt fraction (%)	31%	44.69%	40.03%	2.11%
Sand fraction (%)	34.15%	51%	40.28%	2.88%

Table 2.2: Summary statistics for clay, silt and sand percentages

Maps in Figure 2.8 show, in order, the fraction of clay, silt and sand (in percentage) in the topmost layer of soil of the study area. Values are not provided on the lake (white areas). The overall fractions distributions are also shown.

The summary statistics for the three variables are contained in Table 2.2.

As we can see, there is not much variation in the overall composition of the soil in terms of relative fractions of the three particle-size ranges, the standard deviation is only 1.64 % for clay and does not exceed 3% even for the most variable of the three psf.

Clay fractions are almost normally distributed with a mean of about 20 %, silt fractions present a small tail for lower values but most observations are centered around 40 %, as are sand fractions, although for sand percentages the distribution is slightly bimodal with a second peak around 44-45 %.

Comparing the maps of clay/silt/sand % with the aerial view (Figure 2.7) and with the digital elevation model (Figure 2.1) there seems to be a correlation between elevation and soil composition, in particular in the clay and silt content seems lower in the valleys and on the mountaintops (consequently the sand content is higher). This aspect will be analysed and discussed in Chapter 4.

From a visual inspection of the maps there seems to be a negative correlation between clay and sand fractions, an intuition which is confirmed by the scatter plots in Figure 2.9: the values in the upper triangular part of the matrix are the Pearson correlation coefficients of the variables: it is apparent that a strong negative correlation exists between clay and sand content (-0.69), and between silt and sand content (-0.81).

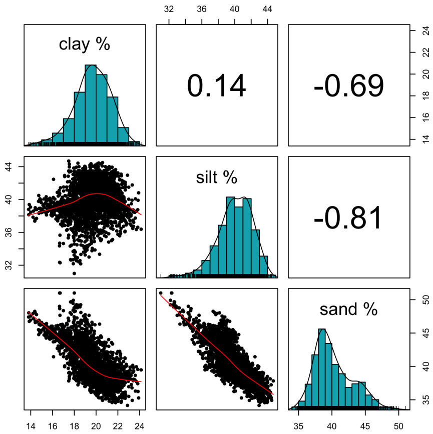


Figure 2.9: Scatter plot of particle-size fractions

This correlation is most likely *spurious*, i.e., it is probably due to the constraints on the data in exam. In fact, the variables we are considering represent percentages of different components of a unit (the soil), so they have to respect the following particular properties: let $z_1(x)$, $z_2(x)$ and $z_3(x)$ respectively represent the percentage of clay, silt and sand in the soil of a particular location x of the spatial domain D, then:

$$\sum_{i=1}^{3} z_i(x) = 100, \quad z_i(x) > 0 \ \ \forall x \in D. \tag{2.1}$$

Data that have to respect these constraints are called **compositional data**. The sample space of compositional data is called the **simplex,** in the case of psf it is 3-dimensional and is defined as:

$$\mathbb{S}^3 = \{(z_1, z_2, z_3) : z_1, z_2, z_3 > 0, \quad z_1 + z_2 + z_3 = 100\} \tag{2.2}$$

Since the sum of the three variables is fixed, when the values of z_1 and z_2 are known so is the value of z_3, hence the 3D-simplex is in fact a 2D subset of $\mathbb{R}^3$. The constrained nature of these variables induces a spurious correlation in the data, not determined by an actual causal relation but rather by their intrinsic nature. This mathematical relationship is called **spurious correlation of ratios** [38].

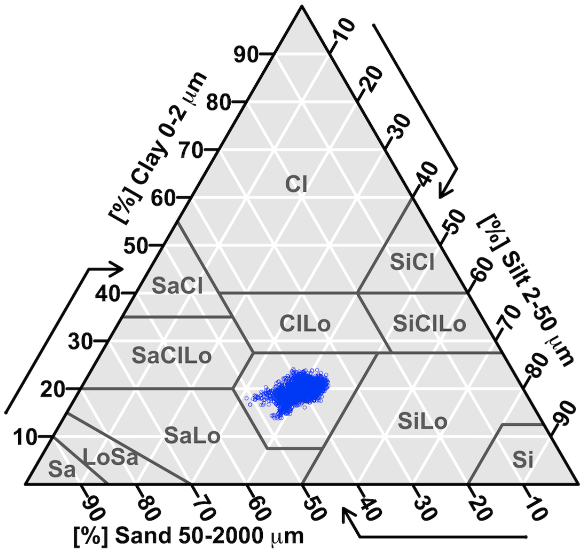


Figure 2.10: Cloud of data in the soil texture triangle (USDA classification)

Finally, we determine soil texture based on clay, silt and sand content. In Figure 2.10 data points are plotted on the soil texture triangle. We observe that the soil of almost the whole study area (except for very few locations with lower sand content) falls into the *loam* category of the USDA classification system. Loam is soil composed mostly of sand and silt and a smaller amount of clay (40–40–20%). The term "loam" generally refers to soil types that are not predominantly sand, silt, or clay. Loam is part of the group **B** of Table 1.3, of fairly permeable soil with moderate infiltration rates and moderate runoff potential. Some locations are characterised by a clay/silt/sand content in between loam and *sandy loam*, a type of soil slightly more permeable which may fall in the group **A** of Table 1.3. On top of that, SoilGrids data have a limited resolution ($\sim 250m$) and are not entirely accurate, so that some areas might actually be characterised by different types of soils like *sandy clay loam* or *clay loam*, a proper characterisation of the incertitude should account for this possibility.

2.5 Absolute and censored depth to bedrock, probability of occurrence of R horizon

Soil thickness exerts a first-order control on the hydrologic response of upland watersheds. Relatively thin soils are more prone to saturated overland flows compared to thicker soils which have greater water storage potential. SoilGrids provides estimates for three variables related to soil thickness: absolute depth to bedrock in meters, which from now on will be indicated with the acronym **ADB**, censored depth up to 200 *cm* (**CDB**), and finally the probability that the soil depth is lower than 2 *m*, namely the probability of occurrence of the *R horizon* (other name for bedrock) within 200 *cm* from the surface (**PRH**).

ADB predictions are based on over 1,500,000 observations worldwide, with values ranging from 0 to 1,250 meters (cf. Table 2.1), the percentage of explained variance of SoilGrids predictions is 54%, considerably lower with respect to particle-size fractions (almost 80% for sand percentage). In particular, SoilGrids tend to overestimate soil thickness in some locations, for instance consider the map of ADB for the study area, reported in Fig. 2.11.

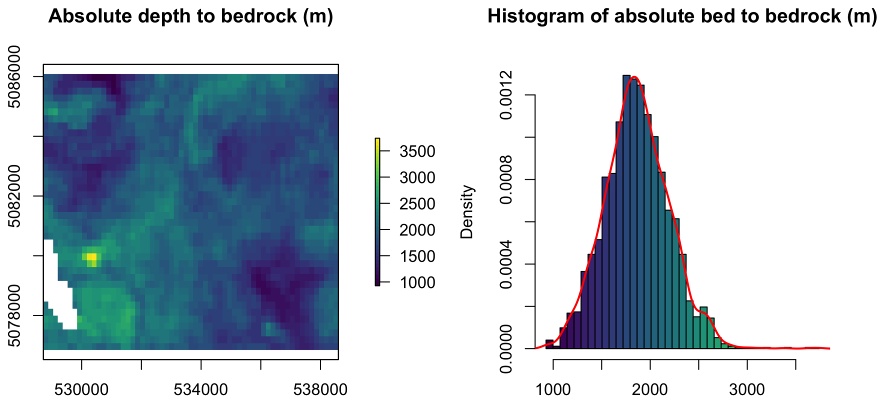
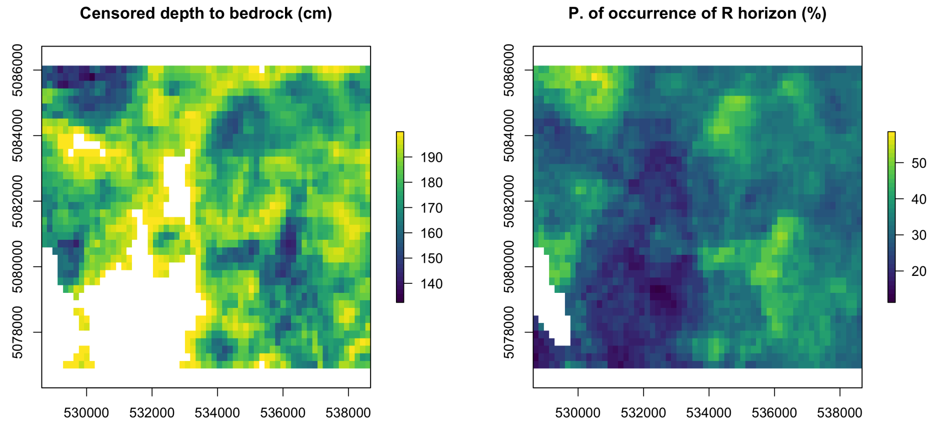


Figure 2.11: SoilGrids prediction of absolute depth to bedrock (in meters) in the study area

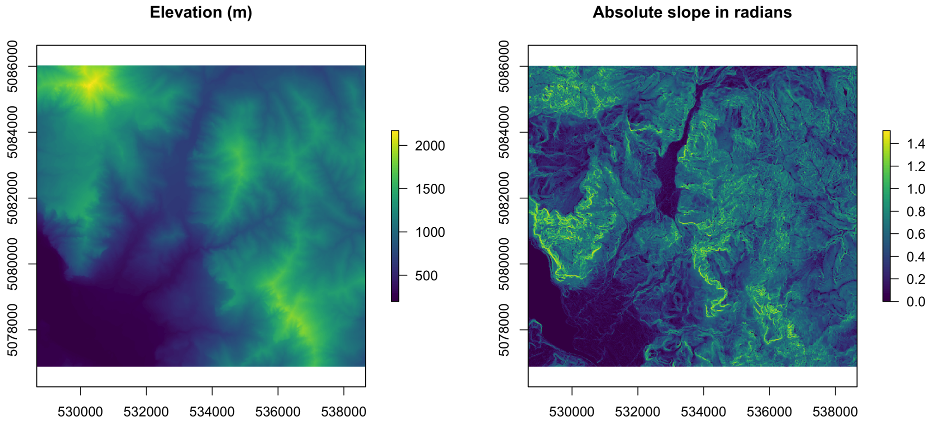
Values range from $\sim$ 1000 *m* to 3500 *m*. Even on hilltops and in very steep locations ADB is estimated to be over 1000 *m*, whereas a basic visual inspection of the aerial view shows that many areas present visible outcrops (i.e., ADB $=$ 0, cf. Figure 2.7). On the contrary, very high values of ADB in the valleys are justified and might even be underestimated.



(a) (b)

Figure 2.12: SoilGrids prediction of the fraction of CDB (a) and PRH (b)

As stated in Chapter 1, knowing the exact value of ADB in the valleys or in general in the areas where soil thickness is expected to be very high (hundreds of meters) is not of primal importance. Instead, a proper identification of area with outcrops and thin soil is more useful for a numerical simulation of sediment transport. Figure 2.12 shows the maps of CDB and PRH, the white areas in the CDB map are locations in which CDB hits the censoring threshold (soil depth $> 2\ m$). As one would expect, the two quantities are highly (negatively) correlated: locations with higher PRH have a lower predicted CDB. The values of CDB are still relatively high for the whole area and almost never go under 130 cm, on the same page PRH hardly ever exceeds 60%, despite the visible outcrops. PRH also never goes below 10%, even at locations with an estimated ADB of 2000+ meters, this inaccuracy might be due to the low resolution of the data and the consequent possible heterogeneity inside the "pixels" or "blocks" of $\sim$ 250 meters side. The objective of this thesis is to *downscale* from a low resolution to a higher one using fine-resolution secondary information and special techniques specifically designed for this task.



(a)

(b)

Figure 2.13: (a) 5 m-resolution digital elevation model (DEM) of the study area (b) Absolute slope in radians computed with the finite differences method

The relation between soil thickness and topographical variables is intuitively obvious and well documented ([61], [41]). Just from a simple visual inspection of the maps of ADB, CDB and PRH in comparison with the elevation and slope maps in Figure 2.13, it is apparent that mountain peaks and steeper areas are those with the lowest estimated ADB and CDB, and those with highest PRH. Fine-resolution maps of topographical variables will be used in Chapter 4 to perform statistical downscaling of CDB and PRH through the **dissever** method.

Chapter 3

Mathematical framework

This chapter contains the mathematical formalisation of the problem of downscaling and prediction/simulation of regionalised variables, and a presentation of the statistical techniques that will be used to perform these tasks.

3.1 Kriging

Clay/silt/sand percentages, ADB, CDB and PRH are all **regionalised variables,** meaning that they vary in space, so that, if we call D the geographical domain of the study area (in our case a square in the two dimensional Euclidean space) each of them could be represented by a *field* over the area $\{z(x),\ x \in D\}$, where x is a generic point or *location* of the domain. In practice, points are identified by minimal geographical units or "pixels", although in the abstract model the domain is still considered as a continuum. Since we don't know the exact values of the variables in exam at each point x, we place ourselves in a statistical context and we model the variables as **random fields**. A random field over a domain D is a stochastic process characterised by a collection of random variables $\{Z(x),\ x \in D\}$ over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

The branch of statistics that focuses on regionalised variables is called **geostatistics**. Generally speaking the main purpose of geostatistics is making inference on the distribution of a random fields using a finite number of observations in some fixed locations of the domain. **Kriging** is the most common and widely used technique in geostatistics. There, we present briefly the basic concepts of the Kriging model, more detailed information can be found in [36].

Suppose we have N observations $\{z(x_i),\ i = 1, ..., N\}$ in N different locations of the domain D. The goal is to use this data to predict the value of $z(x_0)$ on an unobserved location x_0. Since the random variables $Z(x_i)$ are correlated, the known values $z(x_i)$ (usually indicated with z_i) constitute an unique observation of a random vector, and in principle cannot be used to estimate any statistic. To overcome this limitation, certain assumptions are made:

1. **First order stationarity** (constant mean):
$$\mathbb{E}[Z(x)] = m \ \ \forall x \in D,$$
2. **Second order stationarity:** suppose $x_j - x_i = \mathbf{h}$, then
$$C(Z(x_i), Z(x_j)) = C(Z(x_i), Z(x_i + \mathbf{h})) = C(\mathbf{h}),$$
i.e., the correlation between two random variables solely depends on the spatial distance between them, and is independent of their location.
3. **Isotropy:** let $|\mathbf{h}| = h$,
$$C(\mathbf{h}) = C(h),$$
i.e., the spatial correlation only depends on the absolute distance and not on the angle.

Under these assumptions we have:

$$\begin{aligned}
&Var(Z(x)) = C(Z(x), Z(x)) = C(0)\\
&Var(Z(x) - Z(x + \mathbf{h})) = Var(Z(x)) + Var(Z(x + \mathbf{h})) + 2C(Z(x), Z(x + \mathbf{h})) =\\
&= 2C(0) - 2C(h).
\end{aligned}$$

Based on this property we define the **(semi)variogram**:

$$\gamma(h) = C(0) - C(h) = \frac{1}{2}Var(Z(x) - Z(x + \mathbf{h})).$$

The **empirical (semi)variogram** can be computed with the formula

$$\hat{\gamma}(h) = \frac{1}{2|N(h)|} \sum_{i,j:(x_i,x_j)\in N(h)} (z(x_i) - z(x_j))^2, \tag{3.1}$$

where for a certain *lag* Δh:

$$N(h) = \{(x_i, x_j) : h - \Delta h \le |x_i - x_j| \le h + \Delta h\}.$$

The common practice in geostatistics consists in choosing a parametric model for the variogram and fitting it to the empirical variogram.

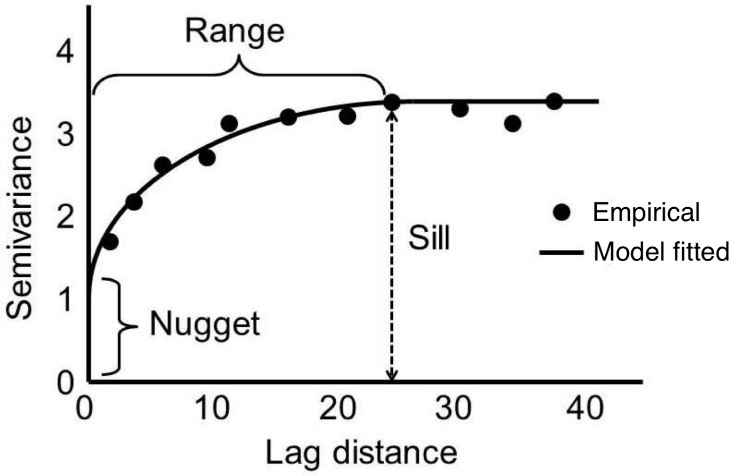


Figure 3.1: Variogram model fitted to some empirical observations

Variogram models are typically characterised by three parameters:

- the **nugget** τ^2: it is the value of the variogram at distance 0, if it is positive it accounts for uncorrelated white-noise in the data; variation at microscales smaller than the sampling distances will appear as part of the nugget effect;

- the **sill**, which is the limit of the variogram as $d \to \infty$; it is equal to the nugget τ^2 plus a positive quantity σ^2, called **partial sill**;

- the **range** a; it is the distance at which the difference of the variogram from the sill becomes negligible. In models with a fixed sill, it is the distance at which this is first reached; for models with an asymptotic sill, it is conventionally taken to be the distance when the *semivariance* first reaches 95% of the sill (i.e., the *practical range*).

Some commonly used variogram models are:

- *Exponential variogram*

$$\gamma^{exp}(h) = \sigma^2 \left(1 - exp(\frac{-\alpha \cdot h}{a})\right) + \tau^2,$$

- *Spherical variogram*

$$\gamma^{sph}(h) = \sigma^2 \left(\left(\frac{3h}{2a} - \frac{h^3}{2a^3}\right) \mathbb{1}_{[0,a)} + \mathbb{1}_{(a,\infty)}\right) + \tau^2,$$

- *Gaussian variogram*

$$\gamma^{gauss}(h) = \sigma^2 \left(1 - \exp(\frac{-\alpha \cdot h^2}{a^2})\right) + \tau^2.$$

In both the exponential and the gaussian variogram the parameter α depends on the definition of range.

Once the covariance structure has been identified, it is possible to perform spatial interpolation with Kriging. Assuming the mean m of the process is known, the **Simple Kriging (SK)** prediction of the value of Z for an unobserved location x_0 is:

$$z^*_{SK}(x_0) - m = \sum_{i=1}^{N} \lambda_i \cdot (z_i - m). \tag{3.2}$$

The parameters λ_i are determined by imposing two constraints:

1. *unbiasedness constraint:* $\mathbb{E}[Z^*_{SK}(x_0)] = \mathbb{E}[Z(x_0)]$;
2. *optimality criterion:* $\lambda = argmin\, \mathbb{E}[(Z^*_{SK}(x_0) - Z(x_0))^2]$;

so that the Kriging predictor becomes the **Best Linear Unbiased Predictor (BLUP)**.

Under all the previous assumptions, the vector of λ's can be computed by solving the *Simple Kriging system*

$$\mathbf{\Sigma} \cdot \lambda = \sigma_0, \tag{3.3}$$

where $\mathbf{\Sigma} = [C(Z(x_i), Z(x_j))]_{i,j=1,\ldots,N}$ and $\sigma_{\mathbf{0}} = [C(Z(x_0), Z(x_i))]^T_{i=1,\ldots,N}$. It is also possible to compute the variance of the prediction error:

$$\sigma^2_{SK} = Var(Z^*_{SK}(x_0) - Z(x_0)) = \lambda^T \cdot \mathbf{\Sigma} \cdot \lambda + C(0) - 2\lambda^T \cdot \sigma_0. \tag{3.4}$$

If m is unknown, we remove m from the equation, the resulting formula is known as **Ordinary Kriging (OK)** Prediction:

$$z^*_{OK}(x_0) = \sum_{i=1}^{N} \lambda_i \cdot z_i, \tag{3.5}$$

to guarantee unbiasedness, we add an additional constraint on the weights λ, namely

$$\sum_{i=1}^{N} \lambda_i = 1,$$

so that

$$\mathbb{E}[Z^*_{OK}(x_0)] = \mathbb{E}[\sum_{i=1}^{N} \lambda_i \cdot Z(x_i)] = \sum_{i=1}^{N} \lambda_i \cdot \mathbb{E}[Z(x_i)] = m \cdot \sum_{i=1}^{N} \lambda_i = m.$$

The vector of λs can then be computed by solving the *Ordinary Kriging system*

$$\begin{bmatrix} \mathbf{\Sigma} & \mathbf{1} \\ \mathbf{1^T} & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} \sigma_{\mathbf{0}} \\ 1 \end{bmatrix}, \tag{3.6}$$

where μ is a Lagrangian multiplier, $\mathbf{1}$ is an N-dimensional vector of 1's, and the terms $\mathbf{\Sigma}$ and $\sigma_{\mathbf{0}}$ are the same as before.

The variance of the prediction error can be computed as

$$\sigma^2_{OK}(x_0) = Var(\hat{Z}^{OK}(x_0) - Z(x_0)) = \lambda^T \cdot \sigma_{\mathbf{0}} + \mu. \tag{3.7}$$

It is important to note that no additional assumption (other than first and second order stationarity and isotropy) needs to be made on the distribution of the random field $Z(x)$ in order for the Kriging estimator to be the BLUP. Kriging is sometimes called *Gaussian Process (GP) regression.* The reason for this is the following: assuming the random field in exam is distributed as a GP with mean m and covariance structure identified by variogram γ (i.e., $Z(x) \sim GP(m, \gamma)$), then the Kriging predictor for $z(x_0)$ is equal to the conditional expectation of $Z(x_0)$ given the observations, namely

$$z^*_{OK}(x_0) = \mathbb{E}[Z(x_0)|Z(x_1) = z_1, ..., Z(x_N) = z_N]. \tag{3.8}$$

Additionally, the variance computed with the formula in (3.7) is the conditional variance of $Z(x_0)$. For this reason, z^*_{OK} and σ^2_{OK} are sometimes referred to as *conditional Kriging mean and variance.*

3.1.1 Regression Kriging (RK)

There exist several extensions of the Kriging method that allow to relax some of the assumptions and to incorporate secondary information into the model. In particular, the assumption of first order stationarity (constant mean over the domain) is often excessively strict: **regression Kriging (RK)** allows to relax this assumption by modelling the regionalised variable in exam as the sum of a *trend component* $m(x)$ (associated directly to the coordinates or to other auxiliary regionalised variables) and the *residual component* $e(x)$, modelled as a stationary random field.

Translated into a formula, the assumption of RK is that

$$Z(x) = m(x) + e(x) = f(u_1(x), ..., u_L(x)) + e(x). \tag{3.9}$$

The residual random field $e(x)$ has 0 mean and unknown covariance structure C, f is an unknown function of the L auxiliary variables $u_1, ..., u_L$, which vary in space and are known at each $x \in D$.

Regression Kriging can be performed in two steps:

1. *Fit of a regression model* to estimate $\hat{f}(u_1, ..., u_L)$;

2. *Kriging on the residuals:* let $\mathbf{u}(x) = (u_1(x), ..., u_L(x))$, residuals are obtained by subtracting the (estimated) trend from the original data:

$$e(x_i) = z_i - \hat{f}(\mathbf{u}(x_i)), \quad i = 1, ..., N,$$

the variogram model is fitted to the obtained residuals $e(x_1), ..., e(x_N)$, simple Kriging is performed on them.

Although this general formulation of RK allows for any regression method to be used at step 1, linear regression (LR) is the default method.

