Spatial Analysis of Pedological and Environmental Features by means of Digital Soil Mapping

PhD Thesis

by Giuliano Langella

DISSPAPA Department, Federico II di Napoli University

Final Version, December 01, 2008

Preface

The thesis deals with the Digital Soil Mapping (**DSM**) techniques in case of limited data about soils, and with the high resolution 4D neurological analysis of precipitation data.

The rationale behind this focus of my PhD thesis has to be related to the extremely difficult task of the spatial inference of soils and rainfall in complex geomorphological settings, which is indeed the case for most if not all hilly and mountain Italian landscapes.

In order to address these topics, after a general introduction, the thesis is articulated into three separate parts. Soils (Part I) and precipitation (Part II) refer to two different case studies, hence they are developed into two separated parts respectively. Part III reports general conclusions and appendix.

Contents is articulated in eight chapters and three appendixes, whose brief overview is the following:

1. The thesis starts with a general introduction (Chapter 1), which describes the importance of handling high resolution and accurate soil and climate spatial information. Conventional soil survey is further investigated to point out its limitations, while highlighting the strength of digital soil mapping in gaining higher resolution and accuracy. Main goals are exposed (§ 1.1) to give the reader the big picture of the entire work.

2. Chapter 2 describe the materials used to address the spatial inference of pedological and climatic features. One can find the generalities of the Telese Valley study area (§ 2.1), information about the soil survey that bore the punctual soil database (§ 2.2.1), and the gathering of auxiliary maps (§ 2.2.2).

3. Chapter 3 describes how the former materials are elaborated to compute further continuous auxiliary information. There is the Digital Terrain Analysis (**DTA**, § 3.1), and the geomorphological segmentation of landscape in landform elements (§ 3.2).

4. Chapter 4 itemizes the methods of spatial analysis used including a brief theoretical introduction, while the specific settings for DSM in my subcase studies can be found in chapter 6.

5. An Explorative Data Analysis (**EDA**) is given in Chapter 5, which also introduces how the **EDASS** tool, specifically designed in VBA under Access, implements several tasks in very few time. This chapter is developed between the theoretical introduction of models of spatial inference (Chapter 4) and the specific settings used to address the spatial inference of selected features (Chapter 6). This position is justified by both the use of some models of spatial analysis within EDASS, and the use of EDASS explorative data analysis and stratification capabilities in delineating case studies.

6. Results about developed models are presented in Chapter 7, where the focus is put on the investigated pedological features, such as clay content and soil colour. The part is also dedicated to the discussion of techniques and procedures.

7. Chapter 8 is entirely devoted to the spatiotemporal rainfall analysis based on Artificial Neural Networks (**ANN**, § 4.3.1), with broadenings towards other disciplines such as Genetic Algorithms (**GA**) and Geostatistics (Indicator Kriging, § 4.2) for the sake of discovering the best path to precipitation interpolation in space-time domain. An autonomy is given to this study due also to the structure of this research theme, that was quite fully developed to be organized as a paper to be submitted.

8. In Appendix a through explanation and concern about tools, functions and scripts is provided. They are designed to speed up the execution of time consuming operations. Examples are the **MultiFieldAdder** tool (Appendix A) which is able to elaborate automatically a lot of raster layers in order to quickly get a matching table for spatial analysis. The **EDASS** tool (Appendix B) which is designed to facilitate patterns discovery from the pedological database (pedo-db), to allow the identification of structured variables in space (existence of autocorrelation), and to enlarge stimuli necessary to build up a mental model about the soilscape at hand. Further, a script named **ANNvsREGR** (Appendix C) is written in the MatLab M-language (with extension '.m') to run a semi automatic spatial analysis with both techniques of artificial neural networks and of multi linear regression, and to compare performance on out-of-sample data.

More satisfying motives are needed for clarifying in readers eye the unusual setting of contents. There are different levels of complexities each of which in turn earn a proper enlightenment. The common goal is to achieve a fluent reading — so I apologize when cross references cause jumping from one chapter to another for tackling the spatial analysis of a soil feature.

Firstly, there is the soil and precipitation separation/aggregation problem. *Separation* is twofold, since a soil database is only available for Telese valley study area, while neurocomputing requires a number of raingauges larger than the Telese gauged network. Hence to enlarge number of cases to a statistical minimum it was chosen a wider area, that is the Campania region raingauge network. There is no *aggregation* between precipitation and soils along this thesis because in precipitation analysis the investigated time support is not compatible with the spatial inference of soil attributes at hand. This means that it is not useful coping with decadal (10-days) singletons in producing spatial predictions of rainfall to support analysis of spatial arrangement of pedological attributes; nevertheless it was unnecessary incommoding neural nets to make predictions at very coarser time resolution. Indeed, in my very little experience, precipitation data show a nice autocorrelation, as pointed out by variography, when for instance one aggregates information to the average yearly precipitation over a study area. As much as time support is coarser the more linear statistical models like multivariate regression and geostatistics express with good accuracy and precision in the spatial analysis of even highly stochastic phenomena. But here I would like to solve the finer time support to deliver precipitation information for both physically based hydrologic models or empirical models at sub-catchment/farm scale that are outside the objectives of my thesis. Furthermore, in literature there is a lack of empirical rain models dealing with relative high space-time resolution (see § 8.1 and Tab. 8.1).

Secondly, there is the separation and consequent nomenclature problem of some chapters within the soil part (Part I). As a matter of fact, all auxiliary information used throughout the thesis should be put in chapter 2 as materials, but a distinguo is preferred to separate the preexistent information from data obtained after processing — so the adjective *postprocessing* given to chapter 3. Another split is made for results by distinguishing between *methodological* and *applied* results. Thus it is kept in readers mind the endeavour (EDASS, chapter 5, and Inference Setup, chapter 6) spent in outlining the arrangement of ***procedures*** for performing spatial analysis, giving at the same time an insight about the sequence of operations fulfilled.

At last, there is the *attribute* problem related to soil mapping. Despite the relative large number of soil attributes stored in the database — which carried towards EDASS implementation — only few selected features are

showed. This meets the obligation to avoid redundancy in the application of techniques, but more importantly a preliminary selection was made by choosing those functional and meaningful variables to soil management. Moreover the selection is at increasing cost of surveying: there are (i) the Munsell soil colour as a field morphological descriptor, (ii) the pH as one of the most simplest soil analytics, then (iii) the fraction of clay content, and finally (iv) the oxalate $Al + \frac{1}{2} \cdot Fe$ extractions.

Aknowledgements

I would like to thank all the persons that make it possible the realization of this PhD thesis. Firstly my gratitude goes to professor Fabio Terribile who built all important opportunities for my scientific growth. Than a particular appreciation is reserved to Angelo Basile for giving me above all an important contribution on the refinement of precipitation analysis. Also I cannot forget the person of Tom Hengl, who introduced myself in the field of Digital Soil Mapping by using the regression kriging technique.

Furthermore I say thank you to my work group for participating my effort spent during this years: Antonello Bonfante, Maurizio Buonanno, Roberto de Mascellis, Michela Iamarino, Piero Manna, Giacomo Mele, Luciana Minieri, Nadia Orefice and Simona Vingiani. They also provided great understanding even if I was only present with my body at work ...

Finally I reserve a special thank to my family since they worked hard in my place when I have had to end up my PhD thesis.

I also remember all the people I omitted here, because this work is only possible with the support of many *invisible* people.

THANK YOU!

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Soil remains one of the most important, yet most abused, natural resources on the planet, indeed a responsible management of soil resources plays a critical role in the survival and prosperity of many nations around the world [White, 2005].

Soil is an unconsolidated or loose combination of inorganic and organic materials. The inorganic components are principally the products of rocks and minerals that have been gradually broken down by weather, chemical action, and other natural processes. The organic materials are composed of debris from plants and from the decomposition of the many life forms that inhabit the soil. It also contains air and water which all together with the solid state forms the complex three-phase system called soil.

The understanding of soil features, properties and behaviour enable sustainable land management. During the last decade, in the world and especially in Europe (e.g. EU Soil Thematic Strategy), a growing attention has been focused on the soil resource, to understand both the internal mechanisms that define its nature and its relationship with the other environmental factors. Together with site studies, large areas of Europe have been mapped at various scales, also within the frame of regional and interregional projects. The released soil maps are suitable for a large range of utilizations, and they are commonly used to plan sustainable land uses. In this way soil maps are strategic tools for land planning at different levels (farm, municipality, province, district, region, nation).

Conventional methodologies used for soil mapping are standardized since long time and, for what is concerning the semi detailed and detailed maps, soil variability is expressed by the soil series. They are the lowest class of the most commonly used soil classification system (Soil Survey Staff, 1993). A soil series is a group of soils, or polypedons, having similar horizons and a very limited variability in their properties, formed from the same parent

material.

Conventional soil survey is also based on the soil-lanscape concept [Hudson, 1992]. Through field investigation and photo interpretation soil mappers establish their mental model about the actual soil-landscape model over the area; therefore they are able to formalize the spatial arrangement of soil series in the different landscape units.

However, standard soil surveys were not designed to provide the high-resolution soil information required by some environmental modeling [Band and Moore, 1995]. Indeed the detail of conventional soil maps are not compatible with other landscape data derived from more detailed digital terrain analysis and remote sensing techniques.

Two major limitations prevent a soil scientist to conduct an accurate and efficient survey, the polygon-based mapping practice and the manual production process. Under the polygon-based mapping practice two major limitations occurs, the generalization of soils in the geographic domain and the generalization of soils in the parameter domain.

The first problem limits the minimum size of a soil body to be delineated as a polygon on map; this force soil mappers filtering the small soil bodies and representing only the dominant ones.

The second limitation arises from the central concepts of soils, for which natural soil bodies are assigned to a set of prescribed soil classes (e.g. Soil Taxonomy) according to a Boolean Classification. This means that a natural soil body is assigned to one and only one *platonic* soil category. This setting force soil spatial variability to be depicted by a step function with constant values within polygons and abrupt soil variation at boundaries.

There are several limitation associated with the manual map production process, but above all the stereoscopic photo interpretation is a subjective, time-consuming and error-prone process.

The conventional approach organizes the soil spatial variability following a *discontinuous and deterministic scheme*. In such a scheme, the change of the soil forming factors (clorpt) [Jenny, 1941] in the landscape units corresponds to the (discontinuous) change between soil series (phases of series).

Soils may not be observed everywhere due to time and money obligations. Therefore, the main objectives of a soil survey are to predict the distribution of soil characteristics and properties influencing the use and management and to transfer the information to land users [Edmonds et al., 1985a]. As a consequence of these obligations and objectives, the large number of studies made to ascertain the variability of officially established soil series have often noticed a larger proportion of included soils than what reported and published in soil reports.

Attention on spatial variability of soil series started to be focused in

the second half of the fifties, when different authors started studying the physical and chemical variability of soil series established in the USA [Jacob and Klute, 1956, Hammond et al., 1958, Thornburn and Larsen, 1959]. From then on, there was a growing interest on the subject, even if until the middle of the seventies the studies were always carried out in the USA. Aljibury and Evans [1961] reported the remarkable variability of water retention and bulk density measurements for two soil series previously considered relatively uniform.

Other studies made in those years [Nelson and McCracken, 1962, Andrew and Stearns, 1963, Mader, 1963, Wilding et al., 1964] remarked the large variability of soil series from the USA. Powel and Springer [1965] studied the composition of three soil series using rectilinear transects and found from 17 to 30% inclusions in the first soil series, up to 30% inclusions in the second soil series, and up to 40% inclusions in the third soils series. Anyhow, the authors highlighted that most soil inclusions were not significantly different from the dominant soils (<15%) for what was concerning their similar interpretation. Wilding et al. [1965], in a study on the variation of soil morphology in three map units in Ohio, found that all map units included 30% or more inclusions of other soils and that only 42% of the 240 studied pedons were correctly classified at series level. Another study made on 48 properties of 220 pedons belonging to 6 soils series from Ohio highlighted that only 37% of the studied pedons were correctly classified [McCormack and Wilding, 1969].

Similar studies made during the seventies concluded that impurity in soil map units, especially at short distance, may reach up to 50% [Beckett and Webster, 1971] and that the problem was present also in Europe. Indeed, Bascomb and Jarvis [1976] found, for a soil series in southern England, that 60% of the studied pedons were satisfying the soil series definition and that soil physical properties were more uniform than soil chemical properties. During the seventies, Cassel and Bauer [1975] and Baker [1978] studied the variability of some physical properties, such as bulk density, water retention at 15 bar tension, and hydraulic conductivity. These properties, even if very important for applications, were and are very rarely measured during routine soil surveys.

Studies made in the last years have reached the same conclusions of past studies, highlighting the current relevance of the problem. Edmonds et al. [1985b], in a study on the variability of 3 soil map units, corresponding to 3 different soil series, found that similar pedons were placed in different taxa, while different pedons were placed in the same taxon. Nettleton et al. [1991] studied about 1500 pedons from soil series recognised in the field and found that 75% of those pedons were taxajuncts of the established soil series. The same authors, working on 56 pedons belonging to 6 soil series, found that

59% of them were out of the taxonomic limits of the series.

Some authors [McBratney et al., 2000] suggest that in order to approach such difficult problems it is required to recognise that the soil continuum cannot be described and analysed using deterministic scheme based on discontinuities (standard soil survey applied to soil series).
The mechanistic Jenny's model [Jenny, 1941] explaining qualitatively soil development is condensed in the equation:

$$\mathbf{S} = \mathbf{f}(\mathbf{cl}, \mathbf{o}, \mathbf{r}, \mathbf{p}, \mathbf{t}, \ldots) \tag{1.1}$$

Much of early work (for a review see Yaalon [1975]) contemplate at most mono-factorial qualitative or semi-quantitative description of relationships between soil-forming factors and soil state. The master aim was to understand and not to predict soil from factors. Therefore climofunctions (e.g. Jones [1973]), organofunctions (e.g. Noy-Meir [1974]), topofunctions (e.g. Anderson and Furley [1975]), lithofunctions, and chronofunctions (e.g. Hay [1960]) have been proposed, disregarding the role of interactions between soil-forming factors themselves (the dots of Eq. 1.1), whose recognition could conversely be a useful work so that more detailed spatial patterns on soils are available.

McKenzie and Ryan [1999] proposed a modified version of Jenny's functional factorial model in which soil is fitted in space domain by means of an environmental correlation model of the form:

$$\mathbf{S} = \mathbf{f}(\mathbf{Cl}, \mathbf{T}, \mathbf{PM}, \mathbf{M}, \ldots) \tag{1.2}$$

where **S** is the soil property or land quality, **Cl**, **T** and **PM** are explanatory variables from spatial layers of climate, terrain and parent material respectively, while M represents other miscellaneous auxiliary layers from multispectral sensing, land management, etc.