When LR is used, the RK estimation becomes

$$z^*_{RK}(x_0) = \sum_{l=1}^{L} \hat{\beta}_l \cdot u_l(x_0) + \sum_{i=1}^{N} \lambda_i \cdot e(x_i). \tag{3.10}$$

Since the residuals are spatially correlated, β parameters should be estimated through a *Generalised Least Squares (GLS)* fitting procedure. This would require to know the correlation of the residuals in advance, which in turn requires the knowledge of the β's to be estimated. An iterative procedure could be used to overcome this problem. However, several authors have shown that this procedure does not improve alter significantly the results obtained from *Ordinary Least Squares (OLS)* fit [48], [33].

3.1.2 Co-Kriging (CoK)

Co-Kriging (CoK) can be considered the multivariate extension of Kriging. For the sake of simplicity, suppose we have two regionalised variables $Z_1(x)$ and $Z_2(x)$, and that we have N observations of Z_1 and Z_2 at the same number of different locations in a study area. The CoK predictor of Z_1 at the unobserved location x_0 is given by:

$$z^*_{CoK}(x_0) = \sum_{i=1}^{N} \lambda_i \cdot z_1(x_i) + \sum_{i=1}^{N} \eta_i \cdot z_2(x_i). \tag{3.11}$$

The λ_i and η_i can be obtained solving a linear system equivalent to the one in (3.3), only this time the covariance matrix contains all the cross-covariances $C(Z_1(x_i), Z_2(x_j))$. To estimate the covariance, usually a *Linear Coregionalization Model (LCM)* is assumed (cf. [25]). Along with the variogram of each individual variable, a **cross-variogram** is also identified by fitting a model to the *empirical cross-variogram*

$$\hat{\gamma}_{12}(h) = \frac{1}{2|N(h)|} \sum_{i,j:(x_i,x_j)\in N(h)} (z_1(x_i) - z_1(x_j)) \cdot (z_2(x_i) - z_2(x_j)). \tag{3.12}$$

More information is provided in Section 7.3 of the appendix.

3.2 Area-to-Point Kriging (ATPK)

Kriging is used when the available observations are located in some precise points of the domain $x_1, ..., x_N$, but as we have seen in chapter 2, in our case data have a lower resolution with respect to the one required and can be therefore considered as **areal data** or **block data** that express an average value over a region.
The problem we are dealing with is the **downscaling** of coarse raster data. Although it might seem different from the problem of spatial interpolation of sparse point observations, the mathematical abstraction of Kriging is perfectly suited for dealing with it. In fact, one of the classical downscaling methods is **Area-to-Point Kriging (ATPK)**, and it is conceptually not different from Kriging. Some basic concepts of ATPK will be presented in this section, all the theoretical aspects are exhaustively presented in [39].

In ATPK coarse resolution data (from now on also called block data) to be downscaled are assumed to be (weighted) averages over the coarse grid cells or *blocks* to which they are assigned. Let $Z(x), x \in D$, be a random field over a geographical domain D, Z representing the regionalised variable of interest with constant mean m and covariance $C(Z(x_i), Z(x_j)) = C(|x_i - x_j|)$ identified by a variogram $\gamma(h)$.

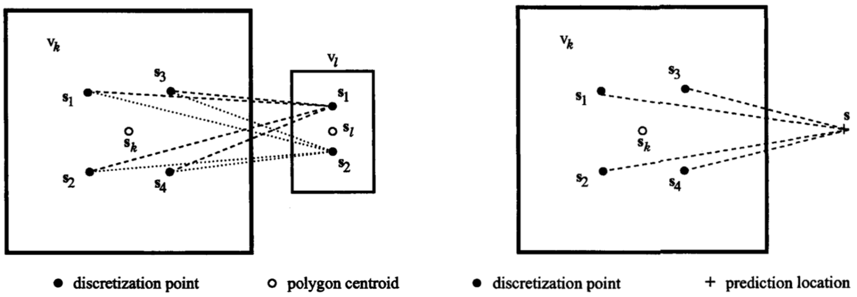
Let ν_k denote the support of areal region k. The $k-th$ observed areal datum $\overline{z}_k$ is assumed to be a realisation of random variable $\overline{Z}_k$, defined as the average of $Z(x)$ over the support ν_k :

$$\overline{Z}_k = \overline{Z}(\nu_k) = \frac{1}{|\nu_k|} \int_{x \in \nu_k} Z(x)dx \approx \frac{1}{P_k} \sum_{i=1}^{P_k} Z(x_i),\ x_i \in \nu_k, \tag{3.13}$$

where, in the discrete case, P_k denotes the number of points within support ν_k. The supports $\nu_1, ..., \nu_K$ are also called *blocks* and in our case are square cells.

Suppose K areal data $\overline{z}_k,\ k = 1, ..., K$, are available; just as Kriging, ATPK consists in predicting $z(x_0)$ as a linear combination of the observed areal data:

$$z^*_{ATPK}(x_0) = \sum_{k=1}^{K} \lambda_k \cdot \overline{z}_k. \tag{3.14}$$



(a) (b)

Figure 3.2: Approximate block-block (a) and block-point (b) covariance computation

Under the usual assumptions, by imposing the unbiasedness and optimality constraints it is easy to derive the system

$$\begin{bmatrix} \mathbf{\Sigma} & \mathbf{1} \\ \mathbf{1^T} & 0 \end{bmatrix} \begin{bmatrix} \lambda \\ \mu \end{bmatrix} = \begin{bmatrix} \overline{\sigma}_{\mathbf{0}} \\ 1 \end{bmatrix}, \tag{3.15}$$

μ is the Lagrangian multiplier used to impose the unbiasedness constraint. This time, $\overline{\mathbf{\Sigma}} = [\overline{C}(\nu_k, \nu_l)]_{k,l=1,\dots,K}$ and $\overline{\sigma}_{\mathbf{0}} = [\overline{C}(x_0, \nu_k)]^T_{k=1,\dots,K}$

The term $\overline{C}(\nu_k, \nu_l)$ denotes the *regularised* covariance between $\overline{Z}(\nu_k)$ and $\overline{Z}(\nu_l)$, also called the block-block covariance, and is computed as:

$$\begin{aligned} \overline{C}(\nu_k, \nu_l) = C(\overline{Z}(\nu_k), \overline{Z}(\nu_l)) = \frac{1}{|\nu_k|}\frac{1}{|\nu_l|}\int_{x\in\nu_k}\int_{x'\in\nu_l} C(Z(x), Z(x'))dxdx' \approx \\ \approx \frac{1}{\tilde{P}_k}\frac{1}{\tilde{P}_l}\sum_{i=1}^{\tilde{P}_k}\sum_{j=1}^{\tilde{P}_l} C(|x_j - x_i|),\ x_i \in \nu_k, x_j \in \nu_l. \end{aligned} \tag{3.16}$$

Analogously, $\overline{C}(x_0, \nu_k)$ is the block-point or point-block covariance, given by the formula:

$$\begin{aligned} \overline{C}(x_0, \nu_k) = C(Z(x_0), \overline{Z}(\nu_k)) = \frac{1}{|\nu_k|}\int_{x\in\nu_k} C(Z(x_0), Z(x))dx \approx \\ \approx \frac{1}{\tilde{P}_k}\sum_{i=1}^{\tilde{P}_k} C(|x_0 - x_i|),\ x_i \in \nu_k. \end{aligned} \tag{3.17}$$

3.2.1 The problem of variogram estimation in ATPK

The implementation of ATPK is fairly straightforward, although it requires the knowledge of point support covariance of the regionalised variable, identified by the variogram $\gamma(h)$. In the presence of punctual observations, the variogram can be estimated by fitting a model to the empirical values, as shown in Equation (3.1). However, when the only available information is given by the areal data, this simple procedure cannot be applied. It is still possible to compute an empirical variogram from low resolution gridded data by simply considering these values as localised in the centroids of their respective areas. The empirical variogram so computed, denoted $\hat{\gamma}_\nu(h)$, is an approximation of the *regularised variogram.* Assuming all blocks have the same size and shape (which is the case in most remote-sensing scenarios, and for SoilGrids data) and following [36], the point support and regularised variograms are related by the general formula:

$$2\gamma_\nu(h) = 2\overline{\gamma}(\nu(x), \nu(x+h)) - \overline{\gamma}(\nu(x), \nu(x)) - \overline{\gamma}(\nu(x+h), \nu(x+h))$$

which, under the assumption of stationarity, becomes:

$$\gamma_\nu(h) = \overline{\gamma}(\nu, \nu_h) - \overline{\gamma}(\nu, \nu). \tag{3.18}$$

The block-to-block variogram $\overline{\gamma}(\nu, \nu_h)$ represents the average value of the point support variogram between an arbitrary point in the support ν and another in the translated support ν_h. The second term, $\overline{\gamma}(\nu, \nu)$ is the within-block variogram value, independent from the block in the case of regular grids.

The problem of inferring the point-support variogram from its regularised version is called **variogram deconvolution**. It is an ill-posed inverse problem, therefore the only way to tackle it is through an empirical procedure paired with the Occam's razor. In [36], Journel and Huijbregts proposed a general approach without explicitly presenting an implementable algorithm. A similar approach was proposed by [10] in the context of remote sensing, yet they didn't clearly discuss the parameter estimation.

In 2008 Goovaerts developed a standard procedure for variogram deconvolution, applicable in general contexts [23], which is hereby presented:

1. Compute the experimental variogram $\hat{\gamma}_\nu(h)$ from low-resolution (areal) data and fit a model $\gamma_\nu^{exp}(h)$, where "*exp*" stands for "experimental", using weighted least-square regression [11] (each lag is weighted to assign more importance to the fitting of variogram values at short distances). The model that yields the smallest deviation between the experimental and modelled curves is selected.

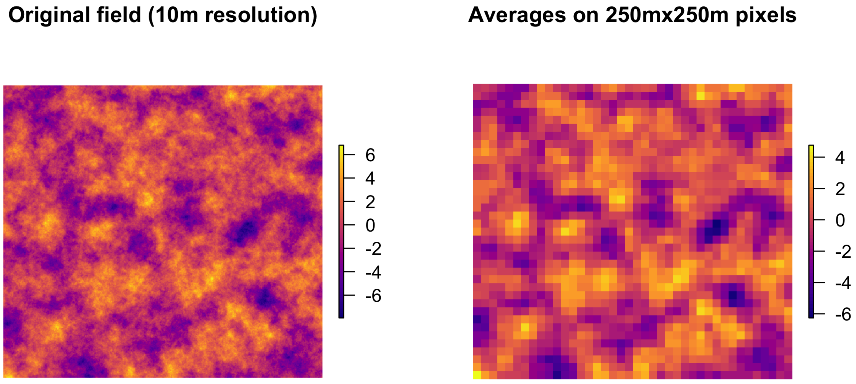


Figure 3.3: Example of the regularisation effect when lowering resolution by taking average values

2. As an initial point support model $\gamma^{(0)}(h)$, use the model fitted to areal data, $\gamma_\nu^{exp}(h)$.

3. Regularise $\gamma^{(0)}(h)$ according to the expression (3.18) to obtain $\gamma_\nu^{(0)}(h)$.

4. Quantify the deviation between the experimental and the regularised variograms using the average relative difference over L lags h_l:

$$D^{(0)} = \frac{1}{L}\sum_{l=1}^{L}\frac{|\gamma_\nu^{(0)}(h_l) - \gamma_\nu^{exp}(h_l)|}{\gamma_\nu^{exp}(h_l)}. \tag{3.19}$$

5. Consider the initial point support model, its regularisation and the associated difference statistic as "optimal" at this stage

$$\gamma^{opt}(h) = \gamma^{(0)}(h), \gamma_\nu^{opt}(h) = \gamma_\nu^{(0)}(h), D^{opt} = D^{(0)}.$$

6. For each lag h_l, compute experimental values for the new point support variogram through a re-scaling of the optimal point support model $\gamma^{opt}(h)$

$$\hat{\gamma}^{(1)}(h_l) = \gamma^{opt}(h_l) \times w^{(1)}(h_l),\ with\ w^{(1)}(h_l) = 1 + \frac{(\gamma_\nu^{exp}(h_l) - \gamma_\nu^{opt}(h_l)}{s_{exp}^2\sqrt{iter}}, \tag{3.20}$$

where s_{exp}^2 is the *sill* of the model $\gamma_\nu^{exp}(h)$, *iter* is the number of the iteration (at this step 1).

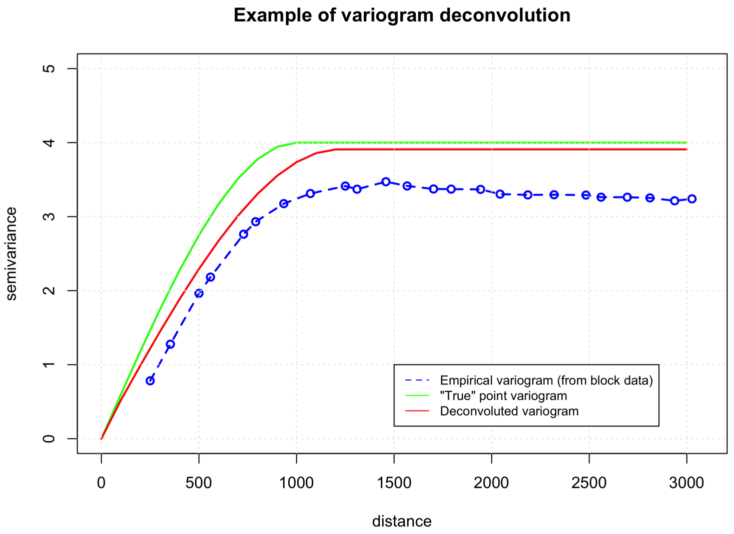


Figure 3.4: Variogram deconvolution of the field in Figure 3.3 using the Goovaerts' method implemented in R

7. Fit a model $\gamma^{(1)}(h)$ to the re-scaled values using weighted least-square regression (same procedure as in step 1).

8. Regularise the model according to (3.18) and obtain $\gamma_{\nu}^{(1)}(h)$.

9. Compute the difference statistic (3.19) for the new regularised model $\gamma_{\nu}^{(1)}(h)$

   - If $D^{(1)} < D^{opt}$, use the point support model $\gamma^{(1)}(h)$ and the associated statistic $D^{(1)}$ as new optimum, repeat stage 6 through 8.
   - If $D^{(1)} \geq D^{opt}$ repeat steps 6 through 8 using the same optimal model but the new re-scaling coefficients computed as

$$w^{(2)}(h_l) = 1 + \frac{(w^{(1)}(h_l) - 1)}{2}$$

10. Stop the iterative procedure after the i-th iteration whenever one of the following three criteria is met: (1) the difference statistic reaches a sufficiently small value; or (2) the maximum number of allowed iterations has been reached; or (3) a small decrease in the difference statistic D was recorded a given number of times;

3.2.2 Area-to-Point Regression Kriging (ATPRK)

Simple ATPK is basically a smoothing technique. In the absence of additional information at a finer scale it is hardly possible to do better than this. However, when auxiliary variables correlated with the one to be downscaled are exhaustively sampled at high resolution, it is possible to use this information to increase the precision of the downscaling procedure.

One of the possible ways to do this is through **Area-to-Point Regression Kriging (ATPRK).** Unsurprisingly, the formulation of ATPRK is almost identical to RK:

$$z^*_{ATPRK}(x_0) = \sum_{l=1}^{L} \hat{\beta}_l \cdot u_l(x_0) + \sum_{i=1}^{K} \lambda_i \cdot \overline{e}(\nu_k). \tag{3.21}$$

This time, as in ATPK, instead of having N residuals $e(x_1), ..., e(x_N)$, each one assigned to a point-location x_i, we have K *areal residuals* associated to each block. The difference is in how we fit the linear regression model for the trend component and how we compute the areal residuals.

The basic assumption is that we dispose of K block data $\overline{z}_1, ..., \overline{z}_K$ and of high resolution maps of the auxiliary variables $u_1(x), ..., u_L(x)$. We also assume that

$$\mathbb{E}[Z(x)] = \sum_{l=1}^{L} \beta_l \cdot u_l(x) \;\; \forall x \in D. \tag{3.22}$$

We don't know any value $z(x_i)$ so we cannot fit the model in the classical way, but from Equation (3.13) we derive

$$\mathbb{E}[\overline{Z}(\nu)] = \mathbb{E}\left[\frac{1}{|\nu|}\int_{x\in\nu} Z(x)dx\right] = \frac{1}{|\nu|}\int_{x\in\nu} \mathbb{E}[Z(x)]dx, \tag{3.23}$$

since $\mathbb{E}$ and $\int$ symbols can be exchanged as long as the expected value is finite. Combining Equation (3.23) with the formula in (3.22), we get

$$\begin{aligned}\mathbb{E}[\overline{Z}(\nu)] &= \frac{1}{|\nu|}\int_{x\in\nu}\left(\sum_{l=1}^{L} \beta_l \cdot u_l(x)\right)dx = \\ &= \sum_{l=1}^{L} \beta_l \cdot \left(\frac{1}{|\nu|}\int_{x\in\nu} u_l(x)dx\right) = \\ &= \sum_{l=1}^{L} \beta_l \cdot \overline{u}_l(\nu).\end{aligned} \tag{3.24}$$

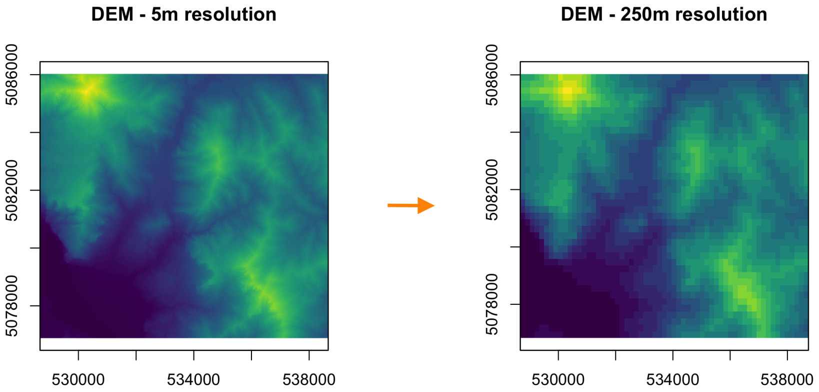


Figure 3.5: Upscaling of the digital elevation model

As a consequence, by computing $\overline{u}_l(\nu_k), \quad l = 1, ..., L, \quad k = 1, ..., K$, we are able to fit a linear model and obtain the $\hat{\beta}_l$ using the areal data $\overline{z}_1, ..., \overline{z}_K$ and subsequently find the residuals

$$\overline{e}(\nu_k) = \overline{z}_k - \sum_{l=1}^{L} \beta_l \cdot \overline{u}_l(\nu_k). \tag{3.25}$$

The procedure then follows the same steps as in Section 3.1.1.
The block averages of the auxiliary variables are computed as follows:

$$\overline{u}_l(\nu_k) = \frac{1}{P_k} \sum_{i=1}^{P_k} u_l(x_i), \tag{3.26}$$

where P_k is the total number of high resolution pixels (minimal geographical units) x_i contained in the block ν_k. This procedure is called **upscaling,** an example is shown in Figure 3.5.

Equation (3.24) holds true only because the relation between Z and the auxiliary variables u_l is assumed to be linear, otherwise any model fitted using the $\overline{z_k}$ and $u_l(\nu_k)$ will not be coherent with the regression Kriging assumption. It is however still possible to fit a model representing a nonlinear relation between Z and the u_l using linear regression; this can be done by increasing the number of features with nonlinear transformations of the original auxiliary variables, like it is done in *polynomial regression.*

3.3 Dissevering

ATPRK is an effective method for downscaling soil information when auxiliary variables are available. However, it has some limitations: block data $\overline{z}_k, \quad k = 1, ..., K$ are not directly compatible with the high resolution maps of auxiliary variables $u_l(x), \quad l = 1, ..., L$, so it is not directly possible to fit a generic regression model to identify a relation of the kind $\mathbb{E}[Z(x)] = f(u_1(x), ..., u_L(x))$. The auxiliary variables can be made compatible by performing upscaling and computing the average value over each block (Equation (3.26)); however, this only allows to fit a linear regression. In fact, Equation (3.24) is only valid because of the linearity of the model, but in general:

$$\frac{1}{|\nu|}\int_{\nu} f(\mathbf{u}(x))dx \neq f\left(\frac{1}{|\nu|}\int_{\nu}\mathbf{u}(x)dx\right). \tag{3.27}$$

When the relation $\mathbb{E}[Z(x)] = f(u_1(x), ..., u_L(x))$ is nonlinear, it is preferable to consider more sophisticated regression methods. A possible solution would be to increase the number of covariates by taking powers or transformations of the auxiliary variables $u_l(x)$, for instance $u_l^2(x)$, or $log(u_l(x))$ thus increasing the number of available features. However, the number of available models remains limited.

Another problem of ATPRK is variogram estimation of the residuals. In the absence of point observations, the variogram can only be obtained by deconvolution.

On top of that, when upscaling is performed on the covariates, some of the original information is lost. In particular, upscaling narrows the range of the variable averaged and eliminates all the local variations (cf. Figures 3.3 and 3.5). The reduction of the range can limit the effectiveness of the regression.

Lastly, ATPRK respects the so called **pycnophylactic** or **mass-preserving** property, meaning that if $z^*(x)$ is the downscaled map obtained applying ATPK to coarsely gridded data $\overline{z}_1, ..., \overline{z}_K$, then $\forall k$:

$$\sum_{i=1}^{P_k} \hat{z}(x_i) = \overline{z}_k. \tag{3.28}$$

The implicit assumption is that block data have no uncertainty associated to them, which is rarely the case, especially when the coarsely gridded data are the result of digital soil mapping algorithm, like SoilGrids.

In 2012, McBratney, Malone et al. came up with a general method to perform downsacling of coarse maps of soil information using available fine gridded covariate data that does not present the aforementioned issues of ATPK. The algorithm is called **dissevering** and is implemented in the R package **dissever** [44].

Let $\overline{z}_k, \quad k = 1, ..., K$, be the target variable values at each coarse resolution grid cell (or block) and $\hat{z}_i, \quad i = 1, ..., P$, denote the estimate of the target variable at each grid cell at fine scale (corresponding to the scale of the available covariates). In the spatial context, there are many i encapsulated by each k, the number of which is determined by the resolution of i and is not necessarily consistently equal for each block. The number of i encapsulated by block k is denoted P_k.

Dissever algorithm has two stages, namely *initialisation* and *iteration:*

- *Initialization.* The *iteration counter* t is set to 0, each $\hat{z}_i^0$ is set equal to the value of its encapsulating target variable $\overline{z}_k$. A nonlinear regression model between $\hat{z}_i^0$ and the suite of available covariates $u_1, ..., u_L$ is fitted to all the grid cells (or a subset, for computational feasibility)

  $$\hat{z}_i^0 = \hat{f}^0(u_1, ..., u_L).$$

  Potentially any regression method could be used;

- *Iteration.* At iteration t, in order to make the average of $\hat{z}_i^t$ estimates of fine resolution grid cells equal to the value of their encapsulating coarse resolution grid cell ($\overline{z}_k$), $\hat{z}_i^{t-1}$ are updated to $\hat{z}_i^t$ using equation

  $$\hat{z}_i^t = \hat{z}_i^{t-1} \times \frac{\overline{z}_k}{\hat{z}_k^{t-1}},$$

  where $\hat{z}_k^{t-1}$ is the average of $\hat{z}_k^{t-1}$ estimates over block $k : (1/\tilde{P}_k) \sum \hat{z}_i^{t-1}$ With the newly adjusted values a new nonlinear regression model $\hat{f}^t$ can be fitted. Iterations proceed until the total variation of the estimations between two consecutive time steps decreases below a given threshold ϵ :

  $$\frac{1}{\tilde{P}} \sum_{i=1}^{\tilde{P}} |\hat{z}_i^t - \hat{z}_i^{t-1}|, \quad \tilde{P} = number\ of\ used\ cells.$$

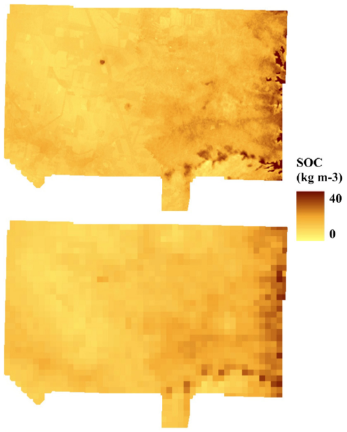


Figure 3.6: Example of application of the dissever algorithm for the downscaling of a *Soil Organic Carbon (SOC)* concentration map. Source: [44]

In the original article [44] **generalised additive models (GAM)** [32] were used in the dissevering procedure to fit a non-linear model for the downscaling of a coarse resolution map of *Soil Organic Carbon (SOC)* concentration (Figure 3.6). GAMs are not the only option, dissever allows to fit *any* non-linear (or linear) regression model for the estimation on the target variable at fine scale using high resolution covariates. Some of the regression methods which are generally used when applying the dissevering procedure are GAMs, *Random Forest (RF)* and **Cubist,** a rule-based model which is an extension of *Quinlan's M5 model tree* [63]. Cubist in particular has proven very effective for the spatial regression of soil properties, and has been used for downscaling in combination with ATPRK ([5]) and dissever ([68]). Cubist is implemented in the homonymous R package, more information on the algorithm is presented in Section 7.5 of the appendix.

Although Dissever has many advantages, it does not take into account the spatial correlation of the target variable, something that ATPRK does instead. In light of the characteristics of the two methods, it is possible to establish a rule of thumb for deciding which one of them to use:

- When the variability of the target variable is mostly explained by its spatial correlation, and its relation with high resolution covariate maps is secondary, use ATPRK.

- When the variability of the target variable is mostly explained by its (possibly non-linear) relation with covariates available at fine scale, and the spatial correlation is secondary/hard to estimate, then use dissever.

3.4 Sequential simulation

Kriging interpolation allows to quantify the uncertainty of the predictions. However, in order to analyse how the uncertainty propagates to the output of a numerical simulation we need to be able to generate stochastic maps that honours the probabilistic model of the variables in exam and the actual observations. This can be done through **conditional simulation**. It is hard to overstate the importance of conditional simulation in geostatistics, especially in reservoir modelling. In fact, several simulation algorithms have been developed, each of them having a specificity making it more effective for a particular problem. Among these algorithms, those based on **sequential simulation** stand out for their incredible flexibility and their ability to handle very large maps. Sequential simulation exploits the **Bayes theorem:** let $Z(x),\ x \in D$ be a random field with a generic covariance structure, let $Z_1, ..., Z_N$ represent the values of Z at N different points of the domain and let $\mathcal{L}(Z_1, ..., Z_N)$ indicate their joint probability distribution, then

$$\mathcal{L}(Z_1, ..., Z_N) = \mathcal{L}(Z_1)\mathcal{L}(Z_2|Z_1)\mathcal{L}(Z_3|Z_1, Z, 2)...\mathcal{L}(Z_N|Z_1, ..., Z_{N-1})$$

From an algorithmic point of view, this allows to sequentially simulate a random field on each point of a gridded map (raster), by conditioning at each step to the observed points and the previous simulations. As long as the range of the correlation function is limited, the conditional simulation at a point only takes into account its neighbourhood, so that the computation time for each individual simulation does not increase with the number of previously simulated points. This key aspect grants good scalability properties.

Sequential simulation relies on Kriging to reproduce the spatial correlation. Typically (in particular in the absence of additional information on the distribution of the random field) an assumption of **multi-Gaussianity** is made, i.e. the random field in exam is assumed to be a Gaussian Process (GP) with given mean and covariance. In this case, the algorithm is called **Sequential Gaussian Simulation (SGS)** and works as follows:

1. Define a random path visiting all the N *nodes* (cells of the grid).

2. For each node $x_i,\ i = 1, ...N$, compute the Kriging conditional mean and variance, given the original information and all the $i - 1$ previously simulated values $z(x_j),\ j = 1, ..., i - 1$.

3. Draw $z(x_i)$ from a normal distribution with mean and variance equal to those computed in step 2.

When the observations are not punctual but averages on blocks, it is still possible to perform conditional sequential simulation, the only difference regarding step 2., where, instead of Kriging, a combination of Kriging Area-to-point Kriging is performed. This allows to condition to both areal data and previously simulated values, so that the Kriging estimate becomes:

$$z^*(x_0) = \sum_{i=1}^{n(x_0)} \lambda_i \cdot z(x_i) + \sum_{j=n(x_0)+1}^{n(x_0)+K} \lambda_j \cdot z(\nu_j) \tag{3.29}$$

where $n(x_0)$ and K are the numbers of surrounding point and areal data, respectively, and $\nu_j,\ j = 1, ..., n(x_0) + K$ represent the areal supports. This is basically the same as ATPK, where point values are considered as *"degenerate"* areal values with infinitesimally small support. The point covariance is computed in the standard way, whereas the covariance between blocks and between points and blocks can be approximated using Equations (3.16) and (3.17) (cf. figure 3.2). The computation of block covariances is computationally intensive and particularly time-consuming. For this reason, even though conceptually the algorithm is the same, a specific implementation is required for fast computation of block-covariances [40], [42]. SGS conditioned to block data is called **Block Sequential Gaussian Simulation (BSGS).**

The multi-Gaussian assumption does not necessarily require the data to be normally distributed. This because samples from a realisation of a random field do not necessarily reflect the point probability distribution of $Z(x)$ ([36], [72]), this is shown in Figure 3.7.

However, there might be some cases in which the normality assumption would be unjustified, for instance when the histogram of the samples presents very heavy tails or when data present some constraints like in the case of particle-size fractions which are positive and with constant sum. The next section focuses on a method to deal with this kind of data.

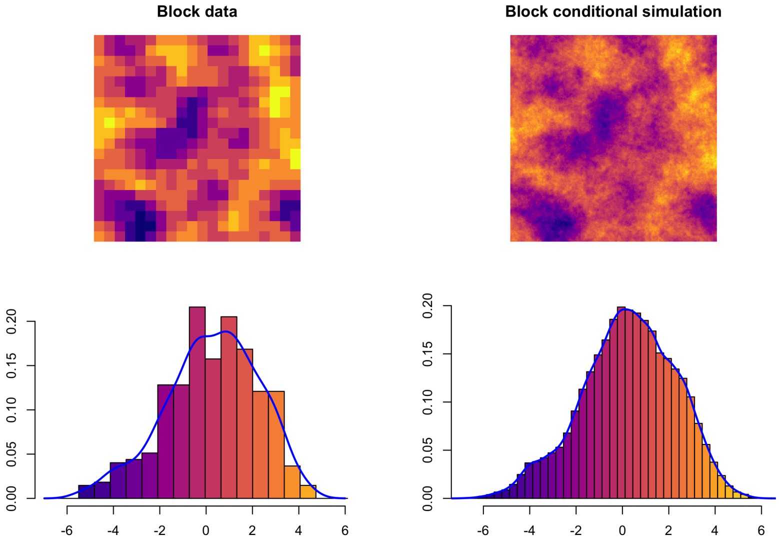


Figure 3.7: Example of block sequential gaussian simulation
The block data were originated from a Gaussian random field using unconditional sequential gaussian simulation on a coarse grid. Notice that the histogram of block data does not have the typical "bell" shape, this is because in conditional simulation the distribution from which each singular point is simulated is not necessarily reproduced at the field scale.

3.5 Compositional data

Particle-size fractions (psf) are **compositional data** which means they have to respect particular constraints (cf. Equation (2.1)).

Because of the range limitation and the spurious correlation of ratios, traditional Kriging and simulation techniques might not be suited for the spatial prediction of psf, although some authors have chosen to implicitly ignore this aspect (like [12]). The authors of [77] tried to account for the particular nature of regionalised variables expressing relative fractions by proposing an extension of Kriging called **Compositional Kriging (CK)**. CK predictions respect the constraints of positivity and fixed sum. However, CK algorithm is based on practical considerations rather than a coherent probabilistic model, and is therefore not suited for stochastic simulation.

The standard approach to the statistical analysis of compositional data is the one proposed by Aitchison in the 1980s [3], and is based on the particular geometry of the **simplex.** An n-dimesional simplex $\mathbb{S}^n$ is defined as:

$$\mathbb{S}^n = \{(z_1, ..., z_n)^T : z_1, ..., z_n > 0,\ z_1 + ... + z_n = c\}. \tag{3.30}$$

In the Aichison's view, the information contained in a set of compositional data is given by the ratios between components, so the information is preserved under multiplication by any positive constant. Therefore, the sample space of compositional data can always be assumed to be a *standard simplex*, i.e. $c = 1$. Normalisation to the standard simplex is called **closure** and is denoted by $\mathcal{C}(\cdot)$:

$$\mathcal{C}(\mathbf{z}) = \left(\frac{z_1}{\sum_{i=1}^n z_i}, ..., \frac{z_n}{\sum_{i=1}^n z_i} \right)^T \quad \forall \mathbf{z} = (z_1, ..., z_n)^T \in \mathbb{S}^n\,. \tag{3.31}$$

Points in a (real) simplex of dimension n have coordinates in $\mathbb{R}^n$ but given $n-1$ coordinates $z_1, ..., z_{n-1}$, then the last one z_n is equal to $c - \sum_{i=1}^n z_i$, so they actually can be uniquely determined by $n-1$ coordinates.

Aitchison (1986) [4] defined a group of operations that give the simplex the structure of a *real vector space,* namely:

- **Perturbation** (sum):

$$\mathbf{x} \oplus \mathbf{y} = \left(\frac{x_1 y_1}{\sum_{i=1}^n x_i y_i}, ..., \frac{x_n y_n}{\sum_{i=1}^n x_i y_i} \right)^T \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{S}^n \tag{3.32}$$

- **Powering** (product with a scalar):

$$\alpha \odot \mathbf{x} = \left(\frac{x_1^\alpha}{\sum_{i=1}^n x_i^\alpha}, ..., \frac{x_n^\alpha}{\sum_{i=1}^n x_i^\alpha} \right)^T \quad \forall \alpha \in \mathbb{R},\ \ \forall \mathbf{x} \in \mathbb{S}^n \tag{3.33}$$

- **Inner product:**

$$\langle \mathbf{x}, \mathbf{y} \rangle_a = \frac{1}{2n} \sum_{i=1}^{n} \sum_{j=1}^{n} log \frac{x_i}{x_j} \cdot log \frac{y_i}{y_j} \quad \forall \mathbf{x}, \mathbf{y} \in \mathbb{S}^n \tag{3.34}$$

The defined inner product induces a **norm** $|| \cdot ||_a := \sqrt{\langle \cdot \rangle_a}$, which in turn induces a **distance** $d_a(\mathbf{x}, \mathbf{y}) = ||\mathbf{x} \ominus \mathbf{y}||_a, \ \mathbf{x}, \mathbf{y} \in \mathbb{S}^n$, where $\mathbf{x} \ominus \mathbf{y}$ denotes the perturbation of $\mathbf{x}$ with the reciprocal of $\mathbf{y}$, i.e., $\mathbf{x} \ominus \mathbf{y} = \mathbf{x} \oplus (-1) \odot \mathbf{y}$. The vector space structure identified by these operations is called **Aitchison geometry,** or **Aitchison simplex** and it is a **finite Hilbert space** [15].

The statistical approach proposed by Aitchison consists in analysing compositional data in the context of the Aitchison geometry. Fortunately it is not required to directly use the previously defined operations. Instead, the standard procedure consist in transforming the original data by applying an **isomorphism** from the n-dimensional Aitchison simplex to the classical Euclidean space $\mathbb{R}^{n-1}$, make all the required assumptions and perform the statistical analysis on the transformed data and, and finally back-transform the results in the original space to ease interpretation.

There are three well characterised isomorphism from $\mathbb{S}^n$ to $\mathbb{R}^n$ or $\mathbb{R}^{n-1}$ that satisfy linearity:

- **Additive log-ratio transformation (alr):** $\mathbb{S}^n \rightarrow \mathbb{R}^{n-1}$

$$alr(\mathbf{z}) = \left(log \frac{z_1}{z_n}, ..., log \frac{z_{n-1}}{z_n} \right)^T \tag{3.35}$$

- **Centered log-ratio transformation (clr):** $\mathbb{S}^n \rightarrow \mathbb{U} \subset \mathbb{R}^n$

$$clr(\mathbf{z}) = \left(log \frac{z_1}{g(\mathbf{z})}, ..., log \frac{z_n}{g(\mathbf{z})} \right)^T, \quad g(\mathbf{z}) = \left(\prod_{j=1}^{n} z_j \right)^{\frac{1}{n}} \tag{3.36}$$

- **Isometric log-ratio transformation (ilr):** $\mathbb{S}^n \rightarrow \mathbb{R}^{n-1}$
  let $\langle ... \rangle_a$ denote the inner product in the Aitchison simplex and $\mathbf{e}_1, ..., \mathbf{e}_n$ be an orthogonal basis of $\mathbb{S}^n$, then

$$ilr(\mathbf{z}) = \left(\langle \mathbf{z}, \mathbf{e}_1 \rangle_a, ..., \langle \mathbf{z}, \mathbf{e}_{n-1} \rangle_a \right)^T \tag{3.37}$$

Differently from *alr*, *clr* and *ilr* transformations are **isometries,** so they preserve distances and angles. However, *clr* maps the simplex $\mathbb{S}^n$ into a linear subspace $\mathbb{U}$ of $\mathbb{R}^n$, whereas *ilr* maps the simplex directly into $\mathbb{R}^{n-1}$.

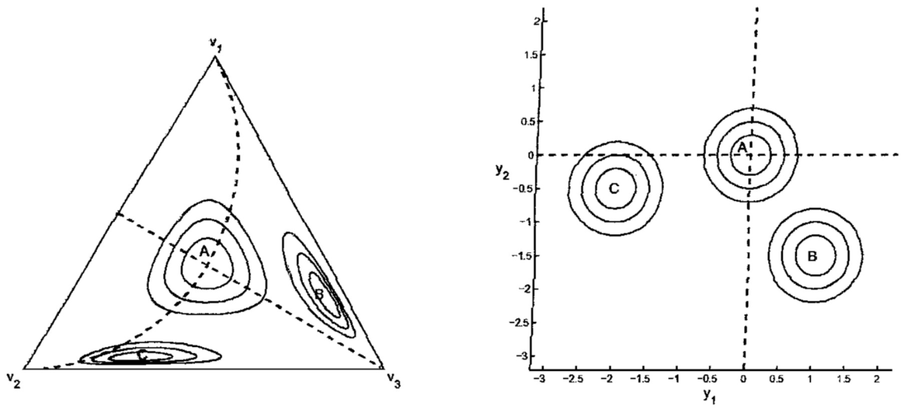


Figure 3.8: This figure shows how circles in the 3D Aichison simplex are transformed in circles of the Euclidean plane through isometric log-ratio transformation. Source: [17]

An algorithm for the *ilr* will be presented in the appendix, for an in-depth presentation of the isometric log-ratio transformations please refer to the articles of Egozcue, Pawlowsky-Glahn et al. [17], [59].

Figure 3.8 shows an example of *ilr* applied to a 3D simplex. The dashed lines represent the coordinates of the orthogonal base used for the transformation, the figure allows to visualise how circles look like in the Aitchison simplex.

In statistics, compositional data are often transformed in order to operate in the context of the Aichison geometry, the analysis is then performed on the new dataset and the results are back-transformed using the inverse transformations (alr^{-1}, clr^{-1} or ilr^{-1}).

In the context of spatial prediction, Kriging of compositional data transformed through one of the formulas (3.35), (3.36) or (7.4) is sometimes called **log-ratio Kriging** [79], and is denoted with different acronyms depending on the transformation used. The most common notation uses the name of the log-ratio transformation in capital letters with a subscript indicating the type of Kriging method, for instance for Ordinary Kriging (OK): ALR_{OK}, CLR_{OK} and ILR_{OK}. Often when comparing different methods, Kriging on the original compositions is denoted is a similar fashion: UT_{OK}, where "UT" stands for "Un-Transformed".

Several authors have addressed the problem of spatial prediction of particle-size fractions and have tried to compare different approaches by evaluating the prediction accuracy. Odeh et al. (2003) [50], compared the results of Compositional Kriging (CK), ALR_{OK}, UT_{OK} and co-Kriging (CoK) for the prediction of psf in the Australian soil. Wang (2017) considered also CLR_{OK} and ILR_{OK}. These studies show that none of the methods consistently outperforms the others; however, UT_{OK} almost always under-performs.

ilr has some important theoretical advantages: consider the three regionalised random variables Z_1, Z_2, Z_3 representing, respectively, the fraction of clay, silt and sand. Suppose we dispose on N observations for each fraction and we want to perform a statistical analysis in the Aitchison geometry. By applying *alr* we would obtain two meaningful transformed variables

$$alr_1 = log(Z_1/Z_3),$$

$$alr_2 = log(Z_2/Z_3).$$

The values obtained depend on which variable is chosen to be the last one, the one that would be discarded in the analysis. If, for instance, we choose to consider the clay fraction as Z_3, the result might change with respect to the one obtained with the initial assumption. The main problem of *alr*, however, is that it is not an isometric transformation from the simplex, with the Aitchison metric, onto the real *alr*-space, with the ordinary Euclidean metric. As a consequence, when one does not take into account the absence of isometry, the interpretation of transformed data is not intuitive at all, as results occasionally do not match properties expected in the simplex [45]. Although *clr* is more symmetric, there is still an arbitrary choice on the variable that will be discarded in the analysis (since the *clr*-transformed variables are linearly dependent and sum up to 0).

Isometric log-ratio transformations do not have this problem, since they reduce the dimensionality of the original data, so that the result won't change based on an arbitrary assumption. Another important advantage of *ilr* is that it allows to choose the reference orthogonal base $\mathbf{e}_1, ..., \mathbf{e}_{n-1}$ of the Aitchison simplex to be used in the transformation. The choice of the base does not influence the results of the analysis, but can lead to practical advantages, depending on the problem at hand. For instance, the basis could be chosen in such a way to grant independence (in terms of correlation) of the resulting transformed data, or in a way that allows to consider particular sub-compositions independently [17].

3.6 Downscaling of compositional data: Isometric Log-Ratio Area-to-Point Kriging

The problem of downscaling of compositional data is still unexplored; [65] used Area-to-Point Kriging to downscale particle-size fractions data transformed with additive-log ratio (alr), using the silt fraction as a reference for the ratios. The alr was used as a practical solution to account for the compositional nature of the data, but the modelling assumptions were not explicitly stated, and the results were interpreted only in terms of prediction accuracy with respect to a given test set. In this section, a general method for the statistical downscaling and simulation of compositional data based on the application of ATPK in the context of the Aitchison simplex is proposed. The method is called **Isometric Log-Ratio Area-to-Point Kriging** (**ILR$_{\mathbf{ATPK}}$**), and, as suggested by the name, it uses ilr to allow to operate within the Aitchison geometry without having to modify the classical statistical techniques based on the Euclidean metric. The emphasis is put on the model assumptions and on the interpretation of the results.

In order to formalise the problem of ATPK in the Aitchison simplex, we must first formalise the statistical concepts of mean, variance and covariance and the integral operator in the simplex embedded with the Airtchison geometry.

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $\mathbf{X} : \Omega \to \mathbb{S}^n$ be a random vector of compositions over $(\Omega, \mathcal{F}, \mathbb{P})$. Following the approach of Fréchet (1948) and [16], who generalised the expected value operator for generic metric spaces, we provide the following definitions.

- The **Aitchison variance around** $\xi \in \mathbb{S}^n$ is the expected value of the squared Aitchison distance between $\mathbf{X}$ and ξ :

$$Var_a(\mathbf{X}, \xi) := \mathbb{E}[d_a^2(\mathbf{X}, \xi)].$$

- The **(Aitchison) centre** of the distribution of $\mathbf{X}$ is the element $\xi \in \mathbb{S}^n$ which minimises $Var_a(\mathbf{X}, \xi)$, and is denoted $Cen(\mathbf{X})$.

- The Aitchison variance around the centre, simply called **Aitchison variance** is denoted

$$Var_a(\mathbf{X}) := Var_a(\mathbf{X}, Cen(\mathbf{X})) = \mathbb{E}[d_a^2(\mathbf{X}, Cen(\mathbf{X}))].$$

  The square root of the AItchison variance is the **Aitchison standard deviation** denoted std_a.

- The **Aitchison trace-covariance** of two random vectors of compositions $\mathbf{X}_1, \mathbf{X}_2$ over the same probability space is defined as

$$C_a(\mathbf{X}_1, \mathbf{X}_2) := \mathbb{E}[\langle \mathbf{X}_1 \ominus Cen(\mathbf{X}_1), \mathbf{X}_2 \ominus Cen(\mathbf{X}_2)\rangle_a].$$

As shown in [16], explicit formulas can be derived for the centre and the Aitchison variance. In particular, the abstract centre defined according to the Fréchet's approach correspond to the original definition of centre of a random vector of compositions $\mathbf{X}$ (Aitchison, 1997), which is equal to the *closed geometric mean:*

Proposition 1.

$$Cen(\mathbf{X}) = \mathcal{C}(\exp(\mathbb{E}[log(X_1)]), ..., \exp(\mathbb{E}[log(X_n)]))^T, \tag{3.38}$$

whereas the Aitchison variance can be expressed as

Proposition 2.

$$Var_a(\mathbf{X}) = \frac{1}{2n}\sum_{i,j=1}^{n} Var\left(log\frac{X_i}{X_j}\right) = \sum_{i=1}^{n} Var\left(log\frac{X_i}{g(\mathbf{X})}\right), \tag{3.39}$$

where g is the geometric mean $g(\mathbf{X}) = \left(\prod_{j=1}^{n} X_j\right)^{\frac{1}{n}}$.

These operators of centre, Aitchison variance and covariance have all the properties of the classical corresponding operators, but with respect to the Aitchison geometry operations. In particular, they can be estimated with the common statistics computed in the Aitchison simplex. So, for instance, the following holds [57]:

Proposition 3. *The best estimator for the centre of a random composition $\mathbb{E}[\mathbf{X}] = \theta$ given a sample of realisations $\mathbf{x}_1, ..., \mathbf{x}_N$ is the sample mean in the Aitchison simplex or the* ***sample centre****:*

$$\hat{\theta} = \frac{1}{N} \odot \bigoplus_{i=1}^{N} \mathbf{x}_i, \tag{3.40}$$

where the symbols $\oplus$ and $\odot$ represent the Aitchison geometry operations of perturbation (3.32)*, corresponding to the sum in the Euclidean geometry, and powering* (3.33)*, corresponding to the multiplication by a scalar.*

With a straightforward computation it is easy to show that the sample centre corresponds to the closure of the geometric mean of the sample:

Proposition 4.

$$\begin{aligned} \hat{\theta} = \frac{1}{N} \odot \bigoplus_{i=1}^{N} \mathbf{x}_i = \mathcal{C}(g_1, ..., g_n), \\ with \;\; g_j = \left(\prod_{i=1}^{N} x_{ij} \right)^{\frac{1}{N}}, \quad j = 1, ..., n. \end{aligned} \tag{3.41}$$

Basically, as the sum operation in the simplex (perturbation) corresponds to the closure of the product, the arithmetic average corresponds to the closed geometric mean.