In 2003 McBratney et al. proposed a review of approaches in making digital soil maps using GIS co-variable layers, and proposed a generic quantitative framework into which various methods of spatial inference can take place. They replaced the Jenny's *clorpt* model with the following one in which factors of soil formation are treated quantitatively and an objective high resolution numerical handling of soil spatial variability is given:

$$\mathbf{S} = \mathbf{f}\left(\mathbf{s}, \mathbf{c}, \mathbf{o}, \mathbf{r}, \mathbf{p}, \mathbf{a}, \mathbf{n}\right) \tag{1.3}$$

where
S: is a soil class or attribute;
s: refers to soil punctual data;

c: climate;
o: organisms;
r: (relief) topography and land surface parameters;
p: parent material;
a: age factor;
n: spatial position on ground;

The **f(.)** function of Eq. 1.3 is an empirical quantitative description linking **S** to auxiliary *scorpan* factors. It may assume as many forms as soil science applications borrow models from statistical techniques. Therefrom there are generalized linear models (GLMs, see Lane [2002]) such as linear regression and logistic regression, tree models (regression and classification trees), geostatistics, fuzzy inference systems, artificial neural networks, genetic algorithms, and knowledge-based systems amongst others. In chapter 4 a group of these statistical techniques are briefly introduced and then explored for the sake of modeling the soil spatial variability of Telese Valley case study.

As a result of all these considerations, the soil spatial variability and its relationship with the usefulness of soil maps is a subject of crucial relevance for the proper management of agri-forestry environments and, more generally, for land planning. Climate is an interlinked topic that largely affects landscape performance. In fact soil data often need to be integrated with climatic data in order to address many practical landscape management issues (e.g. primary production, irrigation, etc.).

Unfortunately the spatial distribution of climatic data, and especially of high resolution rainfall data, is largely affected by many causes of uncertainties.

In such a framework, I believe that it is of great concern to carry out research on these topics of soil and rainfall spatial distribution aiming to set up and test new methodologies enabling a sustainable analysis and representation of soil-climate spatial variability.

1.1 Aims

Environmental modeling is functional to explain the state and/or the dynamic of natural systems, or to make predictions at unknown space-time elements. In soil science Scull et al. [2003] refers to predictive soil mapping.

Although soils are anisotropic both vertically and laterally [Park and Vlek, 2002], the research aim is focused on the lateral variation of soils over the landscape. This variation is here explored by means of different tech-

niques of spatial inference: (i) the generalized linear models (GLMs) with the technique of multiple regression, (ii) geostatistical models with ordinary kriging, regression kriging, cokriging, indicator kriging, and (iii) the group of soft computing (also known as artificial intelligence) with artificial neural networks, genetic algorithms and fuzzy logic.

According to Skidmore [2002] these models can be grouped on the basis of *logic* in inductive and deductive, and on the basis of *processing method* in deterministic and stochastic. Deterministic models can be further divided into empirical, knowledge, and process based models. For instance an artificial neural network is an inductive-stochastic model, a fuzzy inference system is an inductive-knowledge based model, while regression, geostatistics, and genetic algorithms produce inductive-empirical models.

The big picture of this thesis is twofold, inasmuch a particular soil attribute is put in a spatial framework and is analysed by means of more available methods; nevertheless a particular technique is explored and exploited in order to understand its fitness in making accurate predictions about a fixed environmental feature. The former task is the especially objective of the spatial analysis of lateral variability of soil features over the landscape (Part I). The latter one is addressed in the 4D space-time analysis of precipitation data by means of stacked artificial neural networks (Part II).

Part I

Soil Features Mapping

Chapter 2

Materials

2.1 Study Area

The study area called Telese valley is of great strategic interest for both wine and olive oil production. Then it is not surprising that here the knowledge of the soil distribution is crucial to classify the landscape in pedoclimatic homogeneous environments, called terroir, to be considered essential for the high quality cultivation of vineyards and olive trees. In addition to the agronomic features, Telese valley has a large portions of the carbonate relief covered by chestnut forests at medium altitudes and beech forests at the summit (about 1000 m asl). Telese valley have a complex geological, geomorphological and soil setting; possibly it is one of the most complex areas throughout the Campania and probably throughout the south Italy. In this area very ancient soils (paleosols) coexist with very recent soils, with soil chemical and physical properties very different and very complex relations between landforms and soil types.

This area was chosen also because in the years 1996 and 2006 there has been an extensive soil study producing the Soil map of Telese valley, which is the first systematic, comprehensive and scientific knowledge contribution on soils in this area.

There are the particularly fertile volcanic soils (Andosols), that support ecosystems between the most productive of both Italy and possibly of Europe [Di Gennaro et al., 1995, Lulli, 1990]. These soils have unique qualities and behaviors [Maeda et al., 1977, Quantin, 1990] which, overall, confers a remarkable sensitivity and fragility in the considered landscape.

Geography Telsese valley study area coincides with the low valley of the river Calore, bounded on the south by massive Taburno-Camposauro, and

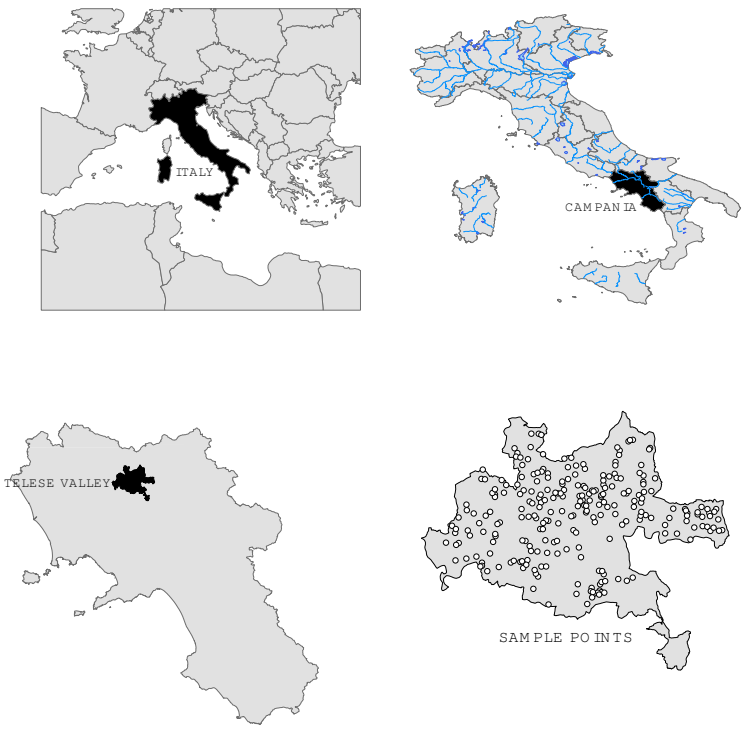


Figure 2.1: Location of the project area with soil profile sample points

Northeast and Northwest, respectively, from the mountain relief of M. Croce - M. Ciesco e Colle Sella (M. Monaco di Gioia) which are the extreme southern extension of the Matese massif.

The area has an extension of 20.000 hectares and lies in the municipalities of Amorosi, Telese, S. Salvatore Telesino, S. Lorenzello, Guardia Sanframondi, Solopaca, S. Lorenzo Maggiore, S. Lupo, Ponte, Paupisi and part of Vitulano.

The minimum altitude are naturally situated at the beds of the main rivers, in particular the Calore River flowing at 35 m asl when connecting into the Volturno reaches 95 m asl at Ponte. The lower energy mountainous relieves have an altitude range between 807 m of Monte Croce and 893 m s.l.m. of Monte Ciesco, 957 m s.l.m. Colle Sella, the dominant relief area is the Camposauro that stands up to 1390 m asl.

The hydrography of study area is mainly related to the river Calore, which lies in the study, the lower valley, with an East-West trend, the area also includes part of of the Volturno River and Torrente Titerno, having respectively North-South and East-West trend.

Geology and Geomorfology The geomorphology highlights very different environments in terms of their genesis and evolution. The main morphological-structural feature being the dominant depression of tectonic origin elongated in east-west direction where the Calore river flows.

The inter-mountain plain mainly develop in the right side of the river, where it is characterized by terraces both recent (Holocene) and in progress, with gradients lower than 5 m, and ancient terraces (upper-middle Pleistocene) located at 150 m asl, these are set on gravel and polygenic deposits including interlayered sand having a yellow-orange colours. These terraces have increasing altitude moving towards Ponte and are broken up by tectonic movement and by the development of the river networks. The tributary streams of the River Calore are generally characterized by short paths ad frequent ramifications. The resulting morphology depicts narrow and elongated plains perpendicular to the river, isolated from each other by narrow V impluvium.

In the left side of the river Calore, the ancient terraces are reduced to residual limbs that are found mainly in the south of Ponte and in the area of the cemetery of Solopaca. The morphology of the foothills detrital band mainly consists of *brecce* cemented and reddened of Mindel genesis. This foothills landscape presents bands having different slopes, reflecting mainly the type of detrital deposit and its degree of hardening. The central area of the detrital deposit is characterized by pseudo-karst phenomena with *doline subdetritiche*.

Areas having volcanic deposits in primary deposition occur at the right side of the Calore river, in the western sector of the valley. The Campanian Ignimbrite (*Tufo Grigio Campano*) occur at the left of the T. Titerno and in central and northern basin of Castelvenere. The volcanic deposits appear only rarely in the eastern sector. The hills are constituted by 3 portions: the first, elongated in east-west direction, mainly arenaceous-marl, divided by Castelvenere T. Titerno, and the other, with NW-SE direction, is set on Mesozoic limestones of the Matese - M. Maggiore, and is located immediately north of settlement Telese Terme; the third, to the east-west trend, shows a calcarenitc lithology as in the villages of Guardia Sanframondi, San Lorenzo Maggiore and San Lupo. In the eastern sector, Glacis flaps are set mainly on *argille varicolori* with poor physical-mechanical quality, which explains the greater tendency of this area to collapse and movement of surface.

The higher slopes and acclivi M. Croce and M. Ciesco are defined by four major faults in NW-SE trend. They are formed by calcareniti Paleoceniche and Eoceniche facies of the Transition (*Flysch Red*) and are surrounded by Miocene flysch.

The massive carbonate of Mount Camposauro is part of the Taburno-

Camposauro relieves, which is made up of different tectonic units. It is conceivable that after tectonic stages which gave the appearance today to the region, the establishment of the new cycle is karst been somewhat delayed by the presence of heat, the lithologic mioceniche waterproof, and has been arrested and rejuvenation to the changing climatic conditions Quaternary and for the deposition of pyroclastic deposited transported by wind.

Agriculture The land use has evolved in recent decades, with the growth of vineyards at the expense of industrial crops such as tobacco. Another important feature is the cultivation of olive trees that can better use the environmental and climate resources and ensure an higher income than pasture, scrub and forest.

Crops are mainly distributed in the town of Amorosi which account for 75% of SAU, to a lesser extent in S. Salvatore Telesino (45%) and Ponte (38%). The fruit trees (mainly apple) are poorly represented with 14% of SAU in the town of Telese, 10% in the municipality of S. Salvatore Telesino and 5% in the municipality of S. Lorenzello.

2.2 Geo-Database

2.2.1 Punctual Soil Database

Telese valley soil survey and mapping, which produced the dataset employed in this thesis, has been performed with the following stages which are typical of the standard soil mapping:

1. Analysis of existing mapping documents.
2. Photo interpretation.
3. Soil survey.
4. Chemical and physical analysis of soils.
5. Classification, correlation and mapping of soils.
6. Soil report.

Soil survey was carried out by means of both hand auger observations and opening of soil profiles in the different soil-landscape units previously identified. The soil profiles and drillings have been described in accordance with the Gardin et al. [1995] methodology and were sampled for chemical and

physical analysis. The studied soils were finally classified according to Soil Taxonomy and returned on a map 1:25.000.

The availability of specialized determinations has, for almost all soils collected, enable to retail the family taxonomic level.

The chemical and physical analysis listed in the database were performed on fine earth (< 2 mm). Analysis were conducted in accordance with the MIPAF et al. [2000] methods with the exception of particle size that was made by the method of pipette at pH 9.5 to reduce the problems of dispersion of volcanic soils due to their charge variable properties [Mizota and Van Reeuwijk, 1989]. This method, while providing better data (greater extraction of clay) of conventional treatment with the sodium hexametaphosphate, however, underestimate the fine fractions. Then as Mizota and Van Reeuwijk [1989] reported, *the particle size analysis on volcanic soils must always be evaluated with suspicion*, and it should be used especially on a relative scale comparing horizons within a profile or between different profiles.

The pH is measured in a soil-water suspension 1/2.5; the content in organic carbon was determined by oxidation with potassium dichromate, the cation exchange capacity (CEC) and the saturation bases were determined with $BaCl_2$, and by analysis with ICP-AES (model Liberty 150, Varian). The extraction of Fe, Al and Si in oxalate (Fe_{ox}, Al_{ox}, Si_{ox}) were performed using the method of Schwertmann [1964] and the content in Fe, Al and Si were determined by ICP-AES. Analysis of water retention was done by measuring the water content, by weight, the potential values of 1.500 KPa.

2.2.2 Continuous Auxiliary Database

The basic approach behind to correlate NDVI with soil colour is that, because NDVI is often considered primarily a function of climate, terrain, vegetation/ecosystem, and soil variables. Therefore, I assumed NDVI is a function of soil colour and surface temperature at specific location. Here, I have taken surface temperature and soil colour together, because surface temperature is greatly affected by soil colour.

From Bishop and McBratney 2001/pag149-150: General description of Auxiliary data
Due to high cost and time-consuming nature of soil sampling, research in developing methods for the creation of soil maps from sparse soil data is becoming increasingly important. In the past 20 years, the development of prediction methods that use cheap secondary information to spatially extend sparse and expensive soil measurements has been a sharpening focus of research e.g. Odeh et al., 1994; Gessler et al., 1995).

2.2.2.1 DEM

The digital elevation model (DEM), an important source of information, is usually used to express a topographic surface in three dimensions and to imitate essential natural geography. This study analyzed digital elevation data sources and their structure, the arithmetic of terrain attribute extraction from DEM (chapter 3) and its applications (chapter 6). DEM is also

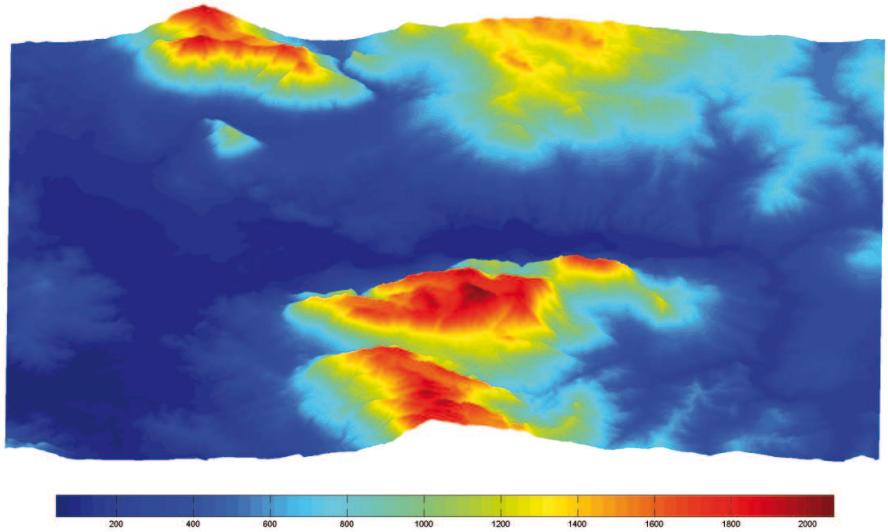


Figure 2.2: Digital elevation model for Telese valley

used to account for the spatial analysis of high resolution precipitation data (chapter 8), in order to make time-series predictions at unknown points, or to make multi–temporal spatial maps of precipitation of a large area (Campania region).