Consistently, the integral of a function $\mathbf{x} : \mathcal{T} \to \mathbb{S}^n$, with values in the simplex, over a set $\tau \subset \mathcal{T}$ is defined as follows:

$$\int_{\tau}^{*} \mathbf{x}(t)dt = \mathcal{C}\left(\exp\left(\int_{\tau} log(x_1(t))dt \right), ..., \exp\left(\int_{\tau} log(x_n(t))dt \right) \right)^T. \tag{3.42}$$

In particular, the quantity $\frac{1}{|\tau|} \odot \int_{\tau}^{*} \mathbf{x}(t)dt$ is called the **centre of $\mathbf{x}(t)$ over τ**, and can be approximated by

$$\frac{1}{|\tau|} \odot \int_{\tau}^{*} \mathbf{x}(t)dt \approx \frac{1}{M} \odot \bigoplus_{i=1}^{M} \mathbf{x}(t_i), \quad \text{for } M \text{ big enough.} \tag{3.43}$$

Now it is possible to formalise ATPK in the Aitchison simplex. Suppose we have a dataset of regionalised compositional areal data

$$\left\{ \overline{\mathbf{z}}_k = (\overline{z}_{k1}, ..., \overline{z}_{kn})^T, \quad k = 1, ..., K \right\},$$

each datum $\overline{z}_{ki}$ representing the relative fraction of the i-th component (of n considered) for the k-th block, or the area of support ν_k according to the ATPK notation. The objective is to predict/simulate the values of the compositional random field $\mathbf{Z}(x) \in \mathbb{S}^n$ over a geographical domain D - typically covering the extent of the available areal data - at a much finer scale. This time, the areal data are assumed to be the **centres** of the unknown regionalised compositions over the blocks to which they are assigned. Hence, the $k-th$ observed areal composition $\overline{\mathbf{z}}_k$ is assumed to be a realisation of random vector of compositions $\overline{\mathbf{Z}}_k$, defined as

$$\overline{\mathbf{Z}}_k = \overline{\mathbf{Z}}(\nu_k) = \frac{1}{|\nu_k|} \odot \int^*_{x \in \nu_k} \mathbf{Z}(x)dx \approx \frac{1}{P_k} \odot \bigoplus_{i=1}^{P_k} \mathbf{Z}(x_i),\ x_i \in \nu_k. \tag{3.44}$$

P_k denoting the number of points (fine-resolution pixels) within block ν_k.

The assumptions on the multi-variate random field $\mathbf{Z}(x)$ are equivalent to those of standard Kriging (cf. Sections 3.1 and 7.3), but with respect to the Aitchison geometry:

1. *Fist order stationarity (constant centre):*

$$Cen(\mathbf{Z}) = \xi, \quad \text{for a fixed } \xi \in \mathbb{S}^n\,. \tag{3.45}$$

2. *Second order stationarity and isotropy:*

$$C_a(\mathbf{Z}, \mathbf{Z}(x + \mathbf{h})) = C_a(|\mathbf{h}|),\ \forall x, \mathbf{h}\ s.t.\ x, x + \mathbf{h} \in D. \tag{3.46}$$

The ATPK prediction of $\mathbf{z}(x_0)$ in the Aitchison simplex is given by:

$$\mathbf{z}^*_{ATPKa}(x_0) = \bigoplus_{k=1}^{K} \lambda_k \odot \overline{\mathbf{z}}_k. \tag{3.47}$$

In order for the predictor to be the BLUP, two constraints have to be imposed allowing for the determination of the λs:

- *unbiasedness:*

$$Cen(\mathbf{Z}^*_{ATPKa}(x_0)) = Cen(\mathbf{Z}(x_0)); \tag{3.48}$$

- *optimality:*

$$\lambda = argminVar_a \left(\mathbf{Z}^*_{ATPKa}(x_0) \ominus \mathbf{Z}(x_0)\right). \tag{3.49}$$

This formulation fully accounts for the special nature of the areal data in exam, allowing to avoid all the possible issues listed in Chapter 2 and in the previous section. Fortunately, it is not necessary to implement a specific algorithm to solve for λ in the context of the Aitchison simplex. In fact, applying an isometric transformation from the Aitchison simplex into a classical Euclidean space allows to obtain an equivalent formulation of the problem that can be solved using the standard ATPK presented in section 3.2, thanks to the following results (detailed in [16]).

Let $\mathbf{X} \in \mathbb{S}^n$ be a random vector of compositions, and $h : \mathbb{S}^n \to \mathbb{R}^m$ be a *linear isometry* from the Aitchison simplex to the m-dimensional real space, m being either equal to $n - 1$ or n. In the latter case, the image of h is a linear subspace of $\mathbb{R}^n$. Then

Proposition 5.

$$Cen(\mathbf{X}) = h^{-1}(\mathbb{E}[h(\mathbf{X})]). \tag{3.50}$$

Proposition 6. *If* $h(\mathbf{X}) = \mathbf{Y} \in \mathbb{R}^m$*, then*

$$Var_a(\mathbf{X}) = \sum_{i=1}^{m} Var(Y_i). \tag{3.51}$$

More specifically, consider an isometric log-ratio transformation *ilr*. The choice of isometric log-ratio over centered log-ratio is motivated by the theoretical advantages of the former, detailed in the previous section. We set $\mathbf{Y}(x) := ilr(\mathbf{Z}(x))\ \forall x \in D$, and we apply *ilr* to the original areal data, obtaining the dataset

$$\{\overline{\mathbf{y}}_k = (\overline{y}_{k1}, ..., \overline{y}_{k(n-1)})^T, \quad k = 1, ..., K\},$$

such that

$$(\overline{y}_{k1}, ..., \overline{y}_{k(n-1)})^T = ilr(\ (\overline{z}_{k1}, ..., \overline{z}_{kn})^T\ )\ \ \forall k.$$

Proposition 7. *All the standard ATPK model assumptions (cf. Section 3.2) made on* $\mathbf{Y}(x)$ *and on the* $\overline{\mathbf{y}}_k$ *imply the previous model assumptions in the Aitchison geometry, and the ATPK predictor for* $\mathbf{Y}(x_0)$ *is equal to the ilr-transformed predictor for* $\mathbf{Z}(x_0)$ *in the Aitchison simplex (cf. Eq.* (3.47)*):*

$$\mathbf{y}^*_{ATPK}(x_0) = ilr(\mathbf{z}^*_{ATPKa}(x_0)) \tag{3.52}$$

Proof. Areal data as centres of the blocks. The $\overline{\mathbf{y}}_k$ are assumed realisations of variables $\overline{\mathbf{Y}}_k = \frac{1}{|\nu_k|}\int_{x\in\nu_k}\mathbf{Y}(x)dx$. This implies that the $\overline{\mathbf{z}}_k$ are realisations of variables $ilr^{-1}(\overline{\mathbf{Y}}_k)$. Since *ilr* is a linear isometry, so is ilr^{-1}, therefore they can be interchanged with integral or summation operators (as long as the sums are finite) [15], hence

$$\begin{aligned} ilr^{-1}(\overline{\mathbf{Y}}_k) = ilr^{-1}\left(\frac{1}{|\nu_k|}\int_{x\in\nu_k}\mathbf{Y}(x)dx\right) = \frac{1}{|\nu_k|}\odot\int^*_{x\in\nu_k} ilr^{-1}(\mathbf{Y}(x))dx = \\ = \frac{1}{|\nu_k|}\odot\int^*_{x\in\nu_k}\mathbf{Z}(x)dx = \overline{\mathbf{Z}}_k. \end{aligned} \tag{3.53}$$

So, property (3.44) is respected.

First and second order stationarity, isotropy. The first order stationarity condition (constant mean) for $\mathbf{Y}(x)$ is equivalent to the constant centre condition for $\mathbf{Z}(x)$ thanks to Proposition 5:

$$\mathbb{E}[\mathbf{Y}(x)] = m \in \mathbb{R}^{n-1} \Rightarrow Cen(\mathbf{Z}(x)) = ilr^{-1}(m) = \xi \in \mathbb{S}^n, \quad m \text{ fixed.}$$

As for the Aitchison trace-covariance C_a, we recall that since *ilr* is a linear isometry, denoting $\langle \cdot, \cdot \rangle$ the inner product in $\mathbb{R}^{n-1}$ we have [16]:

$$\langle \mathbf{X}_1, \mathbf{X}_2 \rangle_a = \langle ilr(\mathbf{X}_1), ilr(\mathbf{X}_2) \rangle, \ \mathbf{X}_1, \mathbf{X}_2 \in \mathbb{S}^n,$$

then, applying Eq. (3.50) we get

$$\begin{aligned} C_a(\mathbf{Z}(x), \mathbf{Z}(x+\mathbf{h})) = \mathbb{E}[\langle \mathbf{Z}(x) \ominus \xi, \mathbf{Z}(x+\mathbf{h}) \ominus \xi \rangle_a] = \\ = \mathbb{E}[\langle \mathbf{Y}(x) - ilr(\xi), \mathbf{Y}(x+\mathbf{h}) - ilr(\xi) \rangle] = Tr(C(\mathbf{Y}(x), \mathbf{Y}(x+\mathbf{h}))), \end{aligned}$$

namely, the Aitchison trace-covariance is equal to the trace of the covariance matrix of the transformed vectors of compositions, hence its name. Second order stationarity and isotropy of the covariance of $\mathbf{Y}(x)$ implies second order stationarity and isotropy of C_a. Since in the Aitchison geometry formulation of ATPK no assumption is made on the covariance structure of the field of compositions, other than the two hypothesis previously stated, then the covariance structure can be estimated directly in the isometric log-ratios setting using the standard techniques presented in the chapter. In particular, if needed, a linear co-regionalization model could be assumed (cf. Section 7.3).

Equivalence of the predictor and of the problem constraints. The ATPK predictor for $\mathbf{Y}(x_0)$ is

$$\begin{aligned} \mathbf{y}^*_{ATPK}(x_0) = \sum_{k=1}^{K} \lambda_k \cdot \overline{\mathbf{y}}_k = \sum_{k=1}^{K} \lambda_k \cdot ilr(\overline{\mathbf{z}}_k) = \\ = ilr\left(\bigoplus_{k=1}^{K} \lambda_k \odot \overline{\mathbf{z}}_k \right), \end{aligned} \tag{3.54}$$

once again we used the fact that *ilr* is a linear isometric transformation to bring it out of the summation operator. Equation (3.54) entails that if the optimal lambdas are the same as those obtained imposing the constraints (3.48) and 3.49, then the thesis holds. In the context of the standard formulation, if the mean m is unknown, the unbiasedness condition can be imposed by setting

$\sum_{k=1}^{K}\lambda_k = 1$. In the Aitchison formulation, this entails

$$\begin{aligned} Cen\left(\bigoplus_{k=1}^{K}\lambda_k \odot \overline{\mathbf{z}}_k\right) &= \bigoplus_{k=1}^{K}\lambda_k \odot Cen(\overline{\mathbf{z}}_k) = \bigoplus_{k=1}^{K}\lambda_k \odot \xi = \\ &= \mathcal{C}\left(\xi_1^{\sum_{k=1}^{K}\lambda_k}, ..., \xi_n^{\sum_{k=1}^{K}\lambda_k}\right) = \mathcal{C}(\xi) = \xi, \end{aligned}$$

hence the unbiasedness constraint is the same. Finally, Proposition 6 and Equation (3.54) entail that

$$Var_a\left(\mathbf{Z}^*_{ATPKa}(x_0) \ominus \mathbf{Z}(x_0)\right) = Var\left(\mathbf{Y}^*_{ATPK}(x_0) - \mathbf{Y}(x_0)\right),$$

so the optimality constraints of the problem in the isometric log-ratio setting is equivalent to the one in the Aitchison geometry, hence the thesis. □

We have shown that by applying an isometric log-ratio transformation on a dataset of compositional areal data it is possible to perform downscaling through Area-to-Point Kriging in the Aitchison simplex setting without having to make a specific implementation. Working directly on the isometric log-ratios leads to consistent results to the initial assumptions in the Aitchison setting.

In particular, it is also possible to perform stochastic simulation through *block sequential simulation*. Also in this case, the assumptions on the distribution from which to generate the simulations can be made on the isometric log-ratios. In fact, because of the properties of *ilr*, any geometric assumption made on the isometric log-ratios in the standard Euclidean space directly translates to an equivalent geometric property in the simplex. For instance, normality of the *ilr*-transformed data entails normality of the compositions in the Aitchison simplex [15].

An important property of ILR_{ATPK} is that, considering the original compositions, since the assumptions are made with respect to the Aitchison geometry the **mass-preserving property** does not hold:

$$\overline{\mathbf{z}}_k \neq \frac{1}{P_k}\sum_{x_j \in \nu_k}\mathbf{z}^*(x_j). \tag{3.55}$$

Instead, since the predictions respect the assumption of Equation (3.44), the following **centre-preserving property** property holds:

$$\overline{\mathbf{z}}_k \approx \frac{1}{P_k} \odot \bigoplus_{x_j \in \nu_k}\mathbf{z}^*(x_j), \tag{3.56}$$

Applying Proposition 3 in Equation (3.56) we get:

$$\overline{\mathbf{z}}_k \approx \frac{1}{P_k} \odot \bigoplus_{x_j \in \nu_k} \mathbf{z}^*(x_j) = \mathcal{C}\left(\left(\prod_{j=1}^{P_k} z_1^*(x_j) \right)^{\frac{1}{P_k}}, \ldots, \left(\prod_{j=1}^{P_k} z_i^*(x_j) \right)^{\frac{1}{P_k}} \right), \quad (3.57)$$

namely, instead of expecting the areal data to be arithmetic averages of the fine-scale predictions, one should expect them to be the closure of the geometric mean of the predicted/simulated values, in accordance with the choice to operate in the context of the Aitchison simplex.

This fact could pose some problems only in the situation in which block-data are *known* to be the arithmetic average of several observations collected in the area, which are somehow no longer available. In this situation it would be required to apply a *bias correction* to the areal data (based on asymptotic estimates of the difference between the mean and the centre of a sample of compositions). Since it is common practice in statistics to treat compositional data within the Aitchison geometry setting, this situation rarely arises.

When the block-data are the results of digital soil mapping at coarse resolution (like the case of SoilGrids), *a priori* there is no reason to require mass-preservation over centre-preservation, but using log-ratio transformations involves all the theoretical advantages listed.

In Chapter 4, the techniques presented will be used for the downscaling and simulation of SoilGrids data.

Chapter 4

Results

This chapter is divided in two parts: in the first part, the results of the statistical downscaling of SoilGrids predictions of particle-size fractions and soil thickness are presented. In the second part, some approaches are proposed to integrate *hard data* (information obtained from local surveys) in the downscaling procedures.

4.1 Statistical downscaling of particle-size fractions

In this chapter particle-size fractions are downscaled from the SoilGrids raster resolution of ca. 250 m to a resolution of 5 m using *isometric log-ratio Area-to-Point Regression Kriging* (ILR_{ATPRK}) (cf. Section 3.6). The 5 m resolution digital elevation model (DEM) is used as auxiliary variable for the ATPRK. *ilr* was chosen over the other log-ratios transformations for the theoretical advantages exposed in the previous chapter (see Section 3.5). We define:

$$\begin{aligned} z_1(x) &= \text{clay \% at location } x \text{ of the domain } D; \\ z_2(x) &= \text{silt \% at location } x \text{ of the domain } D; \\ z_3(x) &= \text{sand \% at location } x \text{ of the domain } D. \end{aligned}$$

A *"point"* x of the map corresponds to a 5 m resolution pixel (the highest resolution considered, to which coarse data have to be downscaled); hence, the area in exam is represented by a finite number of points of a grid $x_i, i = 1, ..., N$ $(= 3654342)$. Maps of SoilGrids predictions over the area only have $K = 2,440$ pixels, with a resolution slightly lower than 250 m (cf. Fig. 2.8), which are considered to be square areal supports or blocks $\nu_k,\ k = 1, ..., K$. Although not all blocks have the same area (some blocks at the borders are smaller), the

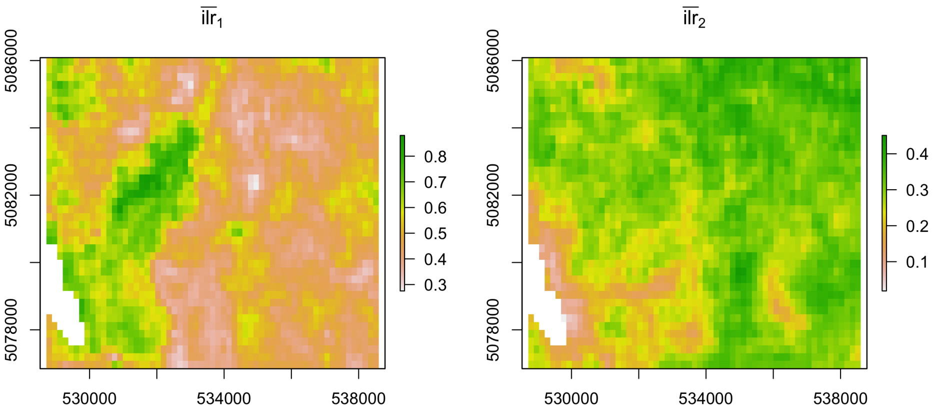


Figure 4.1: Maps of the isometric log-ratios obtained from the original clay, silt and sand fractions

majority of blocks contain ca. $P = 1500$ fine scale pixels. To simplify notation, all blocks will be considered identical and containing P points.

The basic assumption is that that SoilGrids predictions $\overline{z}_{k1}$, $\overline{z}_{k2}$ and $\overline{z}_{k3}$ represent the **centres** (in the Aitchison geometry sense) of the respective variables over the block (= coarse pixel) k to which they are associated, i.e.:

$$\overline{z}_{ki} \approx \frac{1}{P} \odot \bigoplus_{j=1}^{P} z_i(x_j), \quad i = 1, .., 3; \;\; k = 1, ..., K,$$

so that applying log-ratio ATPK will lead to consistent results.

SoilGrids predictions of particle-size fractions are then transformed using the R function *ilr* of the package **compositions** [7]. The base used for the transformation is the one introduced in the original article [17], based on the partition of the vector of compositional variables in two sub-compositions, the first consisting of z_1 and z_2 and the second containing only z_3 (cf. Section 7.4).

The result is two raster maps of regionalised variables

$$(\overline{ilr}_{k1}, \overline{ilr}_{k2}) = ilr(\;(\overline{z}_{k1}, \overline{z}_{k2}, \overline{z}_{k3})\;), \quad k = 1, ..., K,$$

shown in Fig. 4.1. This time, quantities $\overline{ilr}_{k1}$ and $\overline{ilr}_{k2}$ correspond to the *mean values* of the unknown regionalised variables $ilr_1(x)$ and $ilr_2(x)$ over the block k, i.e.:

$$\overline{ilr}_{ki} \approx \frac{1}{P} \sum_{j=1}^{P} ilr_i(x_j), \quad i = 1, .., 2; \;\; k = 1, ..., K.$$

There is no direct physical interpretation of the quantities ilr_1 and ilr_2, they just represent $z_1(x)$, $z_2(x)$ and $z_3(x)$ in the Aitchison simplex.

	min	**max**	**average**	**sd**
$\overline{ilr}_1$	0.275	0.880	0.507	**0.103**
$\overline{ilr}_2$	0.019	0.451	0.289	**0.065**

Table 4.1: Summary statistics for $\overline{ilr}_1$ and $\overline{ilr}_2$

Table 4.1 contains the summary statistics for $\overline{ilr}_1$ and $\overline{ilr}_2$. The average values are respectively, 0.507 and 0.289, roughly corresponding to to the following composition on terms of psf: 20% clay, 40% silt and 40% sand. This is the centre of the SoilGrids predictions for the area in exam, corresponding to a loam type of soil, consistent with Fig. 2.10. The variability (in bold font) is very low for both quantities, but especially for the map of $\overline{ilr}_2$.

In general, before performing ATPRK or block sequential simulation, one should check if auxiliary information is available at fine resolution that might be related to the variables to be downscaled. The use of ATPRK instead of simple ATPK usually improves the precision of the predictions.

In Chapter 2, a link between elevation and soil composition was hypothesised, this intuition is confirmed by the scatter plots in Figure 4.2. In order to study the correlation between the variables the digital elevation model was upscaled to match the resolution of $\overline{ilr}_1$ and $\overline{ilr}_2$ maps, according to the standard ATPRK procedure.

Both $\overline{ilr}_1$ and $\overline{ilr}_2$ seem correlated to the elevation, in particular $\overline{ilr}_2$ presents a concave parabolic behaviour, with lower values in correspondence of peaks and valleys and higher values on the slopes and in general at medium elevations. The behaviour of $\overline{ilr}_1$ is reversed, although the correlation with the DEM is lower.

The correlation between $\overline{ilr}_1$ and $\overline{ilr}_2$ depends on the basis used in the transformation. In this case, the isometric log-ratios obtained from the projection on the basis of the Aitchison simplex defined in [17] show no significative correlation and no particular trend. This does not necessarily imply that the isometric log-ratios are not *spatially cross-correlated.* However, in practical terms, when the variables at the same locations are uncorrelated one would expect no spatial cross-correlation either.

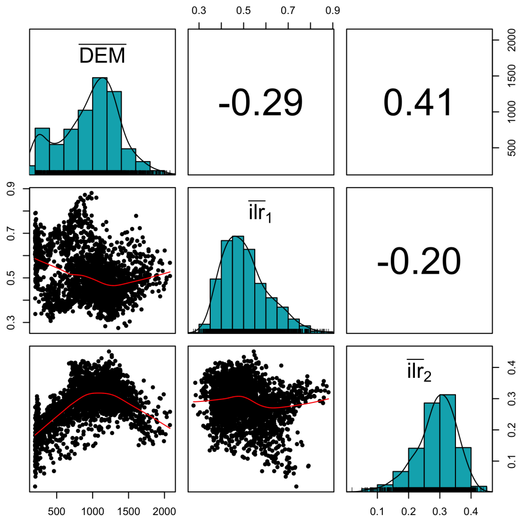


Figure 4.2: Scatter plots and histograms of the isometric log-ratios and the (upscaled) digital elevation model

On the basis of the previous considerations, the unknown maps $ilr_1(x)$ and $ilr_2(x)$ are assumed to be realisations of spatially correlated random fields $ILR_1(x)$ and $ILR_2(x)$, which are modelled as follows:

$$\begin{aligned} ILR_1(x) &= \beta_0^{(1)} + \beta_1^{(1)} \cdot DEM(x) + \beta_2^{(1)} \cdot DEM^2(x) + RES_1(x), \\ ILR_2(x) &= \beta_0^{(2)} + \beta_1^{(2)} \cdot DEM(x) + \beta_2^{(2)} \cdot DEM^2(x) + RES_2(x), \end{aligned} \tag{4.1}$$

where the $\beta_j^{(i)}$, $i, j = 1, 2$ are unknown regression parameters characterising the *trend* component related to the digital elevation, and the terms RES_1 and RES_2 are *residual* spatially correlated 2^{nd} order stationary random fields, characterised by unknown variograms γ_1 and γ_2.

The term $DEM^2(x)$ was also considered to account for the parabolic behaviour relating the ilr and the digital elevation observed in Fig. 4.2. A linear regression model was fitted using coarse maps $\overline{ilr}_i, i = 1, 2$, as target variables, and the upscaled digital elevation model ($\overline{DEM}$) and DEM squared ($\overline{DEM^2}$) as covariates. The fitted values are plotted against the observed values in Figure 4.3. The R^2 index is 0.10 for ILR_1 and 0.46 for ILR_2.

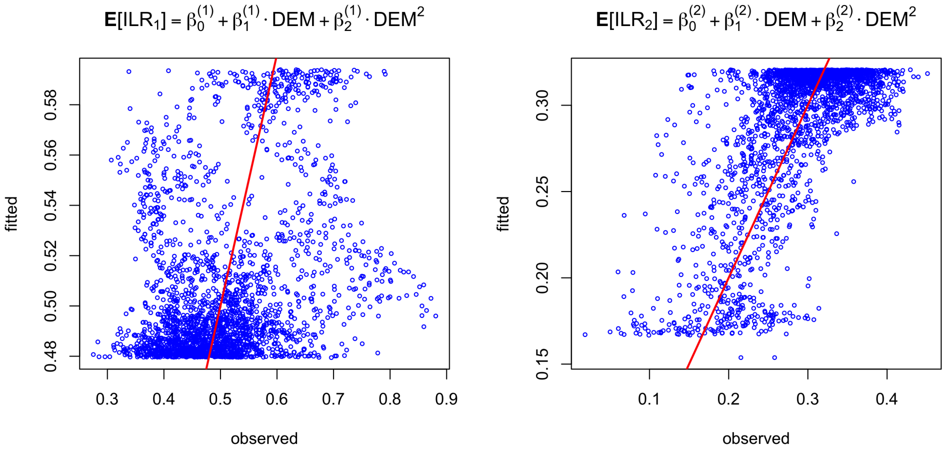


Figure 4.3: Scatter plots of the observed values Vs the fitted values of the regression models in Equation (4.1). In red, the line of equation $y = x$.

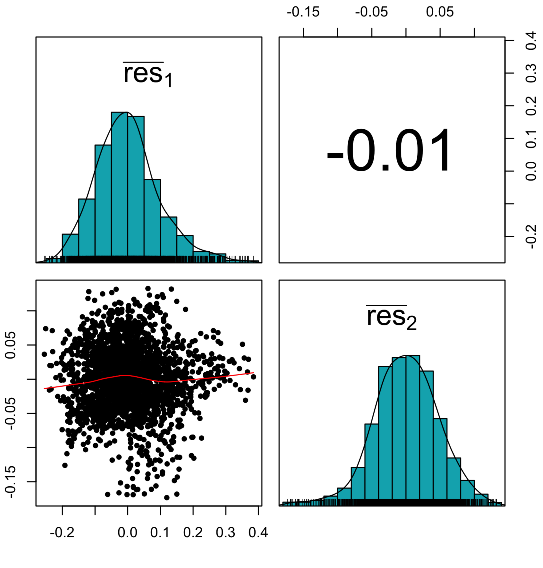


Figure 4.4: Scatter plot and histograms of the residual maps

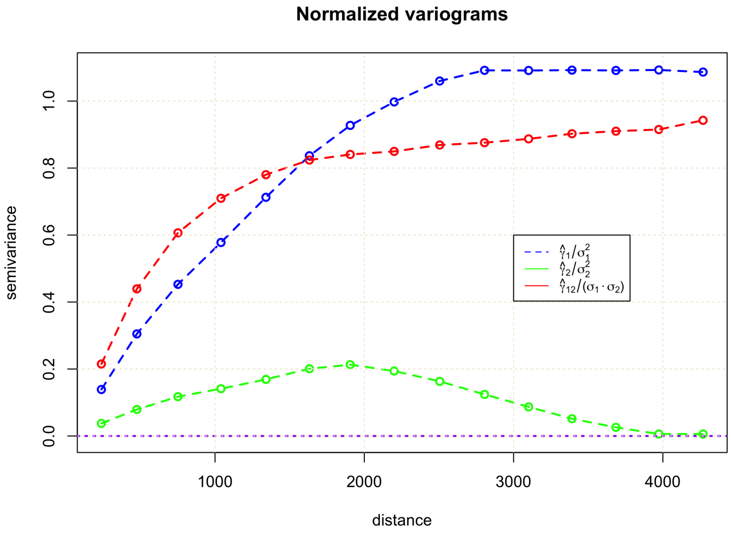
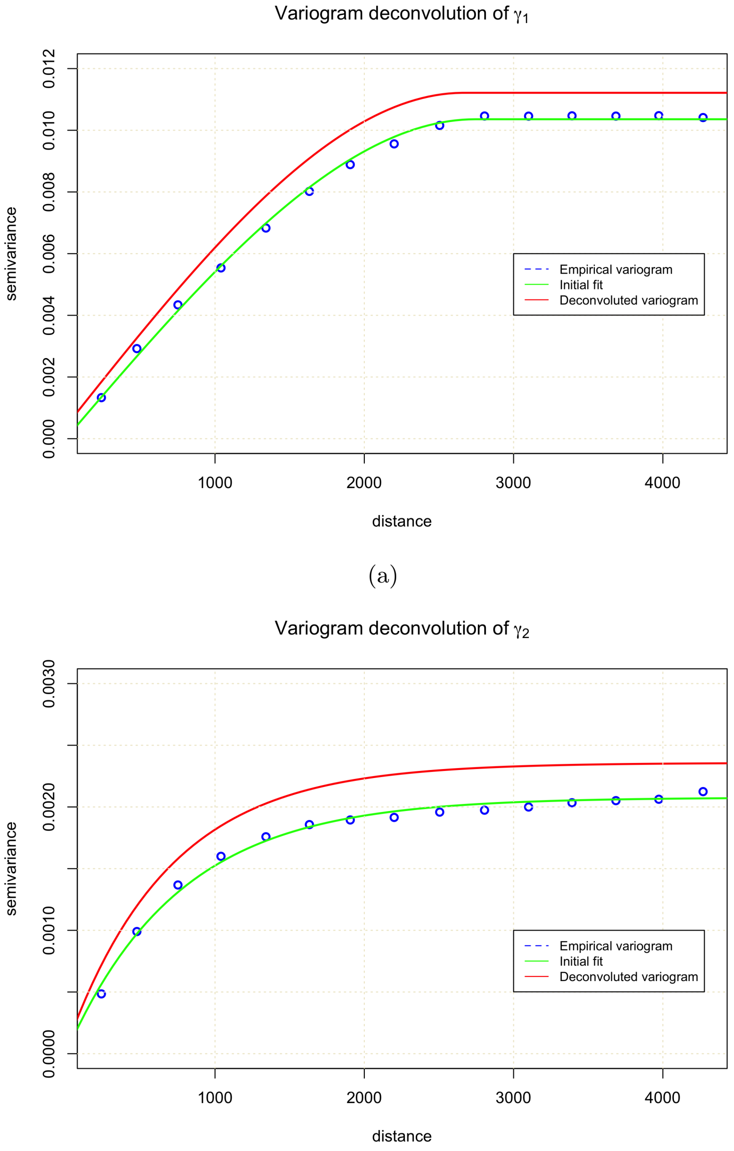


Figure 4.5: Graph of the normalised empirical variograms and of the cross-variogram of $\overline{ilr}_1$ and $\overline{ilr}_2$

Subtracting the identified trend component from the maps $\overline{ilr}_1$ and $\overline{ilr}_2$ we obtain the residual coarse maps $\overline{res}_1$ and $\overline{res}_2$. The distribution of the residuals are shown in Figure 4.4. Comparing with Figure 4.2, the distributions of the residuals are less skewed, in particular $\overline{res}_2$ is almost normal. Moreover, there is no correlation between the two residual components, i.e., the feeble negative correlation (-0.20) between the isometric log-ratios disappears when removing the DEM-related trend component from both variables.

To proceed with the second step of ATPRK, the spatial correlation structure of the residuals $RES_1(x)$ and $RES_2(x)$ has to be estimated. This can be done applying the deconvolution procedure described in Chapter 3 to the empirical variograms obtained from the coarse maps $\overline{res}_1$ and $\overline{res}_2$.

First, the empirical variograms $\hat{\gamma}_1$, $\hat{\gamma}_2$ and the empirical cross-variogram $\hat{\gamma}_{12}$ estimated from the coarse maps are considered. In order to evaluate the magnitude of the cross correlation, the variograms have been normalized (divided by the relative variance) and are plotted in the same graph (Fig. 4.5). As expected from the fact that no correlation between $\overline{res}_1$ and $\overline{res}_2$ was observed, the cross-variogram is negligible, so each variable is considered individually. The model better suited to represent $\hat{\gamma}_1$ is the *spherical* one, whereas $\hat{\gamma}_2$ shows an *exponential* behaviour.



(a)

(b)

Figure 4.6: Results of the Goovaerts' deconvolution procedure applied to the empirical variograms (a) $\hat{\gamma}_1$ and (b) $\hat{\gamma}_2$

Variograms are then deconvoluted using the Goovaerts' procedure, which has been implemented in R using package **gstat** [60], [30] for variogram fitting and variogram regularisation. The results are shown in Figure 4.6.

Now that the regression parameters and the variograms of the residuals have been estimated, it is finally possible to downscale the *ilr* maps using ATPRK:

$$\begin{aligned} ilr_1^*(x_i) &= \hat{\beta}_0^{(1)} + \sum_{l=1}^{2} \hat{\beta}_l^{(1)} \cdot DEM^l(x_i) + \sum_{k=1}^{K} \lambda_k^{(1)} \cdot \overline{res}_{k1},\ i = 1, ..., N, \\ ilr_2^*(x_i) &= \hat{\beta}_0^{(2)} + \sum_{l=1}^{2} \hat{\beta}_l^{(2)} \cdot DEM^l(x_i) + \sum_{k=1}^{K} \lambda_k^{(2)} \cdot \overline{res}_{k2},\ i = 1, ..., N. \end{aligned} \tag{4.2}$$

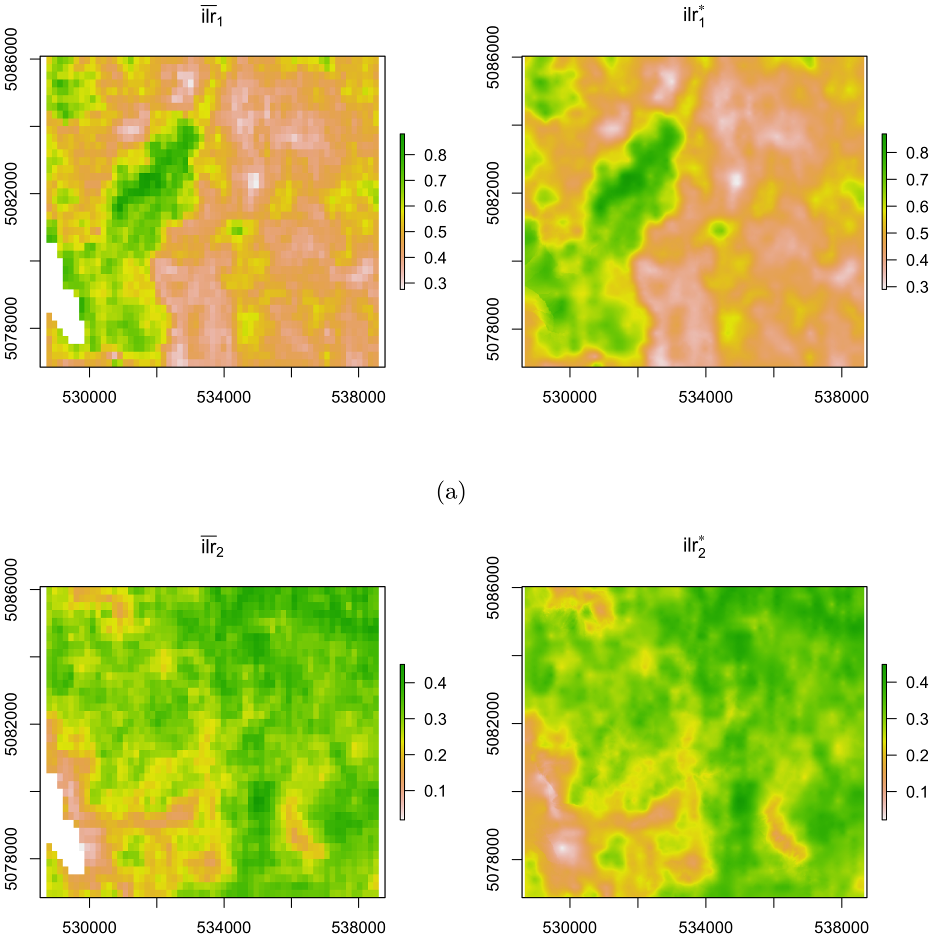
The $\lambda^{(i)}$ are computed by solving the ATPK system. The computation was performed using the function *krige* of the package *gstat*. Results of the downscaling are shown in Figure 4.7.

ATPK prediction have the effect of smoothing the initial coarse maps. The real importance of the probabilistic model of Equation (4.1) is that it allows to perform stochastic simulations of the residual fields conditional to the available data. The stochastic simulations can be used in a Monte Carlo setting to analyse how the uncertainty related to the soil properties affects the SMART-SED model and propagates to the output.

Since the distributions of $\overline{res}_1$ and $\overline{res}_2$ do not present any characteristic that would dismiss the multy-Gaussianity hypothesis (there are no very heavy tails nor multi-modal behaviour), the residual fields are assumed to be Gaussian processes respecting the properties previously mentioned and having covariance structures identified by γ_1 and γ_2, i.e.:

$$\begin{aligned} RES_1(x) &\sim GP(0, \gamma_1), \\ RES_2(x) &\sim GP(0, \gamma_2). \end{aligned}$$

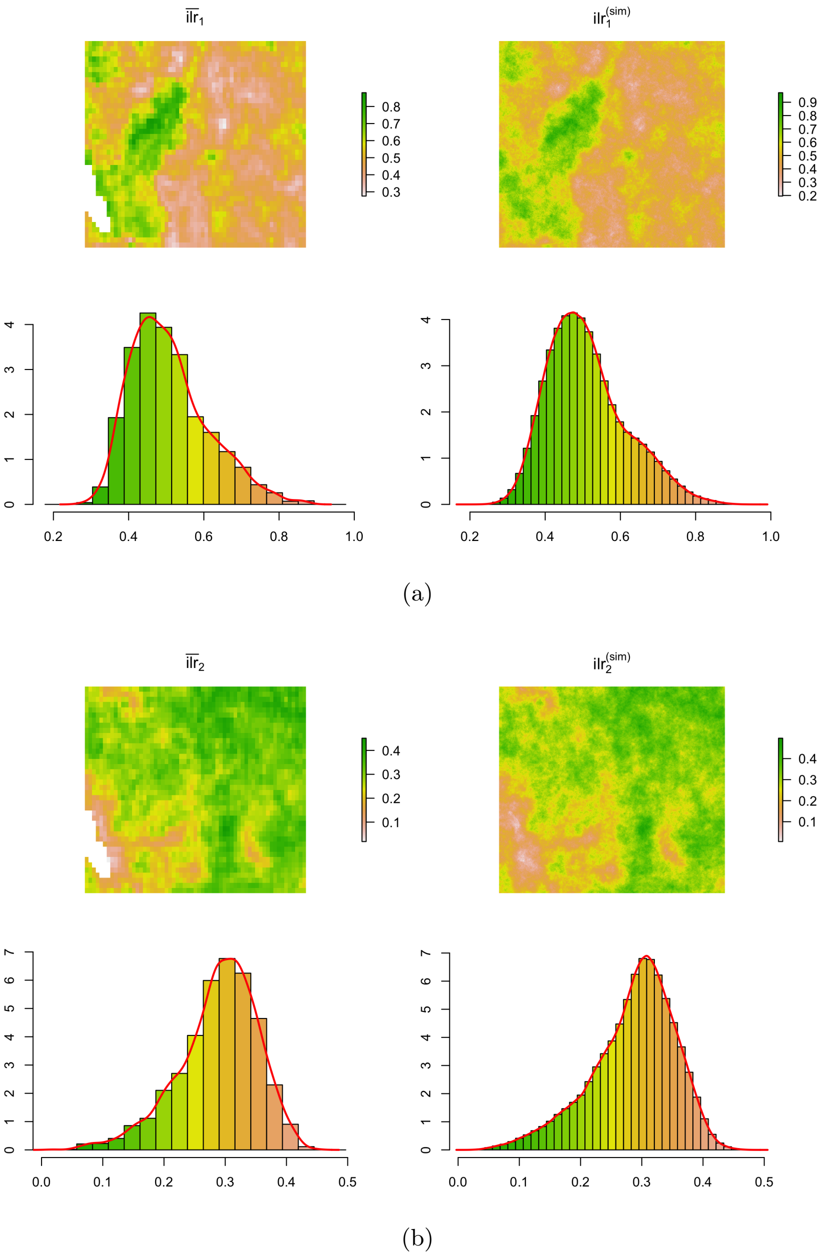
Block Sequential Gaussian Simulation is performed using function *krige* of package *gstat* (the same function used for ATPK). An example of realisation is shown in Figure 4.8. We notice how the simulated values respect the data distribution. The simulations depicts more realistic scenarios on how the fine scale maps would look like (with respect to the smooth map obtained by simple prediction). The results are back-transformed into clay, silt and sand percentages in Figure 4.9.



(a)

(b)

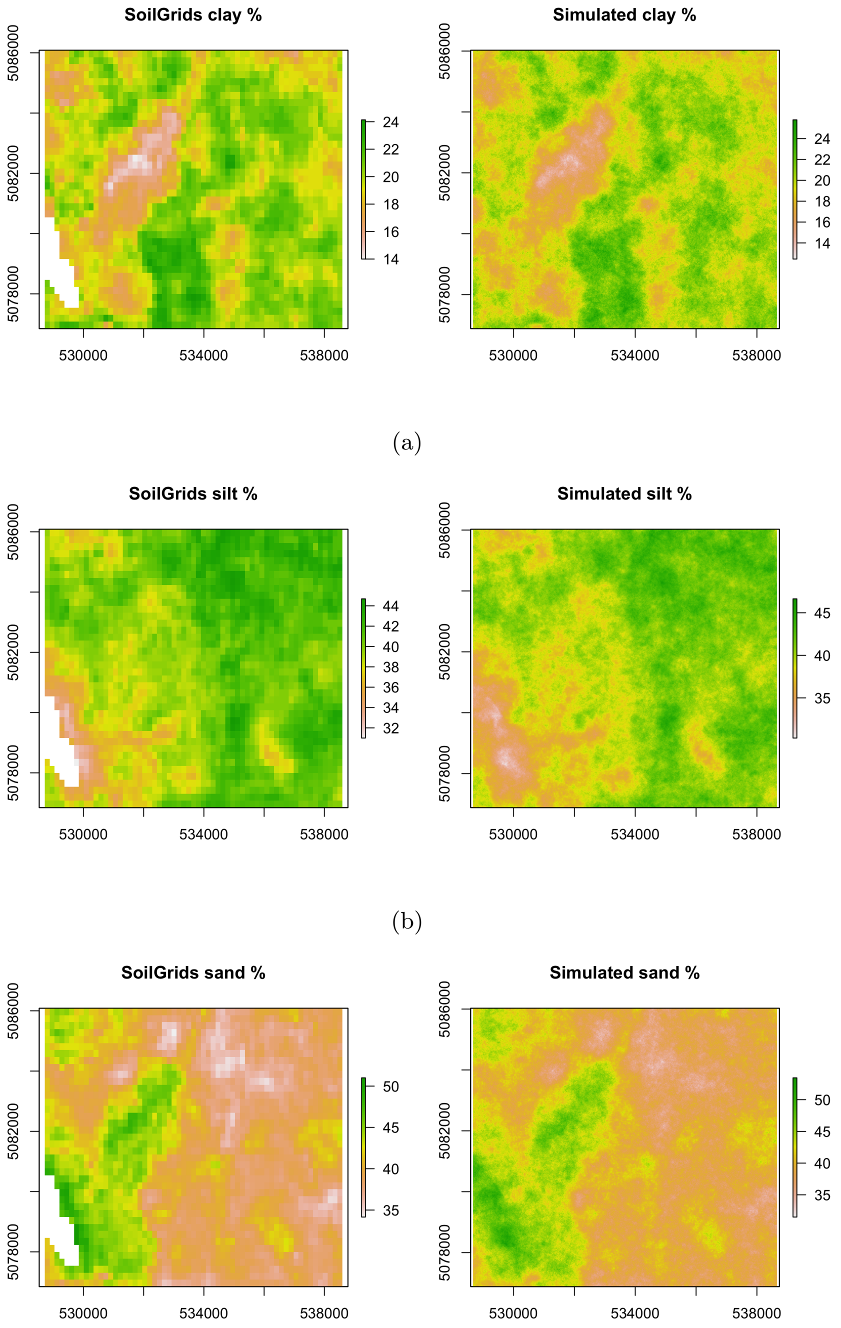
Figure 4.7: Results of the downscaling through ATPRK of the coarse resolution maps $\overline{ilr}_1$ (a) and $\overline{ilr}_2$ (b)



(a)

(b)

Figure 4.8: Block Sequential Gaussian Simulation (BSGS) applied to maps (a) $\overline{ilr}_1$ and (b) $\overline{ilr}_2$. The original histograms are reproduced by the simulations



(a)

(b)

(c)

Figure 4.9: Simulated isometric log-ratios are back-transformed, obtaining stochastic realisations at high resolution of (a) clay, (b) silt and (c) sand percentages in the topsoil

Maps $\overline{ilr}_1$ and $\overline{ilr}_2$ are SoilGrids predictions obtained using Machine Learning algorithm, and therefore they do not reflect the exact composition of the soil, but contain an additional error term $\epsilon(x)$. In the absence of supplementary information (for example hard data), the spatial distribution of the SoilGrids error $\epsilon(x)$ cannot be properly characterised.

Possible ways to add more noise to the simulations when no direct measurements are available are the following:

- Add a white noise to SoilGrids predictions. The value of the variance can be decided on the basis of expert considerations.

- Use a sub-sample of SoilGrids data for the conditional simulations.

At present, SoilGrids does not provide an estimation of the prediction variance for its maps. However, next versions will most likely provide confidence intervals, as it is a stated objective of the authors of the project [34]. With this information it will be possible to adjust the simulations in order to account for the additional error.

In the context of the SMART-SED model, soil composition is used to compute several variables (CN, Y, η, permeability and porosity cf. Chapter 1). The advantage of analysing the variability and performing simulations of the soil texture, instead of each variable considered singularly, is that in doing so the result are more coherent with the model assumptions [54].

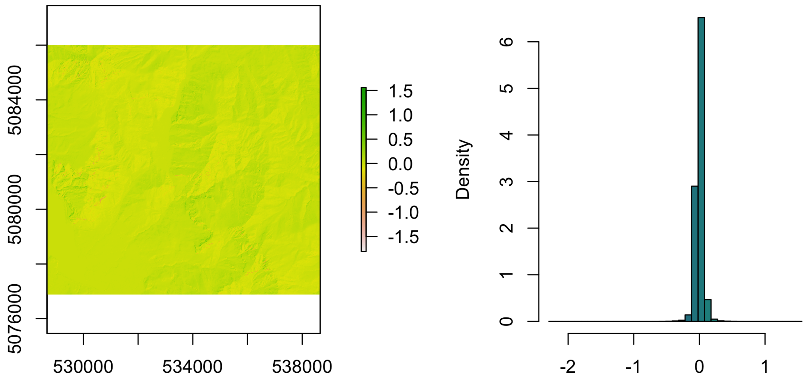


Figure 4.10: Map an histogram of curvature values computed with the McNab method [31]

4.2 Statistical downscaling of soil thickness

In the previous chapters it was explained how soil thickness can be described by three variables: Absolute Depth to Bedrock (ADB), Censored Depth to Bedrock (CDB) and Probability of occurrence of the R Horizon (PRH). The standard censoring threshold for PRH and CDB is 2 m, according to the literature [71], and it is generally set as the maximum depth of the *gravitational soil* (cf. Chapter 1). In the context of SMART-SED, soil thickness is used to compute the maximum quantity of water storable in the gravitational layer of the soil, therefore information of ADB when the thickness is estimated to be over 2 m is not relevant, only variables CDB and PRH will be considered.