2.2.2.2 Vegetation cover

The municipalities most affected in the presence of forests are Solopaca and Vitulano, with the forests of the Camposauro relieves. Here on the slopes at medium-low altitude (below 1.000 m) the mixed forest is particularly rich in arboreal species. This strata is usually composed of black hornbeam (Ostrya carpinifolia), *orniello* (Fraxinus ornus), *Carpinello* (Carpinus orientalis), often oak (Quercus pubescens), *cerro* (Quercus cerris), and chestnut (Castanea sativa). At altitude of about 1.000–1.100 m, where climate becomes cool and moist, beech wood is present. Beech seems to prefer deep soils, fresh and fertile. The undulating summit plains tends to lead towards the development

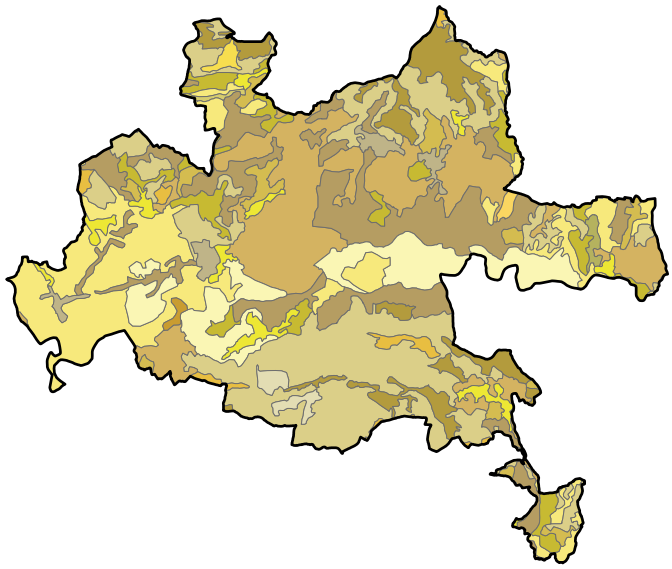


Figure 2.3: Corine Land Cover for Telese valley

of pastures.

On the slopes of Monte Croce and Monte Ciesco in Guardia Sanframondi and S. Lupo, forests are *cerro* (Quercus cerris) and oak (Quercus pubescens) in various stages of degradation.

On the Acero Mount and on the slopes of Colle Sella afforestation are conifers with a recovery of natural vegetation. Particularly interesting the presence of holm oak (Quercus ilex) on the side of Mount Acero.

2.2.2.3 Pedo-landscape units

The survey of Telese valley soils led to the recognition of three major landscape systems (intermountain plains, Preappennine hills and Preappennine mountains), further divided into seven landscape subsystems, characterized by combinations of soil forming factors differentiated and reported in Fig. 2.4. They are:

- Alluvial plain (**PIM**)
- Areas of complex genesis (**AGG**)
- Ancient river terraces (**TET**)

- Foothill Glacis (**GLA**)
- Hills (**CAP**)
- Detrital-colluvium areas (**RAC**)
- Mountain relieves (**MAP**)

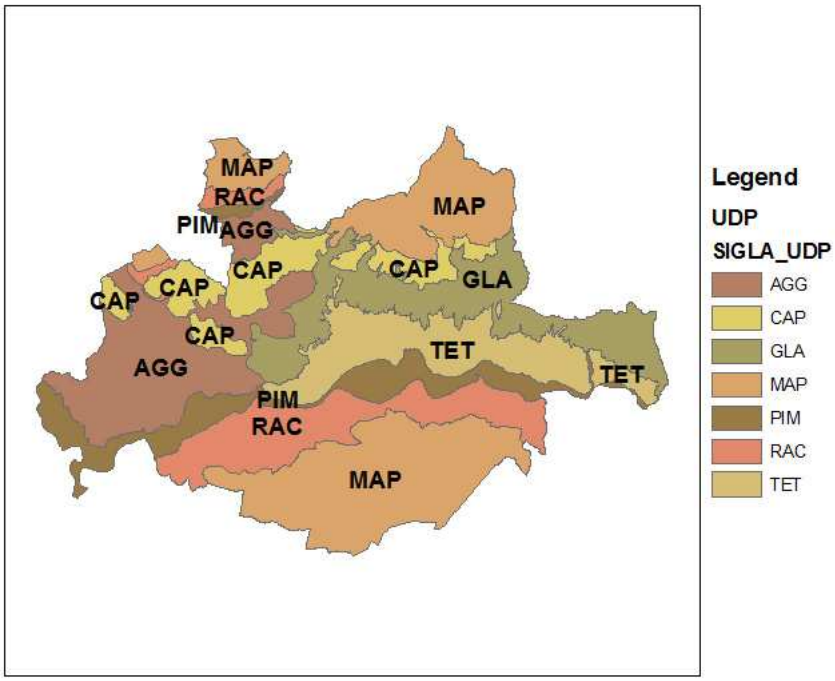


Figure 2.4: Pedo-landscape units for Telese valley

2.3 Workstation Configuration

The delivery of high resolution digital soil maps requires better and better performing workstations. The handling of geo-spatial data made of a lot of georeferenced large matrices to be manipulated, visualised, and processed, the storing ability of complex geo-databases, and the high demand from different software, such as Access, ArcGIS or MatLab, for executing intricate queries in SQL language call for up to date desktop computers.

The computer hardware configuration used to implement the tasks presented in this thesis is illustrated in the following box:

CPU	Intel Pentium Dual Core 3.2 GB
RAM	2 GB
HDD	1x100 GB, 1x200 GB
Graphic Card	ATI Radeon X1050 PCI

The operative system is a Windows XP 64 bits with the following programs installed on:

Microsoft Office: Excel, Access;

GIS: ArcGIS, SAGA, ILWIS;

Statistics: SPSS, R, GSTAT, ISATIS;

Programming: Visual Basic 6;

Miscellaneous: MatLab;

Typesetting: TeXnicCenter.

Chapter 3

Postprocessing Materials

The identification of the spatial and temporal scales over which soil-forming processes operates in a landscape and of the factors that are believed to influence these processes are key issues in soil mapping.

Surface morphology exerts an important role as it governs how the most important driving forces in soil genesis and namely mass (water, mineral nutrients) and energy (light, heat), distributes and flows on the land surface. Therefore quantitative methods of terrain analysis can be used to predict soil properties in space domain by means of deterministic or stochastic predictive relationships which put into relation soil features with auxiliary maps.

I make use of two frameworks of landscape analysis [Zhou et al., 2008]. The continuous framework is represented by the Digital Terrain Analysis, which is based upon an element-wise pixel-by-pixel attribution of values as a function of surrounding values. The discontinuous framework is developed in the task of landform segmentation, which semi-automatically classify landscape into geomorphological entities (facets).

3.1 Digital Terrain Analysis

The process of quantitative description and derivation of topographic attributes from digital elevation data is known as Digital Terrain Analysis (**DTA**).

It can be classified in different ways according to the neighbourhood extension from which to pick surrounding values, or to the source of the computed attributes, or further to the purpose of the analysis. Hence, in order of appearance, tools of terrain analysis can operate on local (3x3, or 5x5 window) or on extended neighbourhood (regional, global); can compute terrain parameters directly from the DEM (the so-called primary attributes)

or involving combinations of primary attributes (secondary attributes); or can compute ecological, geomorphological, hydrological, climatic and so on parameters.

Digital terrain parameters are here grouped according to Wilson and Gallant [2000] in primary and secondary parameters.

Primary attributes describe the geomorphometry itself of a landscape and are calculated from directional first and second order derivatives of topography. They include *slope* (rate of change of elevation along the direction of steepest descent), *aspect* (orientation of the facet of steepest descent), *plan curvature* (rate of change of aspect along a contour line), *profile curvature* (rate of change of slope along a flow line), and *upslope catchment area* (flow contributing area above a certain length of contour) amongst others.

Secondary terrain attributes quantify the effect of the topographic surface on the specific vector field at hand [Shary et al., 2002], such as on the gravitational field in redistributing water in landscape, or on the solar irradiance field in modifying the amount of solar radiation received at surface. They are calculated from the combination of two or more primary attributes, and include the *topographic wetness index* (TPI), the *stream power index* (SPI) and the *sediment transport index* (STI) [Moore et al., 1993].

A wide variety of algorithms for DEM creation and for calculation of the terrain attributes are available, and it is necessary to list the specific algorithms used, to make the research design reproducible. Primary and secondary terrain attributes are given in Table 5.1.

3.2 Fuzzy Landform Segmentation (FLFS)

Soil genesis at a given point in the landscape is the result of the action and interaction of soil forming factors over time [Simonson, 1959]. Geographical allocation of soil taxonomic units in the landscape relies upon the spatial arrangement of type and intensity of the soil forming processes, which in turn and to a lesser degree are controlled by landform shape and position itself.

The Milne's catena concept Milne [1935] was one of the first soil-landscape models in which the relationship between soil attributes and landscape position is pointed out. Up to nowadays several soil-landscape models have been proposed and quite all encapsulate the same premise of Milne's catena, scilicet mass and energy movement are the most important driving forces in landscape evolution and soil genesis. Therefore much endeavour was spent in delineating flow paths and in segmenting areas of low or high flow, accumulation or drainage patterns using the land surface parameters (LSP) such

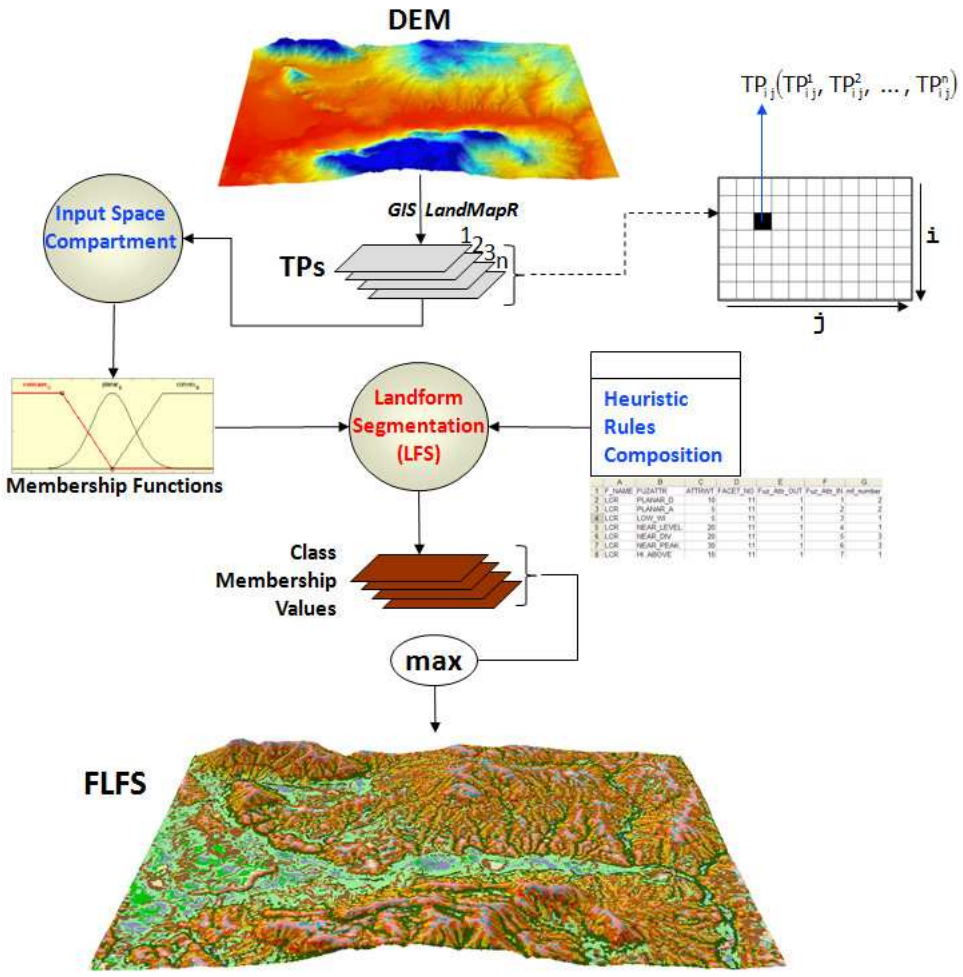


Figure 3.1: Diagram for fuzzy landform segmentation

as those described in § 3.1.

Any scheme of subdivision is arbitrary in principle, but the usefulness of such a model should be evaluated in terms of its ability to explain observed spatial variation in soils, as I make for instance for soil texture (§ 7.1) in Telese Valley study area.

Here is adopted the landform segmentation approach proposed by MacMillan et al. [2000] and successively implemented by MacMillan [2003] in a computer toolkit, LandMapR®. The landform facet segmentation is an heuristic fuzzy rule based procedure that allow continuous (opposed to crisp or hard) classification of the input land surface parameters and afterward of the output landform elements into which landscape is supposed to be divided.
Fuzzy fundamentals will be given in § 4.3.2.

The procedure (Fig. 3.1) is articulated in more steps as summarized below (for more details refers to MacMillan [2003]):

1. application of a low-pass filter for smoothing the raw DEM, with a 3x3 and 5x5 two-dimensional moving average window (under ArcGIS or MatLab);

2. computation of a depressionless DEM by removing pits, and of the upslope contributing area (with the FlowMapR component of LandMapR);

3. selection of a threshold for the upslope area to define the most likely locations for stream channels and ridges (ArcGIS);

4. calculation of the input land surface parameters *slope*, *plan* and *profile curvature*, *topographic wetness index* (as per Quinn et al. [1991]), and of the three relative positional attributes (as per MacMillan et al. [2000]) *PctZ2Str*[1], *PctZ2Pit*[2], and *Z2Pit*[3] (with the FormMapR component);

5. fuzzification of the inputs, that is the conversion of the 7 terrain parameters into fuzzy landform attributes through a set of predefined membership functions (MatLab);

6. conversion of the fuzzy land surface parameters into 15 fuzzy landform facets by means of a fuzzy inference system prepared for the purpose in MatLab, following the specifics of MacMillan et al. [2000] about the joint membership functions and the heuristic weights;

7. extraction for each pixel of a single landform facet amongst 15 as the landform element with highest joint membership function (MatLab).

Steps from 4 to 7 are executed in MatLab employing two private sequential functions which (i) extract through SQL statements the terrain attributes from the Fox Pro tables created by LandMapR, and (ii) fulfill any action from the fuzzification of the input parameters till the extraction for each pixel of the landform facet from the set of the 15 predefined classes.

[1]Percent pixel height relative to nearest stream and divide
[2]Percent pixel height relative to local pits and peaks
[3]Absolute pixel height above the local pit cell

Chapter 4

Methods of Spatial Inference

Merely a very brief summary on some theoretical concepts is given here, sending to thorough and exhaustive material cited in case of need along the chapter.