Soil thickness is strongly related to topography. The dependence from topography is so relevant that often geostatisticians choose regression techniques with topographic covariates over Kriging, since the spatial correlation becomes a secondary aspect ([56], [41]). In particular, the two topographic variables that correlates the most with soil thickness are the elevation and the (absolute) **slope,** which here is intended as the angle (in radians) between the gradient of the DEM and the horizontal plane.

Another variable which is generally associated to soil thickness is the surface **curvature,** that can be computed as the second derivative of the DEM. Although some authors found a consistent relationship between curvature and depth to bedrock ([56]), there exist some issues related to the variable.

Curvature is a complex terrain derivative to compute, since it is sensible to noise in the data [20]. The best results for the computation of curvature in the study area are obtained with the method described in [31]. However, they still present some problems. In particular, most values are around 0, with the exception of very few sparse locations that present curvatures relatively much higher in absolute terms. A map is shown in Figure 4.10. This alters significantly the regression results, therefore curvature is not considered among the covariates.

The objective is to identify relationships of the kind:

$$\begin{aligned} PRH &= f(DEM, slope), \\ CDB &= g(DEM, slope). \end{aligned} \tag{4.3}$$

The use of ATPRK would limit the choice of possible models to the linear ones, but since we might expect nonlinear behaviour this would impact on the accuracy of the result.

Differently from particle-size fractions, for the variables in exam spatial correlation cannot be estimated from SoilGrids maps: in a catchment characterised by a rough surface like the one in exam, one would expect low-range spatial correlation (at most few hundred meters), especially when considered areas with thin soil; this kind of correlation cannot be estimated from coarse maps with a 250 m resolution.

For these reasons, the **dissever** algorithm, presented in the last section of Chapter 3 is used instead.

The procedure will follow these steps:

1. Downscaling of PRH using dissever combined with the **cubist** regression-tree method (cf. Section 3.3 and Section 7.5).

2. Use of the downscaled PRH as a *classification tool* to identify areas with thin soil.

3. Downscaling of CDB.

4. Prediction and unconditional stochastic simulations of CDB at the locations identified at step 2.

Cubist was chosen for its flexibility and because it shows results comparable to bagging methods such as *random forest*, but with significantly shorter execution times.

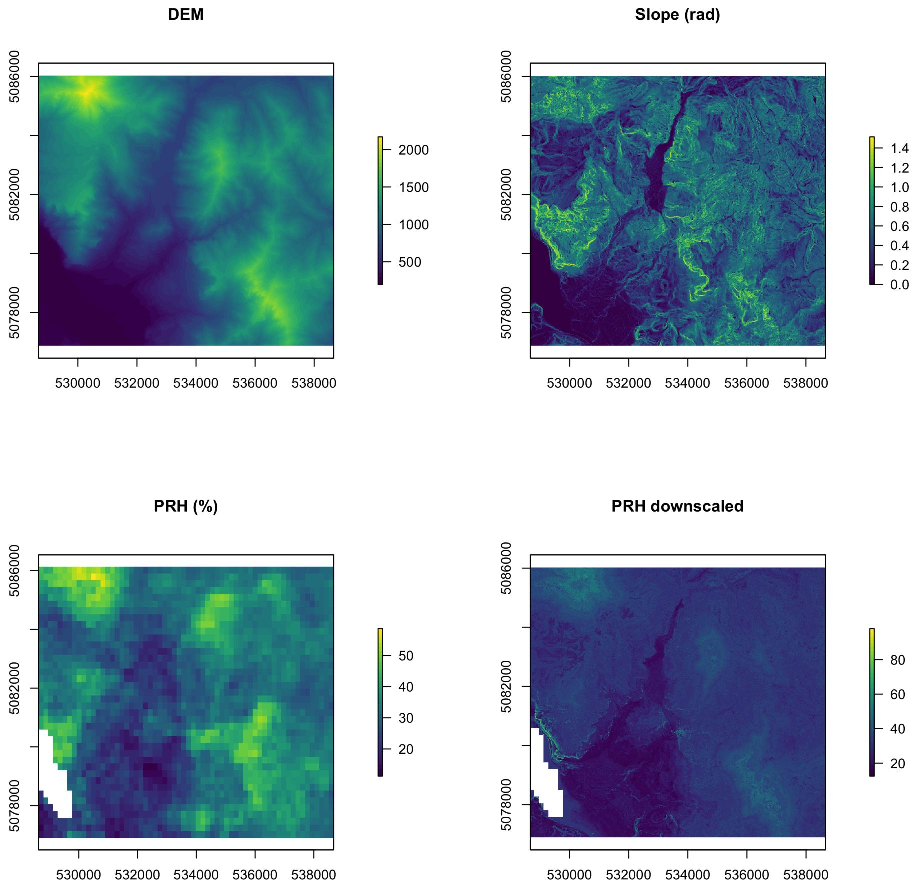
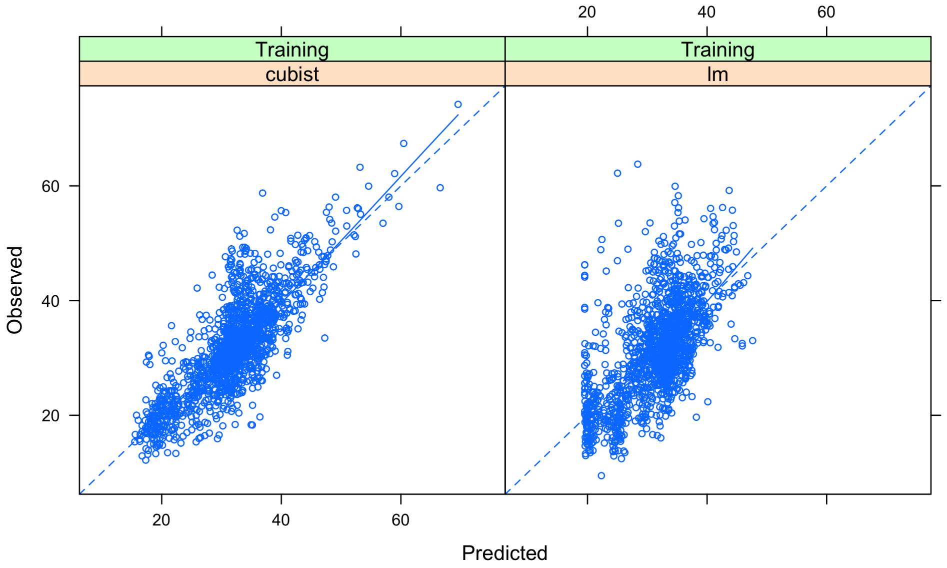
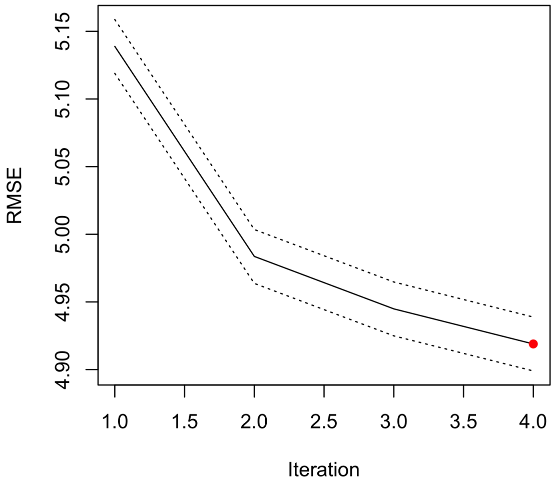


Figure 4.11: Result of the dissever algorithm with the cubist regression tree method for the statistical downscaling of PRH using the digital elevation and the derived slope as fine resolution covariates



(a)



(b)

Figure 4.12: SoilGrids prediction of the fraction of CDB (a) and PRH (b)

The results of the downscaling of PRH are shown in Figure 4.11. DEM and slope maps are reported, to ease the interpretation of the results. Dissever is an iterative procedure; the evolution of the RMSE at different iterations is shown in Figure 4.12, along with the scatter plots of the fitted Vs observed values for a subset of the samples used for the training. The performance of the cubist regression method is compared to that of the linear model (lm).

As expected, areas with steep slopes and/or at high elevations have higher values of PRH. Dissever allows to exploit fine scale topographic information to derive local probabilities at 5 *m* resolution in an effective way, considering both the accuracy of the fitted model with respect to the original coarse maps and heuristic considerations based on aerial views of the catchment (Fig. 2.7).

Considering how hard and expensive it is to measure soil thickness in mountain areas, the result obtained with the downscaling of the PRH map provided by SoilGrids are very promising. In fact, PRH is one of the predicted variables of SoilGrids with best estimated accuracy (AUC = 87%).

Now that a fine scale map with the probabilities of occurrence of the bedrock within the first 2 *m* is available, it is possible to use the map to classify the different locations into one of the two categories:

- **A:** soil thickness $< 2m$;
- **B:** soil thickness $> 2m$.

In order to do that, a possible method is to choose a *threshold* τ, indicating the minimum probability at which a location is considered to belong to group A. The choice of the threshold depends on the weight attributed to the two types of possible misclassification errors. Another possible way to choose τ is to consider the aerial view of the basin and set the threshold in such a way that most visible outcrops fall under the A category. Following this heuristic method, the threshold is set at **35%** of probability, the resulting map is shown in Figure 4.13.

Now that the locations of group A have been identified, it is necessary to estimate the (censored) soil thickness of those areas and evaluate the uncertainty of the estimations.

This time, cubist regression does not seem the be the best choice for the task. Differently from PRH, SoilGrids predictions of CDB are significantly less accurate, they explain only 35% of the total variability of the data used by SoilGrids for the fitting, with an RMSE equal to 1/4 of the total range (50 *cm*). The advantage of using sophisticated regression methods when dissevering is that the results honour more accurately the coarse maps downscaled. However, when the original low resolution data are intrinsically inaccurate, this aspect becomes less relevant.

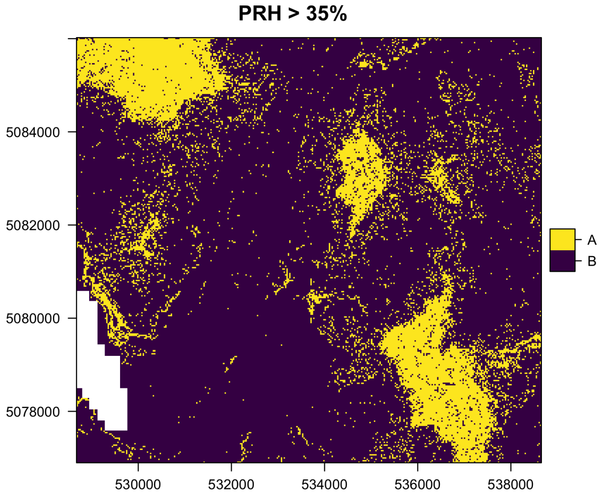


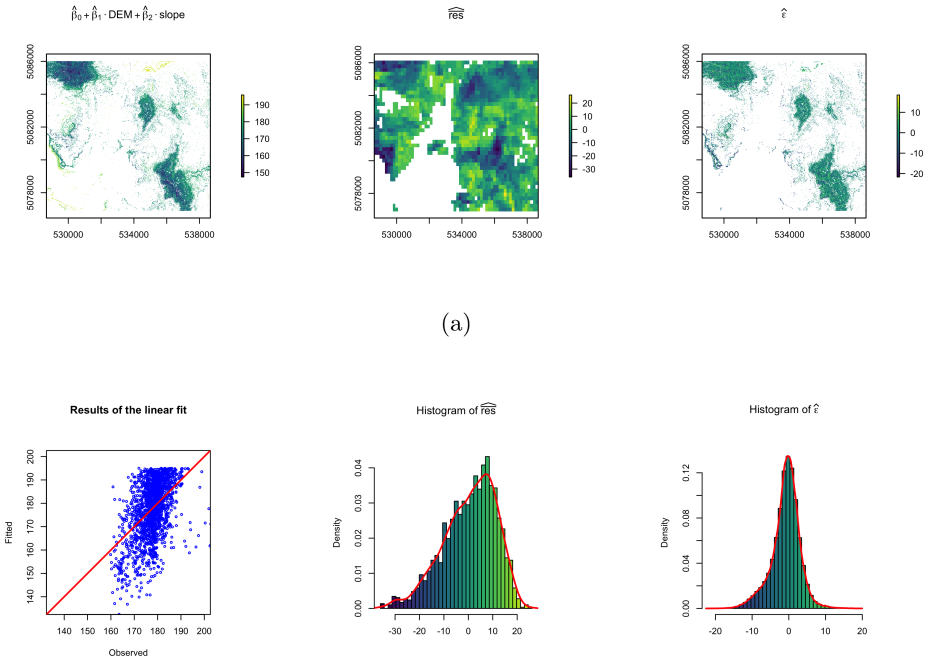
Figure 4.13: Distinction between areas with PRH $>$ 35% (group **A**, and areas with PRH $\leq$ 35% (group **B**)

Downscaled values obtained using dissever+cubist for CDB exceed the threshold of 2 m on several locations, often conflicting with PRH predictions.

For the enumerated reasons, the choice is to use a simpler linear regression, fitted using the un-censored observations, to derive a relation of the kind

$$CDB^*(x) = \beta_0 + \beta_1 \cdot DEM(x) + \beta_2 \cdot slope(x) + \overline{res}(x) + \epsilon(x),$$
$$CDB(x) = \begin{cases} CDB^*(x) & if\ CDB^*(x) < 2\ m \\ 2\ m & otherwise \end{cases} \tag{4.4}$$

In (4.4), $\overline{res}(x)$ is the residual of the linear regression for the block to which point x belongs; ϵ is an additional Gaussian error term, which is assumed to be *spatially uncorrelated*. The model resembles the one used for the isometric log-ratio transformations of particle-size fractions (cf. Equations (4.1)). However this time the residual term in not modelled as a spatially correlated random field, but rather as a white noise with varying means over the blocks ($\overline{res}+\epsilon(x)$), since the spatial correlation cannot be estimated from the coarse map, having a range inferior to the resolution.



(a)

(b)

Figure 4.14: Results of the downscaling of CDB on the areas of group A: (a) maps of the estimated terms of Equation (4.4); (b) scatter plot of fitted Vs observed data, histograms of the block residuals and of the additional error ϵ

The downscaled maps obtained fitting model (4.4) are shown in Figure 4.14, where only locations of group A are considered. The hats over the variables indicate that the values are estimates obtained through linear regression using SoilGrids data. Stochastic simulations can be generated by simulating error maps of $\epsilon(x)$.

Obviously, the precision of the estimates obtained under these model assumptions strongly depends on the precision of SoilGrids predictions. In the presence of direct measurements it would be possible to adjust the models and increase the accuracy. SoilGrids maps, in particular PRH provides information on the areas which are more likely to present thin soil, helping to choose the locations at which conduct local surveys. The observations obtained could be easily integrated in the models presented. Next chapter contains several suggestions on how to carry out this integration.

Chapter 5

Discussion of the results and their use in SMART-SED

In this chapter results are discussed and some proposals are made on how to use them for the Uncertainty Quantification in the SMART-SED model. Some strategies for the integration of direct measurements in the analysis are also proposed.

5.1 Accuracy of the results and integration of hard data

SoilGrids predictions are the result of spatial regression techniques at a very broad scale, they cannot fully capture the variability at a local scale, let alone provide very accurate estimates of soil properties at fine resolution. This aspect has to be taken into consideration when SoilGrids data are used in the context of numerical simulations for risk assessment, especially when the results are involved in the decision making processes of local authorities. For instance, SoilGrids predictions can hardly be used to estimate a *realistic* spatial correlation structure for the variables at local scale. In particular, SoilGrids predictions have a smoother spatial variation than one would expect from data collected with direct measurements, resulting in variograms with greater ranges and lower sills. Variogram deconvolution only accounts for the coarse nature of the data, but not for the fact that they are average predictions, and thus filter out an additional noise component. Moreover, an important limitation of variogram deconvolution is that it does not allow to infer the nugget effect in the correlation [23]. This is a problem especially for the censored depth to bedrock (CDB), and is one of the reasons because spatial correlation of CDB was not considered in Chapter 4. The choice of using SoilGrids maps with variogram deconvolution for the estimation of the covariance of psf, is due to the fact that SoilGrids predictions for psf are more accurate with respect to soil thickness (cf. Table 2.1 in Chapter 2), and variograms with (relatively) high range and low sill are also more realistic in the case of psf [12]. Furthermore, the objective of this work is not to provide extremely accurate high resolution maps of soil properties, but rather to present a mathematical framework for the variability assessment of soil-related quantities that minimises the dependence on direct measurements and expert intervention, but uses easily accessible and widely used online databases. The procedures described in this thesis can also be applied to remotely sensed images, rather than Digital Soil Maps. In this context variogram deconvolution is a powerful and widely used technique [52].

An alternative to variogram deconvolution, when prior knowledge of the area in exam is available, consists in using a Bayesian approach through expert elicitation of prior spatial structure information [74]. This approach has an important advantage: the additional parameter uncertainty considered contributes largely to Area-to-Point Kriging Uncertainty, which is a desirable property in Uncertainty Quantification (UQ).

Another alternative presents itself when *hard data* obtained from local measurements are available. In this case, the variogram structure can be estimated directly using the punctual ”hard” observations. Since direct measurements often require costly interventions, the number of observations from measurements are most of the times limited in number. For this reason, the use of Digital Soil Maps can prove really useful even when hard data are available. An advantage of the downscaling and simulation methods presented in this thesis is that they easily allow for the integration of coarse resolution digital maps and precise punctual observations, without having to modify the algorithms. The process of integration of SoilGrids maps with direct measurements is even suggested by the authors of SoilGrids as an effective way to improve the accuracy of the estimations and the variability assessment (cf. Fig. 5.1). Possible ways to implement the integration are presented in the next subsections.

5.1.1 Combining areal and point data in geostatistical interpolation and simulation

In Chapter 3 it is shown how to combine Kriging and Area-to-Point Kriging (Eq. (3.29)) when dealing with both areal and point data. The procedure was formalised by Goovaerts ([22], [21]), and can be used for prediction and for Block Sequential Simulation (BSS). Unfortunately, at the moment there isn’t any R package allowing to perform this task. One of the best open source softwares allowing to combine areal and point data for geosatatistical interpolation and simulation is **SGeMS (Stanford Geostatistical Modeling Software)**

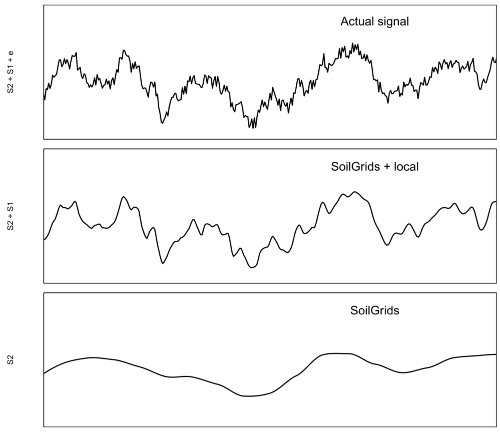


Figure 5.1: SoilGrids can be considered the coarsest component of the global soil variation "signal" curve. Source: [34]

[66]. In particular, SGeMS provides a function to perform fast block sequential simulation (conditioning to both point and areal data) that does not require the multi-Gaussianity assumption, and can reproduce any given point distribution [43].
The function uses the algorithm devised by Soares (2001) called *Direct Sequential Simulation (DSS)* [72], and is the first choice in geostatistics whenever the normality assumption is not valid. The algorithm has been optimised in order to be able to handle numerous block data by using specialised methods to compute the block covariances. Another interesting aspect of the BSS algorithm implemented in SGeMS, is that it allows to account for a noise in the block-data conditioning the simulation. This is done by adding a diagonal matrix containing the variances of the errors associated to each block to the covariance matrix when solving the ATPK system (3.15) at each node.

[53] proposed a more sophisticated two-stage geostatistical integration approach, aiming at downscaling of coarse scale remote sensing data. First, downscaling of the coarse scale secondary data is implemented using Area-to-Point Kriging. This result is used as trend components in the second integration stage: simple Kriging with local varying means that integrates sparse precise observation data with the downscaled data is applied to generate thematic information at a finer scale. This approach not only can account for the statistical relationships between precise observation and secondary data acquired at the different scales, but also to calibrate the errors in the secondary data through the integration with precise observation data.

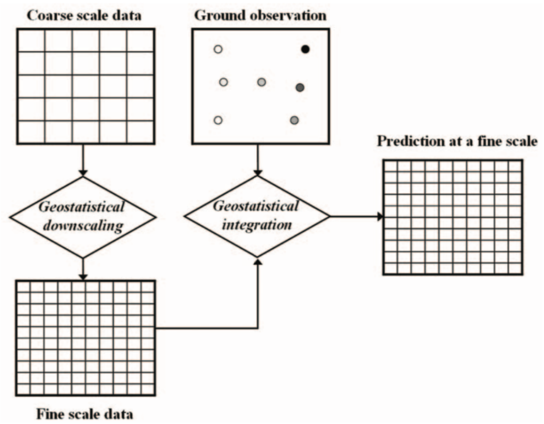

Coarse scale data
Ground observation
Prediction at a fine scale
Geostatistical downscaling
Geostatistical integration
Fine scale data

Figure 5.2: Scheme of the two-stage geostatistical approach presented by Park. Source: [53]

Dissever with inclusion of point data

Dissever derives point values from areal observations and subsequently updates them at each iteration, by fitting a regression model to the point data previously obtained. The integration of precise point information in the algorithm is straightforward: point values at known locations can be set equal to the observed data and kept constant through out the iterations. On top of that, the known values can be weighted more in the regression phase, increasing the accuracy of the estimation.

This approach was suggested by the authors of dissever [51] (cf. Fig. 5.3). At the moment, the R package *dissever* does not allow to integrate point observations, the extension of the function could be the subject of future work.

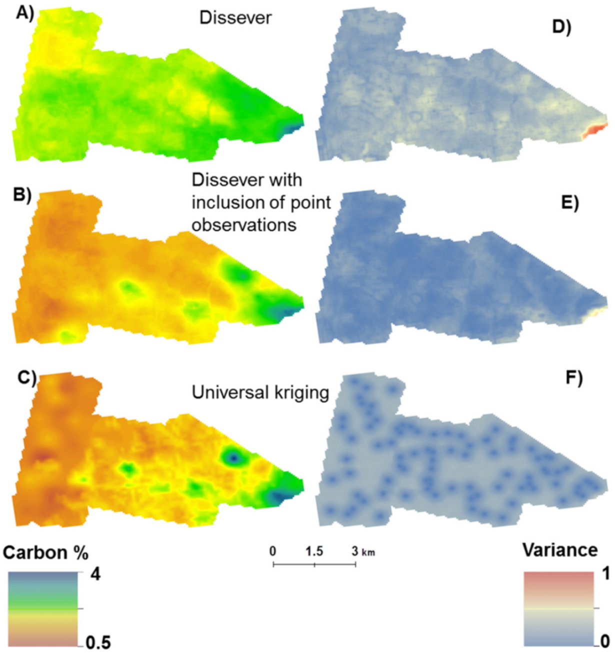


Figure 5.3: Figure taken from [51] showing the effectiveness of the integration of precise point data in the dissever procedure. The results for the downscaling of Soil Organic Carbon (SOC) content in the soil are compared to those of Universal Kriging (other name for Regression Kriging)

5.2 Uncertainty propagation in the SMART-SED model

The analysis of the variability of soil thickness and soil texture is part of the Uncertainty Quantification (UQ) process of the SMART-SED numerical model. In particular, this thesis covers the study of the *parametric uncertainty* (cf. Chapter 1), for the model parameters related to the properties of the soil, namely the parameters of the Gavrilovic model (X, Y and ξ), the infiltration rates in dry and saturated conditions (f_0, f_c), and the *Curve Number* of the *SCS-CN method.*

An important part of UQ is relating the probability distribution of the outputs of the model to the distribution of the input parameters. Consider for instance a computer code performing a numerical simulation of a physical model that takes as input parameters $x_1, ..., x_n$ and generates an output y. Since the simulation is deterministic, each set of parameters $x_1, ..., x_n$ leads to an unique output, which therefore is related to the inputs through an unknown *("black-box")* function

$$y = f(x_1, ..., x_n).$$

If the inputs are random variables $X_1, ..., X_n$ over a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, so is $Y = f(X_1, ..., X_n)$. As long as a method exits allowing to generate N stochastic realisations $x_1^{(i)}, ..., x_n^{(i)},\ i = 1, ..., N$ of $X_1, ..., X_n$ from their original distribution, then this would allow to obtain N samples of Y :

$$\begin{aligned}
y^{(1)} &= f(x_1^{(1)}, ..., x_n^{(1)}),\\
y^{(2)} &= f(x_1^{(2)}, ..., x_n^{(2)}),\\
&...\\
y^{(N)} &= f(x_1^{(N)}, ..., x_n^{(N)}),
\end{aligned}$$

by solving the numerical model N times. The samples could be used to infer the distribution of the output associated to the uncertainty of the inputs, or just to approximate the average response $\hat{y}_\mu = \frac{1}{N}\sum_{i=1}^{N} y^{(i)} \approx \mathbb{E}[Y]$ associated to the variable input parameters (which, in general, is different from the response associated to the average inputs) and its variance $\hat{\sigma}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y^{(i)} - \hat{y}_\mu)^2$. This procedure allows to study how the uncertainty of the inputs *propagates* to the outputs of a deterministic computer code, and since it allows the numerical approximation of an expected value through random simulations (like in Monte-Carlo methods) it is called **Monte-Carlo Uncertainty Propagation** [70]. The advantage of this method is that, given a proper characterisation of the probabilistic distribution of the inputs, it allows not only to approximate the average output, but also to characterise its distribution. However, often the computations involved in the solution of the numerical models are exceedingly time-consuming to allow to generate a sufficient number of realisations of Y. Moreover, this procedure does not provide any insight on the nature of function f, relating inputs and output. Alternative methods have been devised allowing to directly approximate function f. These methods are called **metamodeling techniques** [78], since their goal is to construct a *model of the model* $Y = f(X_1, ..., X_n)$. Some of the most common metamodeling methods are based on *Co-Kriging*. In particular these methods can be applied whenever fast approximations of the computer code in exam are available, which is the case of the SMART-SED model (for instance by considering coarser discretisation grids). The input parameters $x_1, ..., x_n$ are interpreted as *locations* of an n-dimensional Euclidean space. A (relatively) large number N of values $y_{lf}^{(1)}, ..., y_{lf}^{(N)}$ corresponding to the same number of different locations $\mathbf{x}^{(i)} = (x_1^{(i)}, ..., x_n^{(i)}),\ i = 1, ..., N$ is computed by running the fast approximation available, where lf stands for *low-fidelity,* since those outputs are resulting from an approximation of the model. On a subset of $M << N$ of these locations, the (slow) computer code is run to obtain the *high-fidelity* outputs $y_{hf}^{(1)}, ..., y_{hf}^{(N)}$. Cokriging is then performed, using the $y_{lf}^{(i)}$ as secondary information, to derive an approximation of the function f. This technique is called **Multi-Fidelity Co-Kriging (MFCoK)** and was first proposed by [37], a particular formulation allowing for a fast and reliable implementation of the method was proposed by [27]. The method has been extended to the case of multivariate outputs by [55], and finally to the case of an output in a generic Hilbert space (even infinite-dimensional) by [29]. Multi-fidelity Co-Kriging works well when the number of input parameters and their range is contained [37], [27]. However, in the case of the SMART-SED model, the input parameters are *maps of soil properties* $Z(x),\ x \in D$, varying over a domain possibly composed of millions of pixels, therefore MFCoK cannot be used directly. Nevertheless, the approach could be used when considering ulterior **hyperparameters,** for instance, the range and the sill of the variograms of particle-size fractions or of soil thickness. These parameters are uncertain, therefore, instead of providing a punctual estimation, it could be more appropriate to identify some ranges and study how the model responds to different values in those ranges, by performing simulations of the input maps using the variograms identified by the hyperparameters and applying Monte-Carlo uncertainty propagation, thus combining the two techniques presented in this final section.

Conclusions and future developments

In this thesis, some methods were proposed to use SoilGrids maps for the analysis of the variability of soil properties at local scales. These approaches can be extended to any source of soil information in the form of coarse resolution maps, be it satellite images or aggregated data from radars and other data-collection devices. In particular, a novel procedure for the downscaling of compositional data has been presented. Possible future developments could involve the study of the bias introduced when the *centre preserving property* is not valid, and the introduction of a possible correction.
Another obvious improvement would be the extension of the methods presented through the integration of hard data in the analysis, as suggested in Chapter 5. In particular, this feature could be added to the R packages *gstat* and *dissever*.
The next versions of the SoilGrids maps will likely contain confidence intervals for the predictions [34], this information would allow a better characterisation of the variability of the soil properties and could be used to add more noise when performing stochastic simulations.

Regarding the SMART-SED model, the variability of soil-related inputs have been studied, and a probabilistic model allowing to simulate maps of the parameters has been established. The next step for the Uncertainty Quantification is the analysis of the propagation of the variability of the inputs to the outputs, through Monte Carlo (MC) simulation and/or metamodelling techniques.

Appendix

7.3 Linear Coregionalization Model

This sections contains a presentation of the general framework for the estimation of covariance structures in Co-Kriging with multiple variables. The model presented is explained in detail in [25]. Suppose that we have measured the values of p variables at N different locations of a domain $D \subset \mathbb{R}^2$. This gives us a set of p-dimensional vectors, we assume these vectors as a part of a realisation of multivariate random field $\mathbf{Z}(x) = (Z_1(x), ..., Z_p(x))^T$ with the following properties:

$$\begin{aligned} \mathbb{E}[\mathbf{Z}(x) - \mathbf{Z}(x+\mathbf{h})] = (0, ..., 0)^T; \\ \mathbf{C}(\mathbf{h}) = \frac{1}{2}\,\mathbb{E}[(\mathbf{Z}(x) - \mathbf{Z}(x+\mathbf{h}))(\mathbf{Z}(x) - \mathbf{Z}(x+\mathbf{h}))^T] = \sum_{k=1}^{s} \mathbf{S}_k \cdot \gamma_k(|\mathbf{h}|), \end{aligned} \tag{7.1}$$

where the $\gamma_k(h) = \gamma_k(|\mathbf{h}|)$ are known variograms, the $\mathbf{S}_k$ are unknown non-negative matrices and s denotes the number of structures. Each matrix $\mathbf{S}_k$ can be though of as a scalar product matrix between variables. The primary aim is to estimate the unknown matrices $\mathbf{S}_k$. In order to do that, we must first guess what the theoretical basic functions γ_k are. This could be done heuristically by inspecting the sampling matrix visually and obtaining a first approximation of the matrices $\mathbf{S}_k$, as suggested in [25]. The estimation is then carried out through an iterative procedure:

1. Compute the empirical cross-covariance matrix $\hat{\mathbf{C}}(h)$:

   $$\hat{\mathbf{C}}(h) = \frac{1}{2|N(h)|} \sum_{(x_i,x_j)\in N(h)} (\mathbf{Z}(x_i) - \mathbf{Z}(x_j))(\mathbf{Z}(x_i) - \mathbf{Z}(x_j))^T,$$

   where $N(h) = \{(x_i, x_j) : h - \Delta h \leq |x_i - x_j| \leq h + \Delta h\}$.

2. Choose a criterion to measure the goodness of fit between a model $\mathbf{C}(h)$ and $\hat{\mathbf{C}}(h)$. Typically, an Euclidean-like distance is used: let $\mathbf{V}$ be a positive definite matrix and $w(h_j),\ j = 1, ..., m$ be positive weights, then

$$WSS = \sum_{j=1}^{m} w(h_j) \cdot Tr\left(\left(\mathbf{V}(\mathbf{C}(h_j) - \hat{\mathbf{C}}(h_j))\right)^2\right).$$

   WSS stands for *Weighted Sum of Squares,* the weights w are generally proportional to the number of pairs used in the estimate of $\hat{\mathbf{C}}(h)$, and $\mathbf{V}$ is generally the diagonal matrix of inverse variances, to avoid favouring variables with larger variances.

3. Estimate the $\mathbf{S}_k$ by minimising WSS when $\mathbf{C}(h)$ is assumed to be of the form $\mathbf{S}_1\gamma_1(h) + ... + \mathbf{S}_s\gamma_s(h)$. A minimisation algorithm is described in [25].

Choosing a correct number and proper models for the basic structures γ_k is crucial for a "good" estimation of the model. Two situations may arise after a first estimation:

- The plot of the non-parametric covariogram estimates and of their theoretical fitted models shows that the fit is poor. Then, at least one basic structure must be modified or another basic structure must be added into the model.

- The plot shows the model fits well. In this case, it may be worth checking if some basic structure can be omitted, since too many basic structures can cause instability in the estimation results, thus leading to false interpretations.

7.4 Isometric log-ratio transformations

Isometric log-ratio transformations (*ilr*) are a class of linear isometries from the n-dimensional real simplex

$$\mathbb{S}^n = \{(z_1, ..., z_n)^T : z_1, ..., z_n > 0,\ z_1 + ... + z_n = c\},$$

embedded with the Aitchison geometry into $\mathbb{R}^{n-1}$ with the standard Euclidean structure. This class of transformations was first introduced in 2003 by Egozcue and Pawlowski-Glahn [17].Since their introduction, *ilr*s have been one of the most common mathematical instrument for the statistical analysis of compositional data, allowing to operate in the context of the Aitchison geometry without modifying the standard statistical techniques. *ilr* exploits the fact that the simplex embedded with the Aitchison geometry is a finite Hilbert space, so it is possible to find an **orthonormal basis** $\mathcal{B} = \{\mathbf{e}_1, ..., \mathbf{e}_n\}$. Applyng an *ilr* on an element $\mathbf{z} \in Sx^n$ simply consists in projecting $\mathbf{z}$ on the basis $\mathcal{B}$ using the inner product in the Aichison simplex:

$$ilr(\mathbf{z}) = \left(\langle \mathbf{z}, \mathbf{e}_1 \rangle_a, ..., \langle \mathbf{x}, \mathbf{e}_{n-1} \rangle_a\right)^T,$$

thus obtaining a vector of $n-1$ coordinates $\mathbf{y} \in \mathbb{R}^{n-1}$.

Any valid orthonormal basis can be used. Although the Graham-Schmidt algorithm could be used to obtain one, explicit methods exist allowing to determine orthonormal basis with specific properties for any dimension. In particular, in [17] a special class of orthonormal basis is introduced. The idea is to associate the basis to a partition of the of the compositional vector $\mathbf{z} \in \mathbb{S}^n$ into two subcompositions, $\mathbf{z} = (\mathbf{r}, \mathbf{s})$, where $\mathcal{C}(\mathbf{r}) \in \mathbb{S}^r$, $\mathcal{C}(\mathbf{s}) \in \mathbb{S}^s$, $n = r + s$, $r \geq 2$, $s \geq 2$. Note that for $r = n - 1$ we get $s = 1$, which corresponds to a degenerate case in the sense that we have a subcomposition with only one part. In the nondegenerate case, we look for an orthonormal basis such that the compositions of the form

$$\mathcal{C}(\mathbf{r}, \mathbf{c}_s) = \left[z_1, z_2, ..., z_r,\ \underbrace{c, c, ..., c}_{s \text{ components}} \right],$$

$$\mathcal{C}(\mathbf{c}^*_r, \mathbf{s}) = \left[\underbrace{c^*, c^*, ..., c^*}_{r \text{ components}}, z_{r+1}, z_{r+2}, ..., z_n \right],$$

can be expressed by using $r-1$ elements of the orthonormal basis, denoted by $\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_{r-1}$ for the first one,and $s-1$ elements, denoted by $\mathbf{h}_{r+1}, \mathbf{h}_{r+2}, ..., \mathbf{h}_{n-1}$, for the second one. In this way, we associate the first $r-1$ elements of the basis with the subcomposition $\mathcal{C}(\mathbf{r})$ and the last $s-1$ with the complementary subcomposition $\mathcal{C}(\mathbf{s})$.

To complete the orthonormal basis, we need an additional orthonormal vector, which we call **balancing element** and denote by $\mathbf{h}_r$; the orthogonal projection—within the Aitchison geometry— of $\mathbf{z}$ on this element defines the mass ratio between the subcompositions $\mathbf{r}$ and $\mathbf{s}$ to reconstruct $\mathbf{z}$. Such an Aitchison-orthonormal basis can be defined as follows [17].

Proposition 8. *Let $r \geq 2, s \geq 2,$ be such that $n = r + s$. The compositions in $\mathbb{S}^n$ defined as*

$$\mathbf{h}_i = \mathcal{C}\left[\exp\left(\underbrace{\sqrt{\frac{1}{i(i+1)}}, ..., \sqrt{\frac{1}{i(i+1)}}}_{i\ elements}, -\sqrt{\frac{1}{i(i+1)}}, 0, ..., 0\right)\right], \qquad (7.2)$$

for $i = 1, ..., r-1$,

$$\mathbf{h}_r = \mathcal{C}\left[\exp\left(\underbrace{\sqrt{\frac{s}{n \cdot r}}, ..., \sqrt{\frac{s}{n \cdot r}}}_{r\ elements}, \underbrace{-\sqrt{\frac{r}{n \cdot s}}, ..., -\sqrt{\frac{r}{n \cdot s}}}_{s\ elements}\right)\right], \qquad (7.3)$$

and, for $j = 1, 2, ..., s-1$,

$$\mathbf{h}_{n-j} = \mathcal{C}\left[\exp\left(0, ..., 0, -\sqrt{\frac{j}{j+1}}, \underbrace{\sqrt{\frac{1}{j(j+1)}}, ..., \sqrt{\frac{1}{j(j+1)}}}_{j\ elements}\right)\right], \qquad (7.4)$$

are Aitchison-orthonormal and constitute a basis of $\mathbb{S}^n$ associated with a partition into two orthogonal subcompositions with, respectively, r and s components.

Proposition 8 is still valid in the degenerate case $r = n - 1,\ s = 1$. The orthonormal basis consists then of those vectors given in Equations (7.2) and (7.3), whereas the vectors given by Equation (7.4) do not appear at all. Note that the dashed lines in Figure 3.8 represent an orthogonal base corresponding to the degenerate case $r = 2,\ s = 1$, the straight dashed axis (in an Euclidean sense) corresponding to the balancing element and the other axis associated to the subcomposition $(z1, z2)$. The result can be generalised to more than two subcompositions.

Thus, the main consequence of Proposition 8 is that any $\mathbf{z} \in \mathbb{S}^n$ can always be decomposed into three orthogonal parts: two of them in orthogonal subspaces associated with two subcompositions and a balancing one. In particular, it is possible to derive an explicit expression for the *ilr* transformation using the basis defined in Proposition 8: let us define

$$\mathbf{h} = \mathcal{C}\left[\exp\left(\underbrace{\frac{1}{r},...,\frac{1}{r}}_{r \text{ elements}}, \underbrace{\frac{1}{s},...,\frac{1}{s}}_{s \text{ elements}}, \underbrace{0,...,0}_{t \text{ elements}}\right)\right], \quad r+s+t=n.$$

It can be shown that

$$\frac{\langle \mathbf{z}, \mathbf{h}\rangle_a}{||\mathbf{h}||_a} = \sqrt{\frac{rs}{r+s}} log\left(\frac{g(z_1,...,z_r)}{g(z_{r+1},...,z_{r+s})}\right) \tag{7.5}$$

where $g(\cdot)$ denotes the geometric mean. The elements of the basis (7.2)-(7.4) have the form $||\mathbf{h}||_a^{-1} \odot \mathbf{h}$; then, *ilr* transformations with respect to these bases can be expressed using Equation (7.5).

The default basis used by function *ilr* of the R package *compositions* is the one in (7.2)-(7.4). The same basis in the degenerate case $r = 2$, $s = 1$ has been used for the isometric log-ratio transformation of particle-size fractions in Chapter 4.

Other methods to derive possible bases can be found in [19], [58].

7.5 Cubist

Cubist is an empirical learning method based on **regression trees** [8]. It implements an extended version of the **M5** method by Quinlan [63]. The exact algorithm behind the cubist method is proprietary to the author (Quinlan) [9], but the main ideas behind the M5 method are provided in [63]. M5 builds tree-based models but, whereas classical regression trees have values at their leaves, the trees constructed by M5 have multivariare linear models; these *model trees* are thus analogous to piecewise linear functions.

Suppose we have a collection of T training cases (statistical observations), specified by a set set of p values corresponding to the same number of (discrete or continuous) attributes, and the value of the associated target variable. The aim is to construct a model relating the target variable to the attributes, using the information contained in the training cases. The worth of the model is generally measured by the accuracy with which it predicts the target values of a test set of cases. Tree-based models are constructed by iteratively splitting the training set T into subsets using a particular splitting criterion or **test.** The first step in building a regression tree is to compute the standard deviation of the target values of cases in T. Every potential test used to split T is evaluated by determining the subset of cases associated with each of its outcomes. Let T_i denote the subset of cases that have to i-th outcome of the potential test. If we treat the standard deviation $sd(T_i)$ of target values of cases in T_i as a measure of error, the expected reduction in error resulting from this test is

$$\Delta error = sd(T) - \sum_i \frac{|T_i|}{|T|} \times sd(T_i).$$

After examining all possible tests, the one that maximises the expected error reduction is chosen. This procedure is proper of every regression-tree method, the innovations of the M5 method come into play after the initial tree has been grown:

- *Error estimates.* To estimate the error of a model derived from a set of training cases, M5 first determines the average residual (difference between prediction and actual target value) of the model on the training cases. This will generally underestimate the error on unseen cases, so M5 multiplies the value by $(n + \nu)/(n - \nu)$, where n is the number of training cases and ν is the number of attributes used in the model.This increases the estimated error of models with few training cases and many attributes.

- *Linear models.* A multivariate linear model is constructed for the cases at each node of the model tree using classical regression techniques. Instead of using all attributes, only those which are referenced by tests or linear models somewhere in the subtree at this node are used. M5 compares the accuracy of a linear model with the accuracy of a subtree, limiting the attributes ensures a level playing field in which the two types of models use the same information.

- *Simplification of linear models.* After each linear model is obtained as above, it is simplified by eliminating attributes to minimise its estimated error. Although the attribute reduction causes the average residual to increase, it also reduced the multiplicative factor $(n+\nu)/(n-\nu)$, so the estimated error can decrease.

- *Pruning.* Each non-leaf node of the model tree is examined, starting near the bottom. M5 selects as the final model for this node either the simplified linear model discussed above or the model subtree, depending on which has the lowest estimated error. If the linear model is chosen, the subtree at this node is pruned to a leaf.

- *Smoothing.* When the value of a case is predicted by a model tree, the value given by the model at the appropriate leaf is adjusted to reflect the predicted values at nodes along the path from the root to that leaf. M5's smoothed predicted value is backed up from the leaf to the root as follows: (i) the predicted value at the leaf is the value computed by the model at that leaf. (ii) If the case follows branch S_i of subtree S, let n_i be the number of training cases at S_i, $PV(S_i)$ the predicted value at S_i and $M(S)$ the value given by the model at S. The predicted value backed up to S is given by:

  $$PV(S) = \frac{n_i \times PV(S_i) + k \times M(S)}{n_i + k}$$

  where k is a smoothing constant. Smoothing has the most effect on cases when the models along the pact predict very different values.

7.6 Software

In this section, the functions implemented in R and used for the analysis are reported. In particular, a function implementing the variogram deconvolution procedure of Goovaerts (2008), presented in Section 3.2.1, one implementing the linear regression step of the Area-to-Point Regression Kriging procedure, and finally a function implementing Area-to-Point Kriging using the Kyriakidis method (it is essentially a wrapper for *gstat* function *krige*). The full script used for the analysis is available on github at the link: `https://github.com/NiccoloTogni/Downscaling_SoilGrids_SMART_SED`.

```
variogram_deconvolution = function(coarse_raster, vg_type = "Sph", nnblocks = 4,
                  maxiter = 100, cutoff = "default", tol1 = 1e-2, tol2 = 1e-6) {

  ##### Variogram deconvolution procedure (Goovaerts, 2008)
  ##### This is a special version for regular grids (raster data)

  require(gstat)
  require(fields)

  ########### Arguments: ###########################################

  # coarse_raster = coarse resolution raster.
  # vg_type = type of variogram to be fitted.
  # nblocks = the regularized variogram is computed using R function vgmArea,
  # but only at short distances. nblocks is the number of adjacent blocks
  # that are considered near enough to require vgmArea.
  # maxiter = maximum number of iterations.
  # cutoff = cutoff. If = "default" it is set equal to half the raster extent.
  # tol1 = tolerance for Di/D0, if the value is lower the iterations stop.
  # tol2 = tolerance for abs(D_i-D_opt)/D_opt.

  #################################################################

  # Borders and extent of the raster map.
  xmin = coarse_raster@extent@xmin
  xmax = coarse_raster@extent@xmax
  ymin = coarse_raster@extent@ymin
  ymax = coarse_raster@extent@ymax
  coarse_res = res(coarse_raster)
  x_extent = xmax - xmin
  y_extent = ymax - ymin

  if (cutoff == "default") {
    cutoff = min(c(x_extent,y_extent))/2
    # half the raster extent by default.
  }

  # Change the name to a generic "z".
  names(coarse_raster) = ’z’

  # Convert the raster to a list of square polygons corresponding to the coarse pixels
  poly_coarse = rasterToPolygons(coarse_raster, fun=NULL, n=16, na.rm=TRUE,
             digits=12, dissolve=FALSE)