4.1 Multiple Linear Regression

Regression analysis [Hastie et al., 2001] models the relationship between a target variable, also called response or dependent variable, and one or more explanatory variables (the predictor or independent variables) by a least squares function. The vector of response is a linear combination of one or more model parameters, the coefficients of regression. The model is written in matrix notation as:

$$\mathbf{s} = \mathbf{P}\beta + \epsilon \tag{4.1}$$

where $\mathbf{s}$ is the vector of predicted soil attribute, $\mathbf{P}$ is the matrix of predictors, β is the parameter vector, and the error component ϵ represents the unexplained part of the response variable.

The maximum likelihood principle [Dekking et al., 2005] provides a way to estimate the parameter involved in regression equation using usually ordinary least squares (OLS), with following underlying assumptions on error:

(i) independently and identically distributed;

(ii) zero mean and finite variance;

(iii) normally distributed.

Linear regression can be viewed as a particular form of Generalized Linear Models (**GLMs**) where we have an identity link function, a normal distribution and a constant variance [Lane, 2002]. An example can be found in

McKenzie and Ryan [1999], where authors model total organic carbon using terrain parameters such as NDVI, Prescott Index, plan curvature and others.

Because of its ease and availability MLR models have been widely used for the purpose of deriving relationships between soil attributes and ancillary variables. In soil science literature MLR is also known as *scorpan* model (Eq. 1.3) [McBratney et al., 2003].

4.2 Linear Geostatistics

Geostatistics is a collection of statistical methods which where traditionally used in the analysis of mining processes. First developments in the estimation procedure was carried out by the pioneering work of D. Krige in South African gold mines. He quantified spatial correlation between observations through the basic tool in geostatistics, the *variogram*. After a synthetic function was adapted for fitting the experimental variogram, it was used to make predictions at unobserved locations. This procedure, called kriging, was then developed in a more robust statistical theory by a mathematician named G. Matheron.

From theory of regionalized variables [Matheron, 1973] gets down that the spatial variation of any soil variable S can be expressed as the sum of three components: (1) a deterministic trend component m($\mathbf{u}$); (2) a random spatially autocorrelated component R($\mathbf{u}$); and (3) a random residual spatially uncorrelated component ϵ.

$$S(\mathbf{u}) = m(\mathbf{u}) + R(\mathbf{u}) + \epsilon \tag{4.2}$$

where $\mathbf{u}$ is the matrix of ($\mathbf{x}$; $\mathbf{y}$; $\mathbf{z}$) coordinates, and S($\mathbf{u}$) is the *random variable* that governs the realizations of s($\mathbf{u}$) at each point $\mathbf{u}$ over study area A.

The family of random variables over A is a *random function* (**RF**), for which the assumption of *stationarity* has to be kept. That is all the moments of a random function should be invariant under translation [Armstrong, 1998]. Since only the first two moments (mean and covariance) can be verified on experimental data, a weaker hypothesis is maintained, called *second order stationarity*. A further weaker stationarity, the *intrinsic hypothesis*, is developed as in the case of trend phenomena to make the first two moments independent from location $\mathbf{u}$, that is

$$E[S(\mathbf{u}+\mathbf{h}) - S(\mathbf{u})] = 0 \tag{4.3}$$

$$Var[S(\mathbf{u}+\mathbf{h}) - S(\mathbf{u})] = 2\gamma(\mathbf{h}) \tag{4.4}$$

where $\gamma(\mathbf{h})$ is called semivariogram (see Eq. 4.6).

Accounting for a single soil attribute in space three kriging variants can be distinguished as the ways in which the trend component (Eq. 4.2) can be considered. In *simple kriging* $\mathrm{m}(\mathbf{u})$ is constant and known for whole study area A, in *ordinary kriging* it is unknown and constant only within a local neighborhood. In *universal kriging* $\mathrm{m}(\mathbf{u})$ varies even within each local neighborhood $\boldsymbol{W}(\mathbf{u})$ as a multi-linear function of locations $\mathbf{u}$ (kriging with a trend) or of exhaustive secondary information (kriging with external drift, regression kriging) [Goovaerts, 1997, Hengl et al., 2007].

Kriging belongs to the family of generalized least squares regression algorithms, and can be described as a moving weighted average estimating property values at unsampled points based on the relative distance of the neighbouring sampled points. The basic estimator common to all kriging variants is written as

$$S^*(\mathbf{u}) = \sum_{\alpha=1}^{n(\mathbf{u})} \lambda_\alpha(\mathbf{u})[S(\mathbf{u}_\alpha) - m(\mathbf{u}_\alpha)] + m(\mathbf{u}) \tag{4.5}$$

where $S^*(\mathbf{u})$ is the estimated soil attribute, $\lambda_\alpha(\mathbf{u})$ is the vector of weights assigned to the random variable $S(\mathbf{u}_\alpha)$ whose outcomes is the attribute $s(\mathbf{u}_\alpha)$ at sampled points, finally $m(\mathbf{u}_\alpha)$ and $m(\mathbf{u})$ are the expected values of $S(\mathbf{u}_\alpha)$ and $S(\mathbf{u})$ respectively.

To get started with a geostatistical analysis it is necessary to study the main features of regionalization through the so-called *structural analysis*. It involves three main steps, that is the preliminary checking of data, the calculation of the experimental semivariogram (Eq. 4.6) and the fitting of an allowed mathematical model to the experimental variogram. After structural analysis is fulfilled, the successive step consists in kriging or simulation.

$$\gamma(\mathbf{h}) = \frac{1}{2N(\mathbf{h})} \sum_{\alpha=1}^{N(\mathbf{h})} [s(\mathbf{u}_\alpha) - s(\mathbf{u}_\alpha + \mathbf{h})]^2 \tag{4.6}$$

The synthetic model fitted to the experimantal variogram is used to solve the kriging weights ($\lambda_\alpha(\mathbf{u})$), which in turn are used in Eq. 4.5 to compute estimated outcomes of the random variable $\mathrm{S}(\mathbf{u})$ at unvisited locations $\mathbf{u}$.

In conditional Geostatistical Simulation (**GS**) the problem is to build a $\mathbf{RF}_{cs}(\mathbf{u}_\alpha)$ conditional and isomorphic to $\mathbf{RF}(\mathbf{u}_\alpha)$ [Journel and Huijbregts, 1978]. Hence the requirements of a conditional simulation to be satisfied are:

Isomorphism-I — $\mathbf{RF}_{cs}(\mathbf{u}_\alpha)$ has the same expectation of $\mathbf{RF}(\mathbf{u}_\alpha)$ (Eq. 4.3).

Isomorphism-II — $\mathbf{RF}_{cs}(\mathbf{u}_\alpha)$ has the same second-order moment of $\mathbf{RF}(\mathbf{u}_\alpha)$ (Eq. 4.4).

Conditionality — At the experimental data points the simulated and experimental values must be the same.

A web resource for geostatistics and spatial statistics (lattice data, point patterns, geoinformatics, etc.) can be found at http://www.ai-geostats.org/. It is maintained by the Institute for Environment & Sustainability, Joint Research Centre.

4.3 Soft Computing

Soft computing is a consortium of methodologies that model very complex real world phenomena accommodating the guiding principle of tolerance for imprecision, uncertainty, partial truth and approximation to achieve tractability, robustness, low-solution cost and better rapport with reality [Zadeh, 1994]. Its principal constituents are fuzzy logic (FL), neurocomputing (NC), genetic computing (GC) and probabilistic reasoning (PR) [Tsoukalas and Uhrig, 1997]. Only the first three members are used in this thesis. FL is a rule-based system that deals with approximate reasoning, vague, ambiguous, imprecise, noisy, or missing input information. NC tackles problems of system identification, learning or adaptation, while GC is a technique of systematized random research used to find approximate solutions to optimization problems.

4.3.1 Artificial Neural Network

From Haykin [1998] I extract an exhaustive definition of what an ANN is:

> *"A neural network is a massively parallel distributed processor made up of single processing units, which has a natural propensity for storing experiential knowledge and making it available for use. It resembles the brain in two respects: (i) knowledge is acquired by the network from its environment through a learning process, and (ii) interneuron connection strengths, known as synaptic weights, are used to store the acquired knowledge."*

Artificial Neural Networks resemble functionally and topologically the human brain with its billions of neurons and trillion of synapses. The human brain is capable of highly complex, nonlinear, and parallel computing engaging highways of connected neurons. It is able to quickly address functions such as perception, pattern recognition and motility. The brain is the central part of the three-component nervous system [Arbib, 1987], which converts stimuli from the external environment into electrical impulses (through

receptors) that excite brain, and vice versa (through effectors) when brain generates an output response.

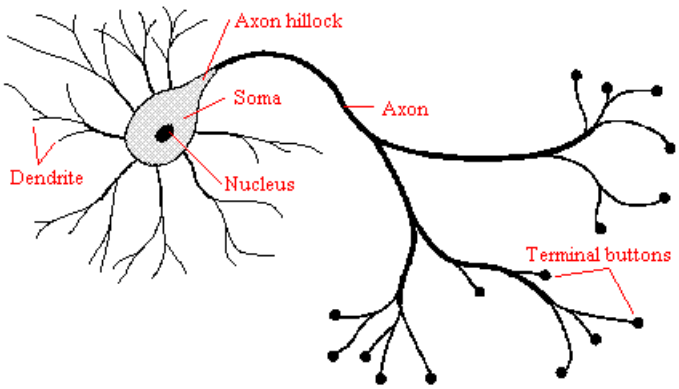


Figure 4.1: Schematic of biological neuron.

A neuron (Fig. 4.1) is the fundamental structural (cellular) unit and processing element of the brain. It receives input signals along a dendrite, but first a signal have to cross the synaptic junction, where its electro-chemical transduction takes place. The stimulus reach the soma (cell body of neuron); if it is strong enough the neuron fires and conveys the mediated stimulus to the other neurons connected through the axon. Plasticity is the ability of neurons to adjust the impedance or conductance of their synapses, and is the process that leads to memory and learning.

The neurobiological analogy is now ready to promote a fully comprehension of the structure and function of an artificial neural network. The basic information-processing unit of an ANN is the artificial neuron (Fig. 4.2).

It is composed of three characteristic elements: (i) a set of links mediated by weights (w) that conveys the input signals to the soma; (ii) the summing junction which generates the so called induced local field (I_k); and (iii) the activation function whose main task consists in squashing the output of the neuron within a preset amplitude range.

Different activation functions, also called transfer functions, exist and they can be grouped in three basic types:

- the threshold function fires the neuron when the induced local field exceeds the assigned threshold;

- the piece-wise linear function embody a linear region of signal transmission;

- the s-shaped sigmoid function is the most widespread transfer function used in the construction of ANNs. Examples are the logistic and the hyperbolic tangent functions defined respectively by

$$\phi(I) = \frac{1}{1 + e^{-I}} \tag{4.7}$$

$$\phi(I) = \frac{e^{I} - e^{I}}{e^{I} + e^{I}} \tag{4.8}$$

The way in which neurons are structured in a multilayer perceptron (MLP) depends upon the learning algorithm used to train the net. When a neural network is stimulated by environment undergoes changes in the synaptic transmittance strength and therefore react in a new way to the environment. The way in which the synaptic weights (the free parameters of the model)

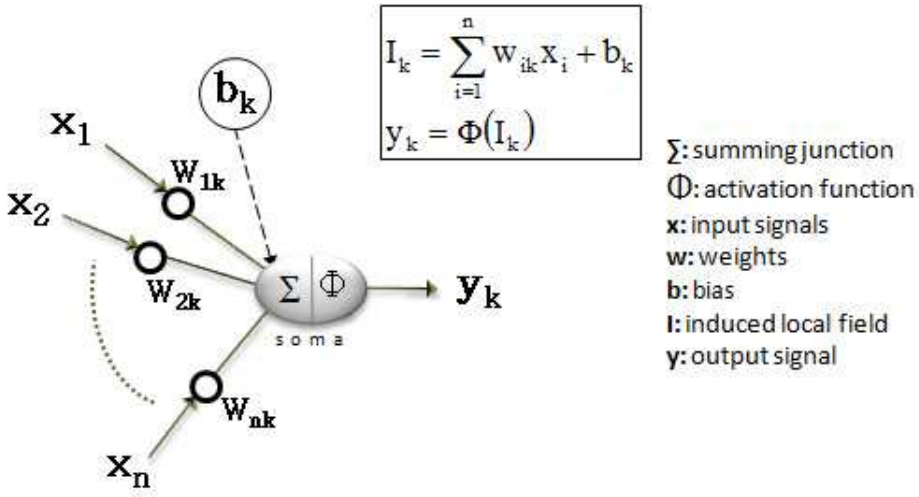


Figure 4.2: Model of artificial neuron.

are adjusted is a task of the training algorithm. In this thesis the class of fully connected back-propagation feedforward neural networks (BP-FFNN) are trained with a teacher (supervised training).

A MLP is made of two or more parallel layers of fully interconnected neurons; this means that a three-layered net presents an input layer, one hidden layer and an output layer. MLPs are generally trained with the error-correction back-propagation learning algorithm. In this framework signals flow following two basic directions; an environmental stimulus generates a forward propagation of function signals while the error signals back-propagate. The error-correction rule may be viewed as a general case of the

ubiquitous least mean squares (LMS) algorithm and generates an adjustment in a synaptic weight proportional to the product of the learning-rate parameter, the error signal and the input signal. put equation?

The training phase is a step-by-step adjustment to the synaptic weights until the system reaches a stopping criteria or a steady state. The nonlinear input-output mapping computed by a network could led to overtraining (or overfitting), a phenomenon in which the net may memorize patterns present in training dataset but not belonging to the underlying function to be modelled. When a network is overtrained, it loses the ability to generalize when unseen input is given during simulation phase.

ANNs require specialized skills to be implemented, even if nowadays softwares offer the opportunity to execute hard tasks through simplified user interfaces which for instance mediate the use of complicated mathematical formulas and computations. Moreover, results entail difficulty of interpretation.

4.3.2 Fuzzy Logic

The concept of Fuzzy Logic (FL) was conceived by Lotfi Zadeh, a professor at the University of California at Berkley, and presented as a way of processing data by allowing partial set membership rather than crisp set membership or non-membership. FL's approach to control problems mimics how a person would make decisions, only much faster.

FL incorporates a simple, rule-based *IF X AND Y THEN Z* approach to a solving control problem rather than attempting to model a system mathematically. The FL model is empirically-based, relying on an operator's experience rather than on their technical understanding of the system. For example, rather than dealing with landform classification in terms such as *Shoulder facet occurs when profile and plan curvatures are less then a threshold, slope angle is equal to zero, etc.*, terms like *IF (profile curvature is planar) AND (plan curvature is planar) AND (slope angle is near level) AND . . . , THEN (landform facet is a Shoulder)* are used. These terms are imprecise and yet very descriptive of what must actually happens.

FL has proven to be an excellent choice for many control system applications since it mimics human control logic. It uses an imprecise but very descriptive language to deal with input data more like a human operator. It is very robust, and often works when first implemented with little or no tuning.

Chapter 5

Methodological Results: EDASS

At this stage a complete set of geospatial information is structured in a georeferenced database. On one hand there is the soil database composed of punctual data about the vertical variability of investigated soil profiles. On the other hand there is the continuous data made of auxiliary maps (Tab. 5.1). To fulfil statistical analysis it is needed a matching table in which collect site and profile data, and values from terrain parameters at sample locations. For the purpose the *MultiFieldAdder* tool (Appendix A) was specifically designed in Visual Basic language under ArcMap. It is downloadable for free from the ESRI website at URL http://arcscripts.esri.com/details.asp?dbid=14826.

The fundamental unit of the matching table is the soil horizon; for each one the table provides a lot of valuable information concerning its label (e.g. Ap1, Bw, etc.), the pedological position within the soil profile and its thickness (upper and lower bounds), the soil matrix colour (Munsell), the geographic location (coordinates), the physical and chemical properties (texture, pH, organic carbon, etc.), the position in the landscape as codified by the terrain parameters, the vegetation cover type (CLC), the photosynthetic rate (NDVI), the pedolandscape unit (UDP), and the landform segmented facet (FLFS). Together more horizons constitute a soil profile.

The qualitative and quantitative analysis of all these features (related to soil horizons and/or profiles) via explorative statistics and visualization of information is a very hard task. Indeed a large number of possible interconnection between variables is possible while a huge amount of feasible stratifications in different domains are kept. For example it is possible to stratify in geographic domain if considering only a cluster of nearby locations, and/or in pedological domain when selecting the topsoil, the subsoil, a particular horizon type (A, B, C), or the whole profile; and/or finally we can

Table 5.1: Environmental explanatory variables available across Telese Valley study area

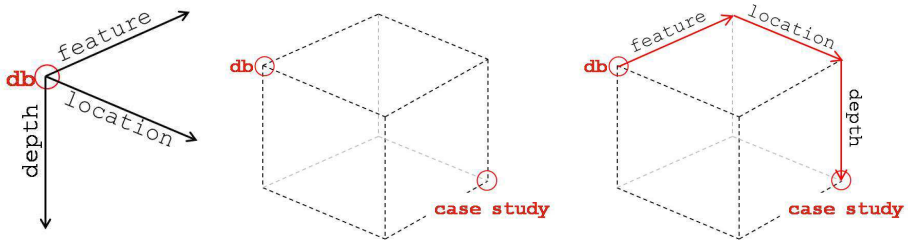
Covariate	Description	Source or reference
ELEV	Elevation above sea level	DEM from 1:25000 contour lines
ASP	Degrees clockwise from north	Burrough and McDonell [1998]
SLO	Measured in degrees	Burrough and McDonell [1998]
PROFC	Generated in ArcGIS	Moore et al. [1991], Zeverbergen and Thorne [1987]
PLANC	Generated in ArcGIS	Moore et al. [1991], Zeverbergen and Thorne [1987]
MEANC	Generated in ILWIS	Shary et al. [2002]
NORTH	Generated in ILWIS	Shary et al. [2002]
ACV	Generated in ILWIS	Hengl et al. [2003]
SPI	Generated in ILWIS	Hengl et al. [2003]
STI	Generated in ILWIS	Hengl et al. [2003]
TWI	Generated in ILWIS	Hengl et al. [2003]
SOLINS	Generated in ILWIS	Shary et al. [2002]
NDVIL	Landsat NDVI	
NDVIS5	MODIS NDVI, sum of June-September (5 layers)	
NDVIS16	MODIS NDVI, sum of March-November (16 layers)	
NDVID5	MODIS NDVI, maximum difference June-September (5 layers)	
NDVID16	MODIS NDVI, maximum difference March-November (16 layers)	
UDP	Pedolandscape units	
FLFS	Fuzzy landform segmentation	MacMillan et al. [2000]
CLC	Corine Land Cover	

stratify in feature space if considering only cases for which particular realizations of one or more attributes occur (e.g. only locations above a certain slope degree and elevation, and belonging to a certain landform element).

Therefore a database can be stratified along three principal axis (Fig. 5.1): *location* (for geography domain stratification), *variable* (for feature space stratification), and *depth* (for pedological stratification along soil profile). Given a complete database at hand, it is possible to extract any particular sub-case study by stratifying through this three main directions.

A priority can be established amongst the cube of feasible directions: firstly a target variable to be investigated is chosen and subsequently both the possible states of other variables and the covariates are set. Secondly, all or a part of locations is selected. Finally the amount and types of horizons per profile are fixed. After that one performs Explorative Data Analysis (EDA, Martinez and Martinez [2005]) through statistical computation and

graphical visualization.



(a) Database complexity.

(b) Stratification cube.

(c) Stratification path.

Figure 5.1: Database exploration across three main components.

Surveying a soil scientist corroborates his own mental model about the spatial arrangement of soil bodies in a landscape relative to the factors of soil formation. Then he use the new acquired awareness to explore relationships between soil properties and environmental covariates. Multitude, volatility and complexity of human reasoning should be followed by as much quick and articulated operative stratification capability on the database.

Unfortunately the delineation of a particular case study from a large database is cumbersome, as one should fulfill the following steps:

1. Stratify the pedo-db following stratification cube; e.g. this step is performed in Microsoft Access through assisted user interface queries.

2. Export to a specific file format; e.g. in a tab delimited text file.

3. Eventually do preliminary adjustments to file header/corpus to make it ready for the statistical software used; e.g. re-codifying missing or no data values (Access/Excel have empty cell, R has *NA*, MatLab has *NaN*, ArcGIS has *-9999*), formatting the file header as required by GSTAT geostatistical software [Pebesma and Wesseling, 1998], or pondering attributes to a certain depth along soil profiles [Meersmans et al., 2008].

4. Import the file in the statistical software for computing; e.g. R (`http://www.r-project.org/`), GSTAT, SPSS or ISATIS.

This list of operations can grow as more specifically-designed tools are used to implement each step. Then the time required to carry out several trials based on different stratifications is exponentially raised, or one should content about very few selected subcase studies.

To make easy the task of exploring vertical and lateral variability of sampled soils, a user friendly tool, **EDASS** (Explorative Data Analysis with Stratification and Statistics), is built in Visual Basic for Applications (VBA) under Microsoft Access (see Appendix B). The objective is to condense in a single window the power of fast and intricate queries as per the stratification cube, the ability to create several graphics such as scatter diagrams, experimental semivariograms, and the possibility to make quick and yet preliminary statistical computations such as correlations, variography and kriging.

The EDASS tool is a handy interface that links in the background the power of SQL statements possible in Access with graphical and statistical capabilities of R statistical software. The interoperability Access-R exploited by the EDASS tool facilitates patterns discovery from the pedological database, allows the identification of structured variables in space (existence of autocorrelation), and enlarges stimuli necessary to build up a mental model about the soilscape at hand.

In Appendix B you will find a detailed description of how the EDASS user interface is designed to solve this preliminary steps and how it works on a selected example.

In the following section EDA is carried out for some soil features and it is pointed out how the pedo-db is inquired by the EDASS tool to extract the subcase studies for the Telese Valley landscape. Therefore, for each target attribute it is showed the peculiar selections made to address spatial analysis by means of specific techniques (chapter 6).

5.1 The set of proposed stratifications

5.1.1 Clay content

As regards the soil texture, the pedo-database can be distinguished into two groups of soil samples; the first was analysed by pipette method during the first phase of soil survey, while the second group was analysed later on by laser method to quickly enlarge the number of cases at disposal for an artificial neural network based model of inference.

In order to address spatial analysis by means of more techniques, three types of stratifications are fulfilled on clay content:

1. pondering for whole soil profile for those samples analysed by pipette method, to allow ANOVA computation across FLFS facets (§ 6.4);

2. topsoil for profiles analysed by pipette method to run universal kriging on sparse locations (§ 6.2.1);

3. all horizons analysed by both pipette and laser methods to accomplish performance comparison between the MLR (§ 6.1.1) and the ANN (§ 6.3.1) techniques.

5.1.2 Soil colour

Quantitative analysis

Soil colour is commonly described qualitatively using Munsell soil colour charts. Colour is therefore decomposed in hue (dominant wavelength), value (overall brightness) and chroma (saturation of colour) (Fig. 5.2), and is assigned to a single Munsell chip under standardized illumination conditions, that is at sunlight on a clear day (the so called illuminant D_{65}, see Wyszecki and Stiles [1982]). Even though the Munsell HVC system is a good choice for handling soil colour, numerical analysis is not possible since the HVC tristimulus describe a perceptual colour space in a discrete form and descriptions include both letters and numbers.

A table (`soil_colors.dat`) was downloaded from the Munsell Color Science Laboratory website, which contains six variables: the three Munsell HVC components, and the x, y, and Y components of the CIE xyY colour system. This file is used to perform multidimensional interpolation from Munsell data to RGB triplets. Operations involved in the conversion between colour spaces are written in an R script file reported below, in which equations are borrowed from the paper of Rossel et al. [2006] and from Bruce Lindbloom website:

```
# You should modify "INSERT INPUT" in order to set your case.
# There exist two numbering convenction for RGB: [0.0, 1.0] or [0.0, 255.0].
# Last modify on 24/nov/2008 by Giuliano Langella

# set working directory INSERT INPUT
setwd("C:/Dottorato/pedometrics/color")

# read in the soil colors: munsell + xyY INSERT INPUT
soil <- read.table("soil_colors.dat", header=TRUE)

#convert xyY --> CIE XYZ
attach(soil)
soil_X <- x * (Y/y)
soil_Y <- Y
soil_Z <- (1 - x - y) * (Y / y)
detach(soil)

### convert XYZ --> RGB with a gamma of 2.4
soil_X = soil_X/100
soil_Y = soil_Y/100
soil_Z = soil_Z/100

soil_R = soil_X * 3.2406 + soil_Y * -1.5372 + soil_Z * -0.4986
soil_G = soil_X * -0.9689 + soil_Y * 1.8758 + soil_Z * 0.0415
soil_B = soil_X * 0.0557 + soil_Y * -0.2040 + soil_Z * 1.0570

fun1 <- function(x) { y <- 1.055 * ( x ^ ( 1 / 2.4 ) ) - 0.055 ; y}
fun2 <- function(x) { y <- 12.92 * x ; y}

R <- ifelse(soil_R > 0.0031308, fun1(soil_R), fun2(soil_R))
G <- ifelse(soil_G > 0.0031308, fun1(soil_G), fun2(soil_G))
B <- ifelse(soil_B > 0.0031308, fun1(soil_B), fun2(soil_B))

#clip values to range {0,1}
R_clip <- ifelse(R < 0, 0, R)
G_clip <- ifelse(G < 0, 0, G)
B_clip <- ifelse(B < 0, 0, B)

R_clip <- ifelse(R > 1, 1, R_clip)
G_clip <- ifelse(G > 1, 1, G_clip)
B_clip <- ifelse(B > 1, 1, B_clip)

# Deactivate this three lines if range is to be [0.0, 1.0] INSERT INPUT
R_clip <- round(R_clip*255)
G_clip <- round(G_clip*255)
B_clip <- round(B_clip*255)

#add back to original table:
soil$R <- R_clip
soil$G <- G_clip
soil$B <- B_clip

# plot the manually converted data
library(plotrix)
library(colorspace)
plot( as(RGB(R_clip,G_clip,B_clip), 'LUV'), cex=0.5)

# write the output to a file: INSERT INPUT
write.table(soil,"soil_colors-RGB.dat",sep="\t")

#Delete temporary objects
rm(soil,soil_X,soil_Y,soil_Z,soil_R,soil_G,soil_B,fun1,fun2,R,G,B,R_clip,G_clip,B_clip)
```

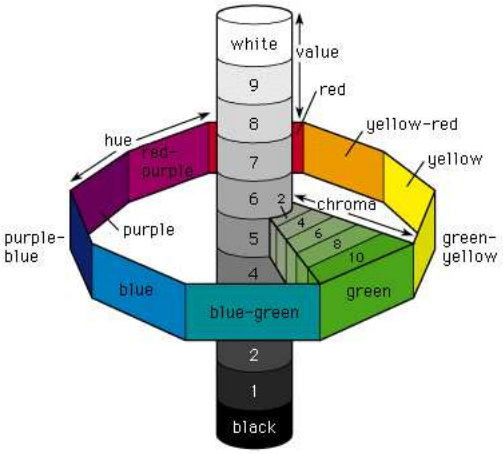


Figure 5.2: The Munsell colour model represented by a cylindrical coordinate system (from http://www.britanica.com/)

A new file is created (`soil_colors-RGB.dat`) with information about three colour systems, the Munsell HVC, the CIE xyY and the RGB. This file is further manipulated to create an ILWIS domain identifier, with which assign to a Munsell soil colour label an RGB triplet in order to display the soil colour as seen in the field (circles, Fig. 5.3). In ILWIS RGB triplets are

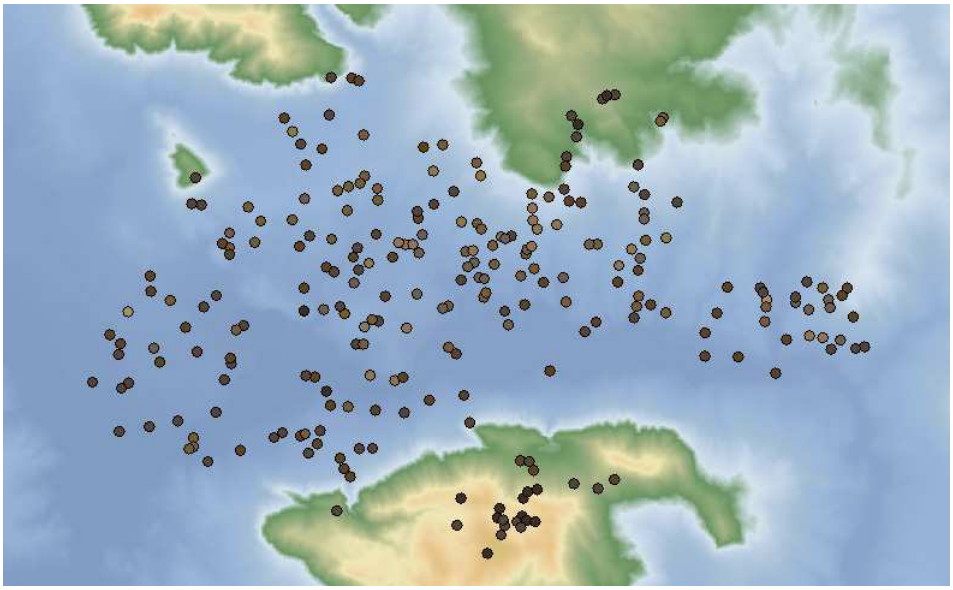


Figure 5.3: Munsell soil colour of topsoil after conversion in RGB

transformed in hue, saturation and intensity, the HSI colour system. RGB

data is highly correlated, hence it is transformed in three statistically decorrelated components, H_{RGB}, S_{RGB} and I_{RGB} [Rossel et al., 2006]. The RGB, HSI and HSI_{RGB} colour systems are spatially investigated in the framework of regression kriging in § 6.2.2.1, using the following stratification:

1. *feature space*: RGB, HSI and HSI_{RGB} triplets as nine dependent variables, and land surface parameters, remote sensing imagery as auxiliary information;
2. *locations*: all samples;
3. *pedological domain*: topsoil.