```

```
# 1. Compute empirical variogram on areal data and fit a model.
gamma_v_hat = variogram(z ~ 1, data = poly_coarse, cutoff=cutoff, width = min(coarse_res)/2)
gamma_v_exp = fit.variogram(gamma_v_hat, vgm(mean(tail(gamma_v_hat$gamma),10), vg_type,
                                         cutoff/2, min(gamma_v_hat$gamma)))

# 2. Initial variogram.
gamma_0 = gamma_v_exp

# 3. Variogram regularization using ’vgmArea’ for low distances, and Journel approximation
# for high distances. This procedure is specific for regular grids and allows to significantly
# speed up the algorithm, avoiding the repeated computation of block covariances,
# or the computation of block covariances at great distances, which can be approximated
# using the Journel formula (Journel, 1978).

new_xmax = xmin + nblocks*coarse_res[1]
new_ymin = ymax - nblocks*coarse_res[2]
inrange = crop(coarse_raster, extent(c(xmin,new_xmax,new_ymin,ymax)))
poly_ref = polygons(rasterToPolygons(inrange, fun=NULL, n=16, na.rm=TRUE, digits=12,
              dissolve=FALSE))
# Now poly_ref contains a square of blocks, given the regularity of the problem,
# we can drop the upper part of the square, since the covariance only depends
# on the distance of the blocks.
polygon_indexes = upper.tri(matrix(1:(length(poly_ref)),nrow = nblocks, ncol = nblocks),
           diag = TRUE)
polygon_indexes = as.vector(polygon_indexes)
poly_ref = poly_ref[polygon_indexes]
# Compute gamma^(v,v_h) for small lags.
gamma_A_0 = vgmArea(x = poly_ref[1],y = poly_ref, vgm = gamma_0, covariance = FALSE)
gamma_vv_0 = gamma_A_0[1] # gamma(v,v), unique since the grid is regular.
coords_ref = as.matrix(coordinates(poly_ref))
short_dist = as.vector( rdist(t(coords_ref[1,]),coords_ref[2:nrow(coords_ref),]) )
# Distances of neighboring blocks.
drop_duplicates = !duplicated(short_dist)
# Drop distance duplicates (the block covariance only depend on distance).
short_dist = short_dist[drop_duplicates]
ordered = order(short_dist)
short_dist = short_dist[ordered]
gamma_v_0 = as.vector( gamma_A_0[2:length(gamma_A_0)] )
gamma_v_0 = gamma_v_0[drop_duplicates]
gamma_v_0 = gamma_v_0[ordered] # Values of regularized variogram for small distances
# Add lags (greater distances).
dist_tail = seq(short_dist[length(short_dist)]+max(coarse_res),cutoff,max(coarse_res))
ref_dist = c(short_dist, dist_tail)
gamma_v_0_tail = variogramLine(gamma_0,dist_vector=dist_tail)$gamma
gamma_v_0 = c(gamma_v_0,gamma_v_0_tail)
gamma_v_0 = gamma_v_0 - gamma_vv_0 # Regularize

# 4. Quantify deviation.
D_0 = ( 1/length(ref_dist) ) *
  sum( abs(gamma_v_0-variogramLine(gamma_v_exp,dist_vector=ref_dist)$gamma)/
        variogramLine(gamma_v_exp,dist_vector=ref_dist)$gamma )

# 5. Define initial optimal variograms.
gamma_opt = gamma_0
gamma_v_opt = gamma_v_0
D_opt = D_0
rescaling_flag = 0

```

```
  # Start loop.
  for (i in 1:maxiter){
    print(paste0("iter: ", i))

    # 6. Compute experimental values for the new point support semivariogram
    # through a rescaling of the optimal point support model.
    if (!rescaling_flag){
      w = 1 + ( 1/(gamma_v_exp$psill[2]* (i^(1/2)) ) )*
        (variogramLine(gamma_v_exp, dist_vector=ref_dist)$gamma - gamma_v_opt)
    }
    else {
      w = 1+(w-1)/2
      rescaling_flag = 0
    }

    # Empirical values to which a variogram model must be fitted.
    gamma_hat_i_values = variogramLine(gamma_opt,dist_vector=ref_dist)$gamma * w
    gamma_hat_i = gamma_v_hat[1:length(ref_dist),]
    gamma_hat_i$np=rep(1,length(ref_dist))
    gamma_hat_i$dist = ref_dist
    gamma_hat_i$gamma = gamma_hat_i_values
    gamma_hat_i$dir.hor[is.na(gamma_hat_i$dir.hor)] = 0
    gamma_hat_i$dir.ver[is.na(gamma_hat_i$dir.ver)] = 0
    gamma_hat_i$id[is.na(gamma_hat_i$id)] = gamma_hat_i$id[1]

    # 7. Fit a new model.
    gamma_i = fit.variogram(object = gamma_hat_i, vgm(mean(tail(gamma_hat_i$gamma),2),
              vg_type, cutoff, min(gamma_hat_i$gamma)))

    # 8. Regularize gamma_i.
    gamma_A_i = vgmArea(x = poly_ref[1],y = poly_ref, vgm = gamma_i, covariance = FALSE)
    gamma_vv_i = gamma_A_i[1]
    gamma_v_i = as.vector( gamma_A_i[2:length(gamma_A_i)] )
    gamma_v_i = gamma_v_i[drop_duplicates]
    gamma_v_i = gamma_v_i[ordered]

    gamma_v_i_tail = variogramLine(gamma_i,dist_vector=dist_tail)$gamma
    gamma_v_i = c(gamma_v_i,gamma_v_i_tail)
    gamma_v_i = gamma_v_i - gamma_vv_i

    # 9. Compute D_i.
    D_i= ( 1/length(ref_dist) ) *
      sum( abs(gamma_v_i-variogramLine(gamma_v_exp,dist_vector=ref_dist)$gamma)/
             variogramLine(gamma_v_exp,dist_vector=ref_dist)$gamma )

    # 10. Stopping criteria.
    if ( (D_i/D_0 < tol1) || abs(D_i-D_opt)/D_opt < tol2) break
    if (D_i < D_opt) {
      gamma_opt = gamma_i
      gamma_v_opt = gamma_v_i
      D_opt = D_i
    }
    else {
      rescaling_flag = 1
    }
  }
  return(gamma_opt)
}
```

```
ATPlm = function(coarse_raster, covariates_stack){

  ### Area-to-Point Linear Regression.
  ### Remark: the target coarse resolution raster and the covariate maps must have
  ### the same coordinate system, and must overlap.
  ### The function returns an object containing a fine resolution map
  ### (matching the resolution of the covariates) of the fitted target variable,
  ### and an lm-type object containig the regression results.

  df = as.data.frame(coarse_raster[[1]])
  xnames = names(covariates_stack) # Covariate names.
  yname = names(coarse_raster)[1] # Name of target variable to be used in the formula.
  nx = length(xnames)
  # Upscale covariate maps for the regression.
  for (i in 1:nx){
    df[xnames[i]] = values( resample(covariates_stack[[i]],coarse_raster, method = "bilinear") )
  }
  f <- as.formula( paste(yname, "˜ .") )
  LM = lm(formula = f, data = df)
  predicted = predict(LM, as.data.frame(covariates_stack))
  map = covariates_stack[[1]]
  values(map) = predicted
  out = list(map = map, LM = LM)
  return(out)
}

ATPK = function(coarse_raster, fine_raster, variogram, npoints = 8, frac = 1.0, nsim = 0,
        beta = NA, nmax = 10){

  ### Area-to-Point kriging for the downscaling of coarse raster data.
  ### Converts the coarse pixels into spatial polygons and uses ATPK (Kyriakydis, 2004).
  ### The function also performs Block Sequential Gaussian Simulation (BSGS).
  ### Returns the downscaled map or a stack of the simulations.

  require(gstat)

  ### Arguments:
  # coarse_raster = coarse resolution raster to be downscaled.
  # fine_raster = fine resolution raster covering the extent of the coarse raster.
  #     It will be used to define the target resolution for the downscaling.
  # frac = fraction of coarse data to use when performing block sequential simulation
  # npoints = number of oints used to approximate the blocks (coarse pixels). Can be 4,8 or 16.
  # For the other parameters consult the documentation of 'krige' function - help(krige)

  # If simulations are required, cosnider only the specified fraction of areal data.
  if (nsim > 0){
    drop = 1.0-frac
    ndrops = floor(drop*length(coarse_raster))
    values(coarse_raster)[sample(1:length(coarse_raster),ndrops,replace = FALSE)] = NA
  }

  BlockMap = rasterToPolygons(coarse_raster, fun=NULL, n=npoints, na.rm=TRUE, digits=12,
                dissolve=FALSE)
  names(BlockMap) = "Z"
  d_new <- SpatialPoints(coordinates(fine_raster), proj4string = crs(BlockMap))
  if(is.na(beta)) {
    downscale = krige(Z˜1, BlockMap, newdata = d_new, model = variogram, nmax = nmax, nsim=nsim)
  } else {
    downscale = krige(Z~1, BlockMap, newdata = d_new, model = variogram, nmax = nmax,
             nsim=nsim, beta = beta)
  }
  if (nsim == 0) {
    downscaled_map = fine_raster
    values(downscaled_map) = downscale$var1.pred
    return(downscaled_map)
  } else {
    sim_data = downscale@data
    sims = stack(replicate(nsim,fine_raster))
    for (i in 1:nsim) values(sims[[i]]) = sim_data[,i]
    return(sims)
  }
}
```

Bibliography

[1] M. A. Martín, Y. Pachepsky, C. Baez, and M. Reyes. "On soil textural classifications and soil texture-based estimations". *Solid Earth Discussions* (Oct. 2017), pp. 1–14.

[2] A. Abbate. "Un modello numerico conservativo di erosione di versante". MA thesis. Politecnico di Milano, 2015.

[3] J. Aitchison. "The Statistical Analysis of Compositional Data". *Journal of The Royal Statistical Society, Series B* 44.2 (1982), pp. 139–177.

[4] J. Aitchison. *The Statistical Analysis of Compositional Data.* Chapman and Hall, 1986.

[5] T. Appelhans, E. Mwangomo, D. R. Hardy, A. Hemp, and T. Nauss. "Evaluating machine learning approaches for the interpolation of monthly air temperature at Mt. Kilimanjaro, Tanzania". *Spatial Statistics* 14 (2015), pp. 91 –113.

[6] L. Bonaventura, D. Brambilla, and L. Longoni. "An efficient and robust distributed model of soil erosion". unpublished. 2019.

[7] K. G. Van Den Boogaart and R. Tolosana-Delgado. ""compositions": A unified R package to analyze compositional data". *Computers & Geosciences* 34.4 (2008), pp. 320 –338.

[8] L. Breiman. *Classification and Regression Trees.* New York: Routledge, 1984.

[9] D. Butina and J. M. R. Gola. "Modeling Aqueous Solubility". *Journal of chemical information and computer sciences* 43 (May 2003), pp. 837–41.

[10] J.B. Collins and C.E. Woodcock. "Geostatistical Estimation of Resolution-Dependent Variance in Remotely Sensed Images". *Photogramm. Eng. Remote Sens.* 65 (Jan. 1995).

[11] N. Cressie. "Fitting Variogram Models by Weighted Least Squares". *J. Int. Assoc. Math. Geol.* 17 (July 1985), pp. 563–586.

[12] M. Delbari, P. Afrasiab, and W. Loiskandl. "Geostatistical Analysis of Soil Texture Fractions on the Field Scale". *Soil and Water Resources* 6 (2011), pp. 172–189.

[13] N. Dragicevic. "Model for erosion intensity and sediment production assessment based on erosion potential method modification". PhD thesis. 2016.

[14] N. Efthimiou, E. Lykoudi, D. Panagoulia, and C. Karavitis. "Assessment of soil susceptibility to erosion using the EPM and RUSLE models: the case of Venetikos river catchment". *Global NEST Journal* 18 (2016), pp. 164–179.

[15] J. J. Egozcue, J. L. Díaz–Barrero, and V. Pawlowsky-Glahn. "Hilbert Space of Probability Density Functions Based on Aitchison Geometry". *Acta Mathematica Sinica, English Series* 22 (July 2006), pp. 1175–1182.

[16] J. J. Egozcue and V. Pawlowsky-Glahn. "Geometric approach to statistical analysis on the simplex". *Stochastic Environmental Research and Risk Assessment* 15.5 (2001), pp. 384–398.

[17] J. J. Egozcue, V. Pawlowsky-Glahn, G. Mateu-Figueras, and C. Barceló-Vidal. "Isometric Logratio Transformations for Compositional Data Analysis". *Mathematical Geology* 35.3 (2003), pp. 279–300.

[18] Y. Fan, M. Clark, D. M. Lawrence, S. Swenson, L. E. Band, S. Brantley, P. Brooks, W. E. Dietrich, A. Flores, G. Grant, J. Kirchner, D. Mackay, J. Mcdonnell, P. Milly, P. L. Sullivan, C. Tague, H. Ajami, N. Chaney, A. Hartmann, and D. Yamazaki. "Hillslope Hydrology in Global Change Research and Earth System Modeling". *Water Resources Research* (Feb. 2019).

[19] P. Filzmoser, K. Hron, and C. Reimann. "Principal component analysis of compositional data with outliers". *Environmetrics* 20 (Sept. 2009), pp. 621 –632.

[20] I. Florinsky. "Accuracy of local topographic variables derived from digital elevation models". *International Journal of Geographical Information Science* 12 (Jan. 1998), pp. 47–61.

[21] P. Goovaerts. "A coherent geostatistical approach for combining choropleth map and field data in the spatial interpolation of soil properties". *European journal of soil science* 62 (June 2011), pp. 371–380.

[22] P. Goovaerts. "Combining Areal and Point Data in Geostatistical Interpolation: Applications to Soil Science and Medical Geography". *Mathematical geosciences* 42 (July 2010), pp. 535–554.

[23] P. Goovaerts. “Kriging and Semivariogram Deconvolution in Presence of Irregular Geographical Units”. *Mathematical geology* 40 (Feb. 2008), pp. 101–128.

[24] C.A. Gotway and L. Young. “Combining Incompatible Spatial Data”. *Journal of the American Statistical Association* 97 (Feb. 2002), pp. 632–648.

[25] M. Goulard and M. Voltz. “Linear coregionalization model: Tools for estimation and choice of cross-variogram matrix”. *Mathematical Geology* 24.3 (1992), pp. 269–286.

[26] A.J. Grass. *Sediment transport by waves and currents.* 1981.

[27] L. Le Gratiet. “Recursive co-kriging model for Design of Computer experiments with multiple levels of fidelity”. *International Journal for Uncertainty Quantification* 4 (Oct. 2012).

[28] J. K. Grudnicki. “Physically based numerical soil erosion model”. MA thesis. Politecnico di Milano, 2018.

[29] O. Grujic, A. Menafoglio, G. Yang, and J. Caers. “Cokriging for multivariate Hilbert space valued random fields: application to multi-fidelity computer code emulation”. *Stochastic Environmental Research and Risk Assessment* 32 (Nov. 2017).

[30] Benedikt Gräler, Edzer Pebesma, and Gerard Heuvelink. “Spatio-Temporal Interpolation using gstat”. *The R Journal* 8 (1 2016), pp. 204–218.

[31] W H. McNab. “Terrain Shape Index: Quantifying Effect of Minor Landforms on Tree Height”. *Forest Science* 35 (Mar. 1989), pp. 91–104.

[32] T. Hastie and R. Tibshirani. “Generalized Additive Models”. In: *Wiley StatsRef: Statistics Reference Online.* American Cancer Society, 2014.

[33] T. Hengl, G.B.M. Heuvelink, and D. G. Rossiter. “About regression-kriging: From equations to case studies”. *Computers & Geosciences* 33.10 (2007), pp. 1301 –1315.

[34] T. Hengl, J. Mendes de Jesus, G.B.M. Heuvelink, M. Ruiperez Gonzalez, and M. et al. Kilibarda. “SoilGrids250m: Global gridded soil information based on machine learning”. *PLOS ONE* 12.2 (2017), pp. 1–40.

[35] Tomislav. Hengl, 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. “SoilGrids1km — Global Soil Information Based on Automated Mapping”. *PLOS ONE* 9.8 (Aug. 2014), pp. 1–17.

[36] A.G. Journel and C.J. Huijbregts. *Mining geostatistics.* 1978.

[37] M.C. Kennedy and A. O'Hagan. "Predicting the Output from a Complex Computer Code When Fast Approximations Are Available". *Biometrika* 87.1 (2000), 1–13.

[38] J.-H. Kim. "Spurious correlation between ratios with a common divisor". *Statistics & Probability Letters* 44.4 (1999), pp. 383 –386.

[39] P. Kyriakidis. "A Geostatistical Framework For Area-To-Point Spatial Interpolation". *Geographical Analysis* 36 (Aug. 2004).

[40] P. Kyriakidis, P. Schneider, and M. F. Goodchild. "Fast Geostatistical Areal Interpolation". *7th International Conference on Geocomputation* (Aug. 2005).

[41] A. Li, X. Tan, W. Wu, H. Liu, and J. Zhu. "Predicting active-layer soil thickness using topographic variables at a small watershed scale". *PLOS ONE* 12.9 (2017), pp. 1–17.

[42] Y. Liu. "Geostatistical integration of linear coarse scale and fine scale data" (Jan. 2007).

[43] Y. Liu and A. G. Journel. "A package for geostatistical integration of coarse and fine scale data". *Computers & Geosciences* 35 (2009), pp. 527–547.

[44] B. P. Malone, A. B. McBratney, B. Minasny, and I. Wheeler. "A general method for downscaling earth resource information". *Computers & Geosciences* 41 (2012), pp. 119 –125.

[45] J.A. Martín-Fernández, R.A. Olea-Meneses, and V. Pawlowsky-Glahn. "Criteria to compare estimation methods of regionalized compositions". *Mathematical Geology* 33.8 (2001), pp. 889–909.

[46] A. McBratney, M.L. Mendonça Santos, and B. Minasny. "On digital soil mapping". *Geoderma* 117 (2003).

[47] A. Menafoglio, L. Guadagnini, and P. Secchi. "A kriging approach based on Aitchison geometry for the characterization of particle-size curves in heterogeneous aquifers". *Stochastic Environmental Research and Risk Assessment* 28 (2014), pp. 1835–1851.

[48] B. Minasny and A. Mcbratney. "Spatial prediction of soil properties using EBLUP with the Matérn covariance function". *Geoderma* 140 (2007), pp. 324–336.

[49] S.K. Mishra and P. Singh Vijay. *Soil Conservation Service Curve Number (SCS-CN) Methodology.* Vol. 42. Springer, 2003.

[50] I.O.A. Odeh, A.J. Todd, and J. Triantafilisl. "Spatial Prediction of Soil Particle-Size Fractions as Compositional Data". *Soil Science* 168.7 (2003), pp. 501–515.

[51] B. P. Malone, Q. Styc, B. Minasny, and A. McBratney. “Digital soil mapping of soil carbon at the farm scale: A spatial downscaling approach in consideration of measured and uncertain data”. *Geoderma* 290 (Mar. 2017), pp. 91–99.

[52] E. Pardo-Igúzquiza and P. M. Atkinson. “Modelling the semivariograms and cross-semivariograms required in downscaling cokriging by numerical convolution–deconvolution”. *Computers & Geosciences* 33.10 (2007), pp. 1273 –1284.

[53] N.W. Park. “Spatial Downscaling of TRMM Precipitation Using Geostatistics and Fine Scale Environmental Variables”. *Advances in Meteorology* 2013 (Dec. 2013).

[54] N.W. Park and D.H. Jang. “Comparison of Geostatistical Kriging Algorithms for Intertidal Surface Sediment Facies Mapping with Grain Size Data”. *The Scientific World Journal* 2014 (2014).

[55] L. Parussini, D. Venturi, P. Perdikaris, and G. Karniadakis. “Multifidelity Gaussian process regression for prediction of random fields”. *Journal of Computational Physics* 336 (May 2017), pp. 36–50.

[56] N.R. Patton, K.A. Lohse, S.E. Godsey, B.T. Crosby, and M.S. Seyfried. “Predicting Soil Thickness on Soil Mantled Hillslopes”. *Nature Communications* 9 (2018).

[57] V. Pawlowsky-Glahn and J. J. Egozcue. “BLU Estimators and Compositional Data”. *Mathematical Geology* 34 (Apr. 2002), pp. 259–274.

[58] V. Pawlowsky-Glahn, J. J. Egozcue, and R. Tolosana-Delgado. “Principal balances”. *4th International Workshop on Compositional Data Analysis (Codawork 2011)* (Aug. 2011).

[59] V. Pawlowsky-Glahn, J.J. Egozcue, R.A. Olea, and E. Pardo-Iguzquiza. “Cokriging of compositional balances including a dimension reduction and retrieval of original units”. *Journal of the Southern African Institute of Mining and Metallurgy* 115 (2015), pp. 59–72.

[60] E. J. Pebesma. “Multivariable geostatistics in S: the gstat package”. *Computers & Geosciences* 30 (2004), pp. 683–691.

[61] J.D. Pelletier and C. Rasmussen. “Geomorphically Based Predictive Mapping of Soil Thickness in Upland Watersheds”. *Water Resources Research* 45 (2009).

[62] A. Quarteroni, R. Sacco, and F. Saleri. *Numerical Mathematics (Texts in Applied Mathematics)*. Berlin, Heidelberg: Springer-Verlag, 2006.

[63] J. R. Quinlan. “Learning With Continuous Classes” (1992), pp. 343–348.

[64] R Development Core Team. "R: A Language and Environment for Statistical Computing" (2008). URL: `http://www.R-project.org`.

[65] M. R. Dobarco, T. G. Orton, D. Arrouays, B. Lemercier, J.B. Paroissien, C. Walter, and N. Saby. "Prediction of soil texture using descriptive statistics and area-to-point kriging in Region Centre (France)". *Geoderma Regional* 7 (Apr. 2016).

[66] N. Remy. "S-GeMS: The Stanford Geostatistical Modeling Software: A Tool for New Algorithms Development". In: *Geostatistics Banff 2004*. Ed. by Oy Leuangthong and Clayton V. Deutsch. Springer Netherlands, 2005, pp. 865–871.

[67] R. Rosso. *Caratterizzazione idrologica del regime di piena in Lombardia: bacini tributari del lago di Como mappatura dell'indice di assorbimento e del massimo volume specifico di ritenzione potenziale del terreno.* April 2004.

[68] P. Roudier, B.P. Malone, C.B. Hedley, B. Minasny, and A. McBratney. "Comparison of regression methods for spatial downscaling of soil organic carbon stocks maps". *Computers and Electronics in Agriculture* 142 (Nov. 2017), pp. 91–100.

[69] J. Sacks, W. Welch, T. J. Mitchell, and H. Wynn. "Design and analysis of computer experiments". *Statistical Science* 4 (Jan. 1989).

[70] T.J. Santner, B.J. Williams, and W.I. Notz. *The Design and Analysis of Computer Experiments*. Springer Verlag, 2003.

[71] W. Shang-Guan, T. Hengl, J. Mendes de Jesus, H. Yuan, and Y. Dai. "Mapping the global depth to bedrock for land surface modeling". *Journal of Advances in Modeling Earth Systems* 9.1 (2017), pp. 65–88.

[72] A. Soares. "Direct Sequential Simulation and Cosimulation". *Mathematical Geology* 33 (Nov. 2001), pp. 911–926.

[73] A. Talebi, P.A. Troc, and R. Uijlenhoet. "A steady-state analytical slope stability model for complex hillslopes". *Hydrological Processes* 22 (2008), 546–553.

[74] P.N. Truong and G. Heuvelink. "Bayesian Area-to-Point Kriging using Expert Knowledge as Informative Priors". *International Journal of Applied Earth Observation and Geoinformation* 30 (Apr. 2013), pp. 2291–.

[75] M. Vågen and G. Winowiecki. "Predicting the Spatial Distribution and Severity of Soil Erosion in the Global Tropics using Satellite Remote Sensing". *Remote Sensing* 11 (July 2019), p. 1800.

[76] W. Walker, P. E. Harremoës, J. Rotmans, J. P. Van Der Sluijs, M.B.A. Van Asselt, P. Janssen, and M.P. Krayer Von Kraus. “Defining Uncertainty: A Conceptual Basis for Uncertainty Management in Model-Based Decision Support”. *Integrated Assessment* 4 (Mar. 2003).

[77] D.J.J. Walvoort and J.J. De Gruijter. “Compositional Kriging: A Spatial Interpolation Method for Compositional Data”. *Mathematical Geology* 33.8 (2001), 951–966.

[78] G. Wang and S. Shan. “Review of Metamodeling Techniques in Support of Engineering Design Optimization”. *Journal of Mechanical Design - J MECH DESIGN* 129 (Apr. 2007).

[79] Z. Wang and W. Shi. “Mapping soil particle-size fractions: A comparison of Compositional Kriging and log-ratio Kriging”. *Journal of Hydrology* 546 (2017), pp. 526 –541.

[80] Z. Zhang, A. Ward, and J. Keller. “Determining the Porosity and Saturated Hydraulic Conductivity of Binary Mixtures”. 10 (Feb. 2011).