PDI

Another type of manipulation of Munsell soil colour is addressed by means of a quali-quantitative assignment of scores to Munsell HVC colour system. The main goal is to build an indicator of degree of pedogenesis (**PDI**, Pedogenetic Degree Index). We know that soil colour definitely does not account for only its evolution, however the idea behind the proposed decodification is that each colour component partly hold an information about the power of soil-forming factors a soil body experienced over time. This way each of the three components is studied in order to understand for instance the way in which it could be profitably put in decreasing order of PDI.

The Munsell HVC components are treated separately into two groups, the hue component on one hand and the value and chroma components on the other hand. **Hue** is defined categorically by an alphanumeric label, i.e. *7.5YR*, in which letters are abbreviations of the colours of the spectrum (i.e. YR for yellow-red), and preceding number (i.e. 7.5) which range from 0 to 10 return a more yellow and less red hue as number increases. **Value** is specified on a numerical scale from 0 (absolute black) to 10 (absolute white). **Chroma** is also described numerically beginning at 0 for neutral greys (the achromatic point) to a maximum value of 20, which is never approached with soil.

For Telese valley landscape it is proposed the set of assignments reported in Tab. 5.2 for hue and Tab. 5.3 together for value and chroma. Here only the range of values found in the study area are analysed and decodified in terms of pedogenetic degree index.

It is assumed that *hue* goes from R to G as soil obliteration by soil-forming factors over time decreases, and consequently a larger PDI_H is assigned. Similarly, decreasing values of *value* and *chroma* are assigned to smaller PDI_{VC}. Particular attention is payed for simultaneous occurrences of relative low values for *value* and *chroma*; indeed the dichotomy $\&/xor$ is adopted in order to distinguish between simultaneous/exclusive occurrences respectively for value and chroma.

Munsell HUE	
PDI_H	Component
1	10R
2	2.5YR
3	5YR
4	7.5YR
5	10YR
6	2.5Y
7	5Y
8	7.5Y
9	10Y
10	...G ...

Table 5.2: Attribution of *pedoclass* values to hue component

Munsell Value and Chroma				
PDI_{VC}	from	to	Component	Relation
1	0	2	VALUE	&
	0	2	CHROMA	
2	0	2	VALUE	xor
	0	2	CHROMA	
3	3	5	VALUE	&
	3	4	CHROMA	
4	3	5	VALUE	xor
	3	4	CHROMA	
5	6	7	VALUE	&
	5	6	CHROMA	
6	8	10	VALUE	&
	7	8	CHROMA	

Table 5.3: Attribution of *pedoclass* values to value and chroma components

This means for instance that PDI_{VC} equal to 1 is assigned only to those combinations of value and chroma between range [0, 2], which are solved by & relation. When only one of the two components is within the aforementioned range a larger PDI_{VC} (=2) is assigned thanks to the xor operator. Note that the last two assignments of Tab. 5.3 include only the & operator since the xor combination for these intervals are already solved by lower PDI_{VC} xor statements. The PDI for the Munsell soil colour is thus defined by

$$PDI^{+} = PDI_V + PDI_{VC} \tag{5.1}$$

where the plus sign of PDI^{+} is adopted to highlight the sum of the two

quantities. The PDI^+ is inserted into the matching table to start explorative

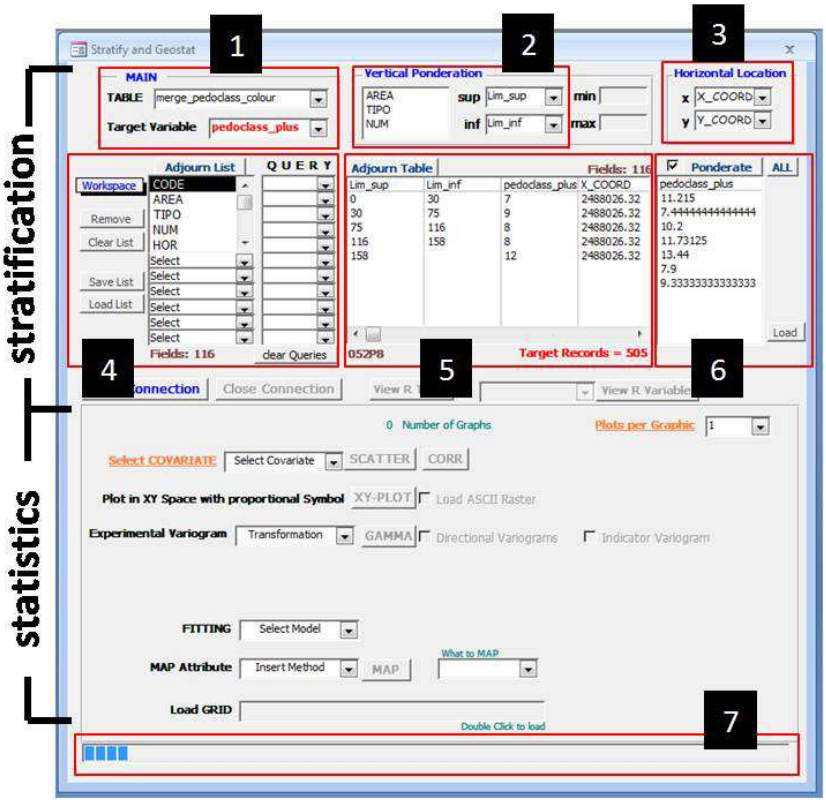


Figure 5.4: EDASS while pondering PDI^+. Steps from 1 to 7 are explained in the text.

analysis with EDASS. Main steps are exposed in the following to take insight about the usefulness of EDASS toolbox in studying for instance the PDI^+ variable. Firstly, it is highlighted the stratification power of the tool with the help of Fig. 5.4, in which numbers correspond to the following list of operations:

1. The matching table is selected from the database and then is the turn of the target variable PDI^+.

2. Key fields are loaded to allow pondering through selection of all horizons belonging to a soil profile. Also upper and lower bounds variables are selected for a thickness weighted pondering.

3. Spatial coordinates are loaded.

4. In this box selections on whatsoever attribute is possible to stratify data.

5. This Visual Basic listbox lists fields and cases; it is adjourned with selections made in step 4 and with operations user is running (as the pondering task itself depicted in Fig. 5.4).

6. Here user can start the pondering task. It computes a weighted pondering proportional to the thickness of soil horizons considered per profile. Pondering can be made along the whole soil profile or stratifications at step 4 allow different upper–lower bounds possibilities. Once operation is ended user can save results to a text file or continue with the statistical computation through EDASS.

7. The progress bar activates to give an idea of elapsed and remaining time.

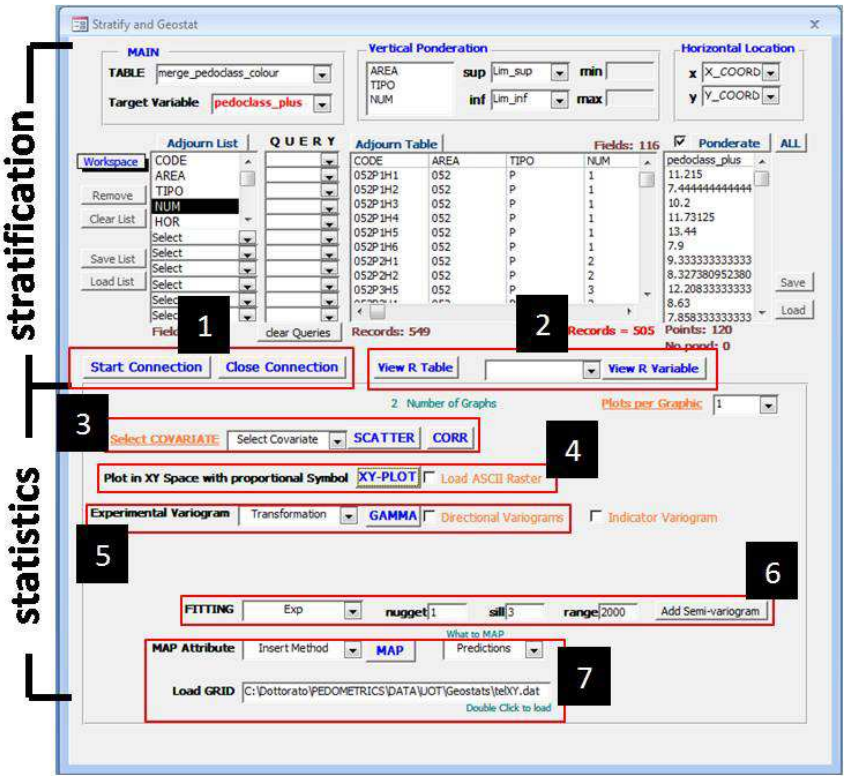


Figure 5.5: Statistics in EDASS on PDI^+. Steps from 1 to 7 are explained in the text.

The first part of operations is now executed, and pondering PDI^+ is calculated. Anyway user might carry on with the statistical analysis by means of EDASS tool also without pondering. Steps involved in explorative statistical analysis (Fig. 5.5) are listed below:

1. By pressing the 'Start Connection' button the R software is run in the background and receive in the form of a table data that user have selected.

2. Here one can handle the R objects (Fig. 5.6 a), by directly viewing them as tables.

3. Once goodness of R objects is ascertained (discretionary step), it is possible to perform a bivariate analysis with selected covariate by means of Pearson correlation coefficient and scatter diagrams (see Fig. 5.6 b).

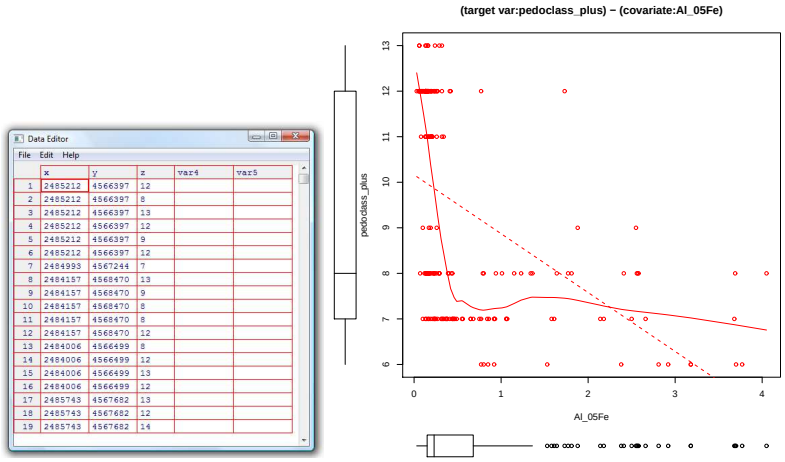
4. An xy scatterplot with coordinates is made to investigate spatial patterns of target variable (Fig. 5.6 c). To facilitate spatial analysis through visualization the tool takes care of three aspect: (i) marker size is plotted proportional to the target variable magnitude; (ii) if a covariate is selected the marker is filled with a grey scale color proportional to the magnitude of the covariable; (iii) it is possible to load an Arc-Info ASCII Raster in order to investigate spatial patterns of target also in relation to a continuous auxiliary information.

5. In this section of EDASS, a semivariogram of target can be plotted omnidirectional (Fig. 5.6 d) or along preferred directions (Fig. 5.6 e)

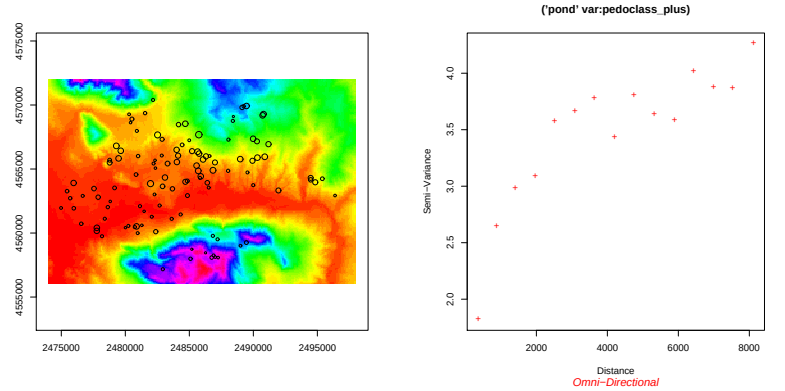
6. User can explicit the synthetic variogram which assumes to best fit experimental variogram.

7. Here a kriging map (Fig. 5.6 f) is produced accounting for the variogram model previously selected at unknown locations loaded in the form of an xy table.

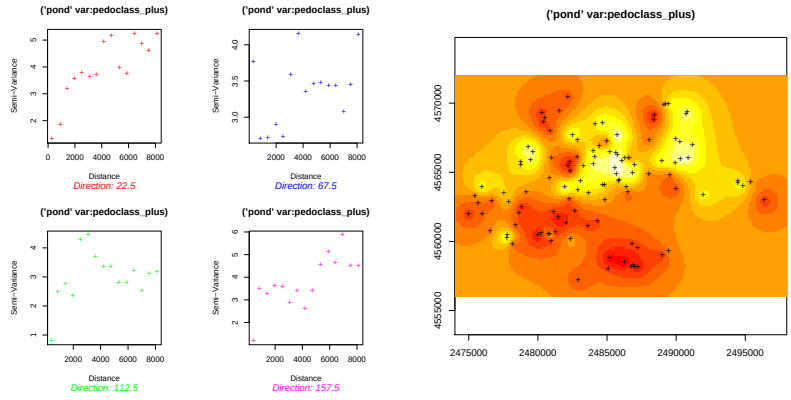
Explorative data analysis with EDASS is intended only as a preliminary task, however it revealed to be very powerful in highlighting the good structure of the omnidirectional semivariogram of PDI^+ (Fig. 5.6 d). In § 6.2.2.2 will be presented the spatial inference of pondering PDI^+ by means of ordinary kriging.



(a) View R object (b) Scatter with $Al + 1/2 \cdot Fe$



(c) XY-Plot (d) Semivariogram



(e) Directional Variogram (f) Kriging

Figure 5.6: Main steps in explorative analysis of PDI^+ through EDASS

Chapter 6

Methodological Results: Inference Setup

In chapter 4 linear and non-linear statistical models of spatial interpolation are introduced with a brief theoretical description. Here the focus is on the specific applications developed.

The development of quantitative soil-landscape models is addressed by using GIS technologies, a digital elevation model, terrain analysis (§ 3.1), remotely sensed imagery, the powerful exploration capacity of EDASS tool (chapter 5 and appendix B), and statistical analysis.

A range of statistical data analysis methods are applied to develop models for spatial prediction using environmental correlation [McKenzie and Ryan, 1999]. These include neural nets (§ 4.3.1), fuzzy logic (§ 4.3.2), generalized linear models (§ 4.1), and geostatistics (§ 4.2).

6.1 Multiple Linear Regression

Regression models are developed in a stepwise manner. The multiple linear regression equations are calculated using the significance F test of Fisher Snedecor at 0.5% level for entering (**pIN**) an independent variable and at 1% for removing (**pOUT**) a predictor already in equation.

The stepwise method starts entering the most significant independent variable (lowest p-value) if and only if its p-value is less than pIN; each time a new eligible predictor with lowest p-value is entered (if p-value $<$ pIN) the stepwise procedure reexamine all the predictors in equation and the variable with the largest probability of F is removed if the value is larger than pOUT. This process continues until no variables in the equation can be removed and no variables outside the equation are suitable for entry.

Regression analysis is carried out in SPSS or in MatLab; the results were preliminary compared between softwares to capture possible differences in parameter estimation. The *ANNvsREGR* script of MatLab (app. C) is preferred when the dependent variable is also analysed by means of artificial neural networks, such as the case of clay content.

6.1.1 Clay content

In literature the spatial prediction of clay content is found for instance in the work of Odeh et al. [1994] in which landform attributes derived from a DEM were used, and in the work of Odeh and McBratney [2000] where authors put remotely sensed data in a multivariate spatial prediction model by using NOAA-AVHRR (National Oceanic and Atmospheric Administration – Advanced Very High Resolution Radiometer) data.

In this work the complete set of horizons is used for each soil profile as stated in § 5.1.1. The matching table with clay content and auxiliary maps is elaborated with the MatLab script *ANNvsREGR*. The script is printed in Appendix C where also a useful description about its handling in MatLab environment is supplied.

6.2 Linear Geostatistics

Geostatistical analysis on soil attributes is performed employing the EDASS tool only for preliminary discovery of spatial structures (chapter 5), but a finer study is tuned by means of specifically designed softwares, such as R, GSTAT stand alone and a MatLab tool called mGstat, which implements among others the GSTAT package.

6.2.1 Universal Kriging of clay content

Universal Kriging (**UK**) is mathematically equivalent to Regression Kriging (**RK**) [Hengl et al., 2007]. UK solves kriging weights using directly the auxiliary predictors, while RK interpolation method profit both by a separate interpretation of the trend component m(**u**) and residual stochastic component R(**u**) (see Eq. 4.2), and by the possibility to extend the study of m(**u**) to a broader range of regression techniques.

The analysis of clay content is here carried out within the stand-alone GSTAT package. GSTAT requires a batch file with extension *cmd* that contains a series of commands executed in order. Various information is provided to the program, such as the matching table file path (e.g. `text052.eas`),

the set of predictors (`X=18&20&25`) and the synthetic semivariogram model (`variogram(clay)`) with which fitting R(**u**).

Here below the Windows command file `'UK_clay_res.cmd'` necessary to execute universal kriging of clay in Telese valley study area is fully reported (see GSTAT examples for more insight):

```
# UK of clay content in Telese valley study area
points(clay): 'text052.eas',x=5,y=6,v=8,X=18&20&25;
variogram(clay): 298.01 Nug(0) + 5486.25 Exp(7377.63);
mask: 'acv.asc','fill_telese20.asc','north.asc';
predictions(clay): 'Pred_UK_clay_res_2.asc';
variances(clay): 'Var_UK_clay_res_2.asc';
```

The `#` symbol indicates that the current line does not contain command statements, but a user description of file content. Auxiliary predictors are provided in the same order in `X` (from matching table) and in `mask` (from the ASCIIGRID maps). Outlets of universal kriging are the maps of predictions and variances.

Under the DOS command prompt the command file is addressed to gstat.exe for computation, through the following statement:

```
>GSTAT UK_clay_res.cmd
```

GSTAT is located under `root\windows\system32\`. Another way to do this, consists in calling gstat from MatLab environment using the mGstat package

```
>> gstat('UK_clay_res.cmd');
```

The two statements produce an identical result in quite the same time, 61 seconds for GSTAT and 62 seconds for mGstat. In fact the two ASCIIGRID maps of predictions, the one by GSTAT and the other one by mGstat, were imported in MatLab and compared pixel-by-pixel. The ARC/INFO rasters in ASCII format are imported in MatLab through a personal function, the ImportAsciiRaster. By clicking the link you open the MatLab central file exchange from which watch or download the code.

In order to properly design the variogram model to put in command file, it is necessary to calculate the regression of the dependent variable on auxiliary variables first. This way the matching table (`'texture.dbf'`) is firstly imported in SPSS

```
GET TRANSLATE
     FILE='C:\Dottorato\PEDOMETRICS\DATA\UOT\RegrKr\TEXTURE\texture052.dbf'
     /TYPE=DBF /MAP .
```

and secondly a correlation matrix is computed to evaluate the performance of available environmental covariates:

```
CORRELATIONS
      /VARIABLES=argilla limo_tot sabbia_t acv catch fill_tel glf_chan
      glf_plai glf_ridg meanc north planc profc spi sti twi aspect_f
      slope_ft
      /PRINT=TWOTAIL NOSIG
      /MISSING=PAIRWISE .
```

The final step is regression analysis. The stepwise method provides a way to select eligible predictors on the basis of p-values (`pIN` and `pOUT`) for inserting or removing independent variables:

```
REGRESSION
      /MISSING LISTWISE
      /STATISTICS COEFF OUTS R ANOVA COLLIN TOL
      /CRITERIA=PIN(.05) POUT(.10)
      /NOORIGIN
      /DEPENDENT argilla
      /METHOD=STEPWISE acv catch fill_tel glf_chan glf_plai glf_ridg
      meanc north planc profc spi sti twi aspect_f slope_ft
      /SCATTERPLOT=(*SDRESID ,*ZPRED )
      /RESIDUALS HIST(ZRESID)
      /PARTIALPLOT ALL
      /SAVE ZPRED COOK LEVER .
```

Although stepwise method can be applied when experimental design does not contemplates preexistent relationships amongst variables, it should not be used as a standard method. Indeed a physical relationship between target and predictors is not always ensured specially for sparse data, since it could be a result of chance.

Note also that optional functions are activated in the script, such as the histogram of regression residuals for normality check, and the set of partial regressions plots to evaluate graphically the contribution of each predictor to the explanation of target variance.

Residuals are computed using regression coefficients from stepwise method. Variographic analysis on R($\mathbf{u}$) is accomplished in `gstatw.exe` (Fig. 6.1). The spatial prediction of clay content is given by the sum of the predicted drift and residuals:

$$s^*_{CLAY}(\mathbf{u}_0) = m(\mathbf{u}_0) + R(\mathbf{u}_0) =$$

$$= \mathbf{q_0^T} \cdot \hat{\beta}_{gls} + \lambda_\mathbf{0}^\mathbf{T} \cdot (\mathbf{s} - \mathbf{q_0^T} \cdot \hat{\beta}_{gls}) \tag{6.1}$$

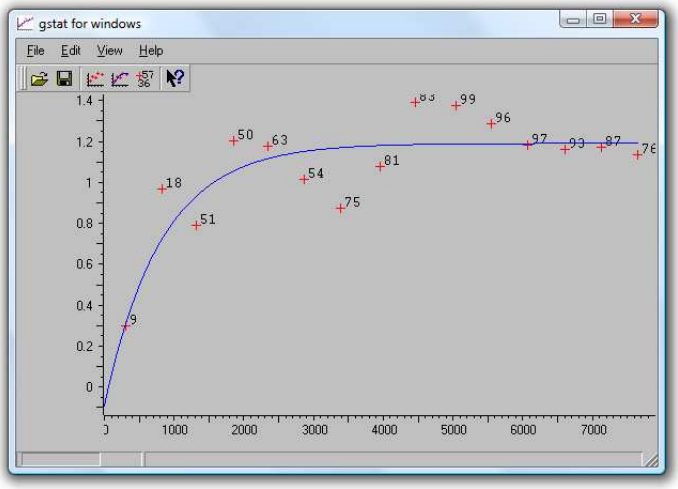


Figure 6.1: Variography of clay $R(\mathbf{u})$ in *gstatw*

were $s^*_{CLAY}(\mathbf{u}_0)$ is the predicted clay content at unvisited location $\mathbf{u}_0$, $\mathbf{q_0}$ is the vector of $p+1$ predictors at $\mathbf{u}_0$, $\hat{\beta}_{gls}$ is the vector of $p+1$ drift model coefficients estimated using the generalized least squares (GLS) method, $\lambda_\mathbf{0}$ is the vector of $\mathtt{n}$ kriging weights, and $\mathbf{s}$ is the vector of $\mathtt{n}$ sampled points.

The thorough procedure of spatial inference of clay outlined above is also performed using the logit transformation of clay content. This preliminary step on target is very precious if one would bound the map of predictions within a predefined range of possible values. Commonly logit is a transformation used for linearizing sigmoid distribution of proportions. First the target variable is standardized to the 0 to 1 range

$$s^+ = \frac{s - s_{min}}{s_{max} - s_{min}}; \qquad s_{min} < s < s_{max} \tag{6.2}$$

where s is the proportion of clay content (%), then the standardized s^+ clay is logit transformed with

$$s^{++} = ln\left(\frac{s^+}{1 - s^+}\right); \qquad 0 < s^+ < 1 \tag{6.3}$$

The bounds $[s_{min},\ s_{max}]$ can represent the physical minimum and maximum of s, i.e. $[0, 100]$ in the case of clay content, or can represents the limits of s within sampled points at hand as the $[0.4, 68.8]$ range of clay across surveyed Telese valley. Here the range $[0, 70]$ is adopted with limits outside sampled points bounds, inasmuch where dependent s^+ is zero or one s^{++} will be missing value.

Universal kriging of logit clay is fulfilled with a GSTAT command file similar to the one reported in the box on page 47. The file structure is the same but `v=...`, `variogram(clay)` and output ascii rasters are quite different. The map of predictions $\widehat{s}^{++}(A)$ made with Eq. 6.1 is imported in MatLab with the function ImportAsciiRaster, and then is back-transformed to the original scale:

$$\widehat{s}(A) = \frac{e^{\widehat{s}^{++}(A)}}{1 + e^{\widehat{s}^{++}(A)}} \cdot (s_{max} - s_{min}) + s_{min} \tag{6.4}$$

where $\widehat{s}(A)$ is the map of UK predictions of clay content.

In order to compare results, conditional geostatistical simulation [Goovaerts, 1997] is run on one hundred isomorphic realizations $\mathbf{RF}_{cs}(\mathbf{u}_\alpha)$ (pag. 25) using the following command file under GSTAT:

```
# GS of clay content in Telese valley study area
points(ARGILLA): 'text052.eas',x=5,y=6,v=8,X=18&20&25,max=40;
variogram(ARGILLA): 298.01 Nug(0) + 5486.25 Exp(7377.63);
mask: 'acv.asc','fill_telese20.asc','north.asc';
method: gs;
predictions(ARGILLA): 'Pred_GS_clay.asc';
set nsim=100;
```

Simulated maps are imported in MatLab with the ImportAsciiRaster function:

```
>> [Z R] = ImportAsciiRaster(NaN, 'r', 'd');
```

which creates a three dimensional double array of size [nrows, ncols, nsimulations] = [1200, 1500, 100]. The mean value for each pixel is calculated, shrinking the stack of GS maps into a single average 2–D [1200, 1500] simulated clay map.

6.2.2 Soil colour mapping

6.2.2.1 Regression kriging of colour triplets

An example of regression kriging on quantitative soil colour components can be found in Hengl and Langella [2007]. Unfortunately does not exist a computer program in which exhaustively solving RK technique. For instance in R it is possible to separately solve the trend component within its basic environment, while the residual stochastic component can be analysed by internal additive packages, such as geoR or GSTAT.

Table 6.1: Summary of colour components submitted to statistical analysis

Munsell	RGB	HSI	HSI_{RGB}
Hue	Red	Hue	H_{RGB}
Value	Green	Saturation	S_{RGB}
Chroma	Blue	Intensity	I_{RGB}

Here the analysis is carried out in MatLab environment. I have nine quantitative colour components (Tab. 6.1) pertaining to three colour systems computed in § 5.1.2 from Munsell HVC. For each colour component is pointed out a multilinear regression analysis on auxiliary information followed by a variographic analysis on residuals (see Eq. 6.1). The matching table is imported in MatLab[1]:

```
%LOAD TABLE 'col_rk'
    cd('C:\Dottorato\PEDOMETRICS\DATA\UOT\RegrKr\COLOR\Layers\MatLab')
    load col_rk.mat
```

Then coordinates, target and predictors data are extracted:

```
%PREPARE DATA
    %--coordinates
    xy = cell2mat(col_rk(2:end,[24:25]));
    %--dependent: [R G B HUE SAT INT Hrgb Srgb Irgb]
    dep_h = col_rk(1,[96:104]);
    dep = cell2mat(col_rk(2:end,[96:104]));
    %--independent
    indep_h = col_rk(1,[38:57 90:93 105:127]);
    indep = cell2mat(col_rk(2:end,[38:57 90:93 105:127]));
```

Now the system is ready to start the stepwise multiple regression using the 'stepwisefit' MatLab function. The $p+1$ drift model coefficients $\hat{\beta}_{gls}$ are computed, and used to obtain the $(\mathbf{s} - \mathbf{q_0^T} \cdot \hat{\beta}_{gls})$ residuals (Eq. 6.1):

```
%STEPWISE REGRESSION
    %--current colour component [1,9]
    dependent = 9;
    %--stepwisefit
    [n,n,n,inmodel,stats,n,n] = stepwisefit(indep,dep(:,dependent));
    %--organize output of statistical analysis
    inmodel = find(inmodel==1);
    stats.inmodel = inmodel;
    clear n inmodel
    %--initialize vector of predictions PLUS intercept
    pred_regr = zeros(size(dep(:,dependent),2),1) + stats.intercept;
```

[1]Note that % symbol codify for non-command lines, such as eplanatory text.

```
    %--compute predictions: 'pred_regr'
    for pred = 1:size(stats.inmodel,2)
        pred_regr = pred_regr + stats.B(stats.inmodel(pred))*
         *indep(:,stats.inmodel(pred))';
    end
    clear pred ans
    pred_regr=pred_regr';
    %--compute residuals
    res = dep(:,dependent)-pred_regr;
```

Experimental variograms are computed with Eq. 4.6 for both the colour components and the residuals of colour components after detrending them with regression analysis.

```
%VARIOGRAPHY
    %--selected colour component
    [hc_d,garr_d]=semivar_exp(xy,dep(:,dependent));
    %--residuals selected of colour component
    [hc,garr]=semivar_exp(xy,res);
```

For each colour component three plots are performed: (i) the semivariogram of the colour component at hand, (ii) the scatter diagram of the dependent variable on the MLR stepwise fit, and (iii) the semivariogram of residuals. The MatLab code is:

```
%PLOT
    %--fig_1
    figure(1)
    scatter(hc_d,garr_d,'r+');
    title(dep_h(dependent),'FontWeight','b', 'FontSize',20)
    ylabel('Variance','FontWeight','b', 'FontSize',18)
    xlabel('Distance (m)','FontWeight','b', 'FontSize',18)
    %--fig_2
    figure(2)
    scatter(hc,garr,'r+');
    title(strcat('RES (', dep_h(dependent), ')'),
    'FontWeight','b', 'FontSize',20)
    ylabel('Variance','FontWeight','b', 'FontSize',18)
    xlabel('Distance (m)','FontWeight','b', 'FontSize',18)
    %--fig_3
    figure(3)
    scatter(pred_regr,dep(:,dependent))
    title(strcat('Stepwise - MLR (', dep_h(dependent), ')'),
     'FontWeight','b', 'FontSize',18)
    xlabel('Predictions','FontWeight','b', 'FontSize',18)
    ylabel('Measurements','FontWeight','b', 'FontSize',18)
```

6.2.2.2 Ordinary kriging of PDI

A preliminary analysis in EDASS (5.1.2) confirmed that pondering PDI^{+} exhibit a good spatial structure. A variographic analysis and kriging is computed in ISATIS and GSTAT for comparison purpose.

6.3 Artificial Neural Network

In their work McKenzie and Ryan [1999] reported the statement that the use of suitable environmental variables is more important than the choice of prediction method. Here I oppose the contrary assertion, because even though a plethora of suitable auxiliary data is used, I generally observe better performance of non-linear models build with ANNs on linear models like regression and kriging.

Neurocomputing is performed in MatLab environment by means of a set of scripts suitably written to shape the Neural Network Toolbox and MatLab Base functions to accomplish analysis on pedological data.

6.3.1 Clay content

From McKenzie and Ryan 1999: I start from following statement to prove the contrary, i.e. NN are more performing then generalized linear models... Neural nets were unsatisfactory because of the difficulty of interpretation and requirement for specialised skills. They also noted that the use of suitable environmental variables was more important than the choice of prediction method. The generality of these results is not clear because soil data are often noisy and conditional relationships appear more common.

6.4 Spatial analysis of clay using Fuzzy logic

Fuzzy logic is used in § 3.2 to implement a heuristic fuzzy segmentation of the landscape in 15 predefined geomorphological elements. The role of landform segmentation is to provide experimental facets that can be used to stratify the study site into functionally distinct units and compare the values of a soil property across these units.

As an example of possible applications, two-dimensional spatial occurrence of clay content over the complex Telese Valley landscape is developed. Clay particles undergo lateral surface and subsurface traslocation according to gravitational field and water kinetics. Clay particles are expected to be depleted from facets such as shoulders (surface erosion, subsurface eluviation)

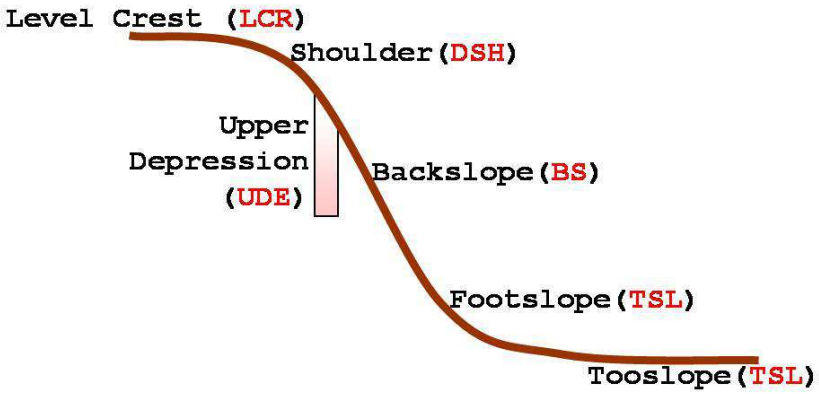


Figure 6.2: Simple soil-landscape model

and backslopes (mechanical transportation), and to accumulate in facets as upper depressions, footslopes (colluvium), and toeslopes (alluvium) [Park et al., 2001].

A one-way analysis of variance (ANOVA) is accomplished in SPSS environment. One-way ANOVA is used to test for significant differences between means of two or more independent groups, thus it is used to discriminate for average clay content amongst landform facets. In figure 6.2 a simple hill slope soil-landscape model is depicted, representing the facets found for locations stratified as follows:

1. *feature space*: clay content as target variable, FLFS as auxiliary map;
2. *locations*: all samples analysed by pipette method (about 70 profiles);
3. *pedological domain*: pondering along whole soil profile.

Chapter 7

Applied Results and Discussion

7.1 Clay content

Clay content (%) is put in a spatial framework and is analysed by means of a range of data analysis methods. These models of spatial inference are (i) one-way ANOVA, (ii) multiple linear regression, (iii) universal kriging, and (iv) finally artificial neural networks.

ANOVA The one-way analysis of variance is conducted to test for trends across the landform facets segmented by the heuristic fuzzy logic procedure pointed out in § 3.2. The conditional probability that all group means are different is 0.050 (Table 7.1). This implies that one-way ANOVA rejects the null hypothesis that the group means are equal.

Table 7.1: One-way ANOVA on pondering clay across FLFS facets

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	3641,725	11	331,066	1,814	,050
Within Groups	67514,410	370	182,471		
Total	71156,136	381			

Once differences amongst facets are not a result of chance, it is investigated if average amounts measured are prone to be explained by driving forces mechanics. In Fig. 7.1 it is possible to evaluate how amount of pondering clay content changes across geomorphological elements.

As pointed out before (pag. 54), there is an expectation on how clay particles distributes across landscape, as gravitational field exerts directly but

mostly indirectly through water movement an important role. To understand if FLFS facets can give reason of lateral redistribution of clay particles, average clay computed for each geomorphological element should be compared with expectation about that.

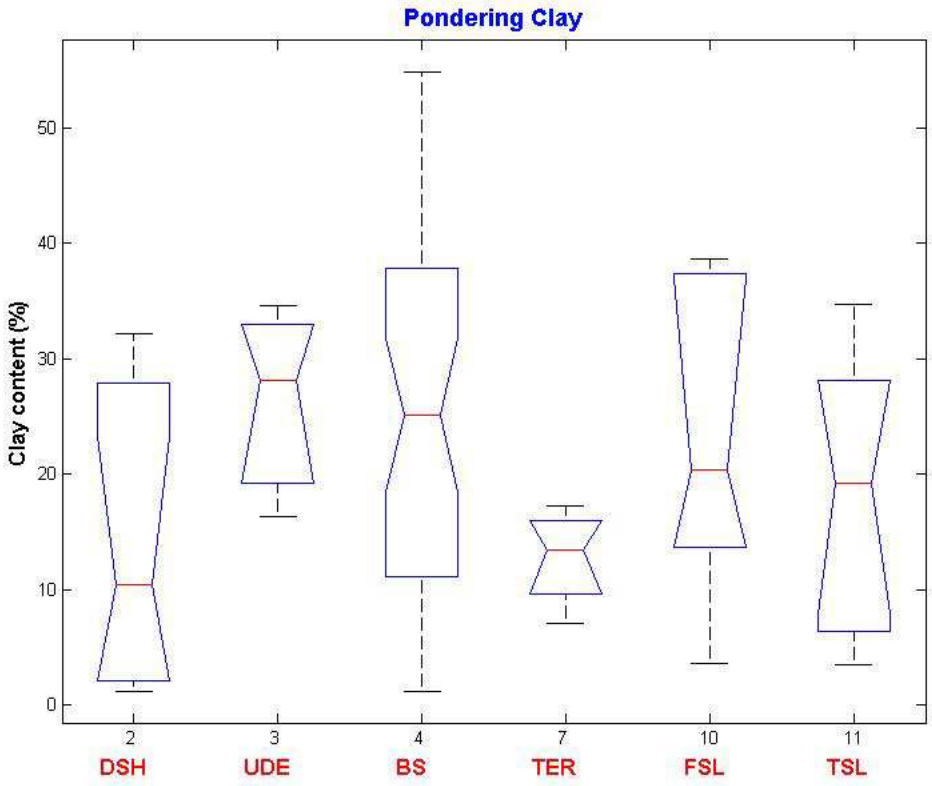


Figure 7.1: Clay content across geomorphological elements.

Expectation is confirmed on shoulders (DSH), upper depressions (UDE), and terraces (TER) inasmuch it is found higher clay content (%) across UDEs and lower mean values on the other two.

Situation is more complicated on backslope (BS), footslope (FSL), and toeslope (TSL) facets for different reasons. Backslopes have a consistent variability as depicted by boxplot; even if any outlayer can be found it is quite evident that other source of variability occur and should be investigated further.

Footslope and toeslope cases are instead less clear because those geomorphological elements relies in positions where landscape was obliterated by much varying soil-forming factors over time.

The result is that it is not possible to find justifications for clay redistribution across BS, FSL and TSL elements.

Multiple Linear Regression Spatial analysis of clay content is here investigated by means of a multilinear equation in which DEM derived Land Surface Parameters (LSP) and remote sensing imageries are employed.

The main difference between this technique and ANOVA relies on the configuration of the explanatory variables. Both LSPs and FLFS facets are continuous in geographic space, but while LSPs are measured on a continuous scale in the attribute domain, FLFS procedure (§ 3.2) only produces a nominal variable. At most geomorphological elements could be organized in a ordinal variable when facets are ordered in the sense of a specific problem at hand, as was done for the ANOVA sub-case study.

Regression analysis is run with the assistance of *ANNvsREGR* MatLab script (appendix C), which also performs neurocomputing on the same subcase study. Figure 7.2 depicts clay pattern (see colour variation) along profiles in locations stratified as specified in § 5.1.1.

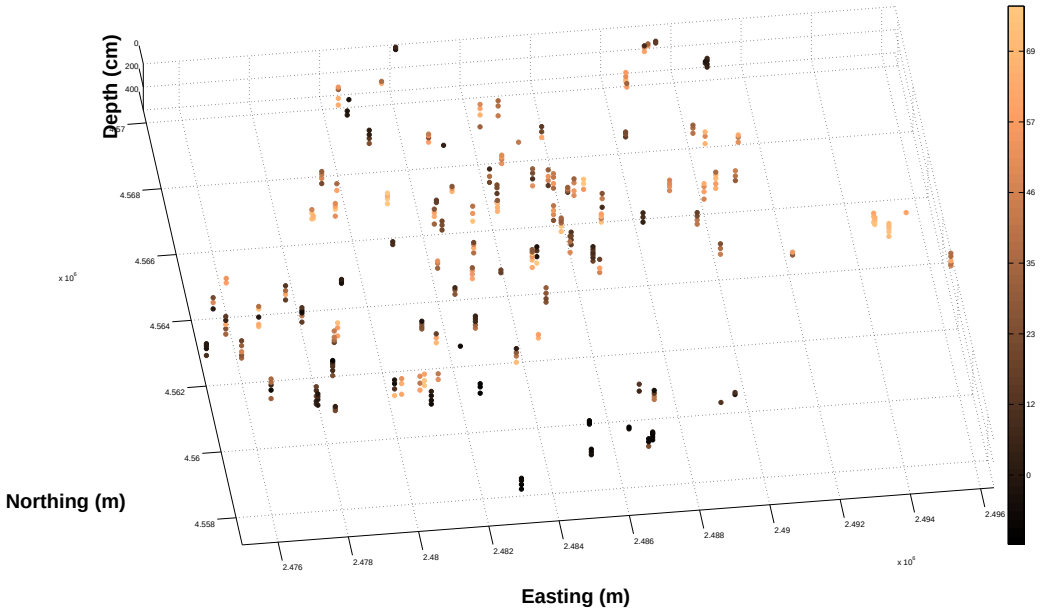


Figure 7.2: Clay content in 3-D space. Color bar indicates clay values. Note that nearby vertically aligned circles belongs to the same soil profile.

The general settings for multiple regression is pointed out in § 4.1. Taking advantage of *ANNvsREGR* script, few manual selections enabled the outlin-

ing of regression modeling (see App. C):

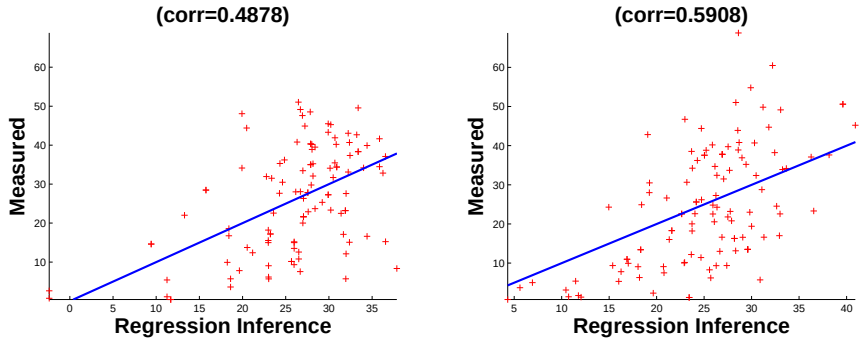
-1. `col_target = 15 and Threshold = 0.5`.

-2. `subsets = [0.70 0.30]`.

The resulting multi linear equation is reported below

$$\begin{aligned} CLAY &= -0.4241 + 0.0816 \cdot ASP - 0.4071 \cdot NDVIS5 \\ &\quad - 0.2971 \cdot NDVID16 \end{aligned} \tag{7.1}$$

It is highlighted the role of MODIS NDVI (NDVIS5, NDVID16) and of ASP (Tab. 5.1) in explaining the spatial distribution of clay in soils. Maybe the vegetation cover type codified by multitemporal-enhanced vegetation indices derived from the MODIS imagery takes partly account for soil types underneath, and consequently of the state and distribution of this soil physical property in space. However this model shows bad goodness of fit as revealed by scatter diagram of Fig. 7.1 a.



(a) (b)

Figure 7.3: Clay predictions are made with (a) Eq. 7.1, and (b) Eq. 7.2

One limitation of model in Eq. 7.1 is that it does not consider a predictor to account for vertical variability of clay, whose pattern inevitably exhibit a 3-D spatial variation (Fig. 7.2). For this reason a new model is built, considering the horizon number as an ordinal variable. The stepwise fitting produced the following model:

$$\begin{aligned} CLAY &= -0.3021 + 0.1346 \cdot ASP - 0.341 \cdot NDVIS5 \\ &\quad - 0.176 \cdot NDVID16 + 01531 \cdot HOR \end{aligned} \tag{7.2}$$

Though the use of more auxiliary maps even to account for vertical variation along soil profile, regression modeling demonstrate a low level of performance in analyzing the spatial variation of clay content.

Universal Kriging

7.1.1 Regression kriging of colour triplets

Munsell HVC system was designed to arrange colours according to equal intervals of visual perception, thus the primary advantage of the Munsell system is its ease of interpretation. However, Munsell HVC coordinates are based on subjective perception and comparison, thus the system is not uniform from quantitative viewpoint. Transformations of hue, value and chroma in more colour systems are performed in § 5.1.2.

Spatial analysis of RGB, HSI and HSI_{RGB} colour systems by means of regression kriging is tried in § 6.2.2.1. Unfortunately neither the semivariogram of colour triplets nor the semivariogram of residuals show good spatial structures to justify the use of a geostatistical inference technique for mapping purpose (Figures 7.1.1, 7.1.1 and 7.1.1).

7.1.2 Ordinary kriging of PDI

My scores construct with tales of codification/decodification, that is decodec (i) variography and kriging -investigate relationship between PDI and Al05Fe, to state robustness of decodec -compare estimated out-of-sample PDI with Al05Fe to state robustness of prediction of degree of pedogenesis in landscape.

(i) multiple regression; (ii) ANN

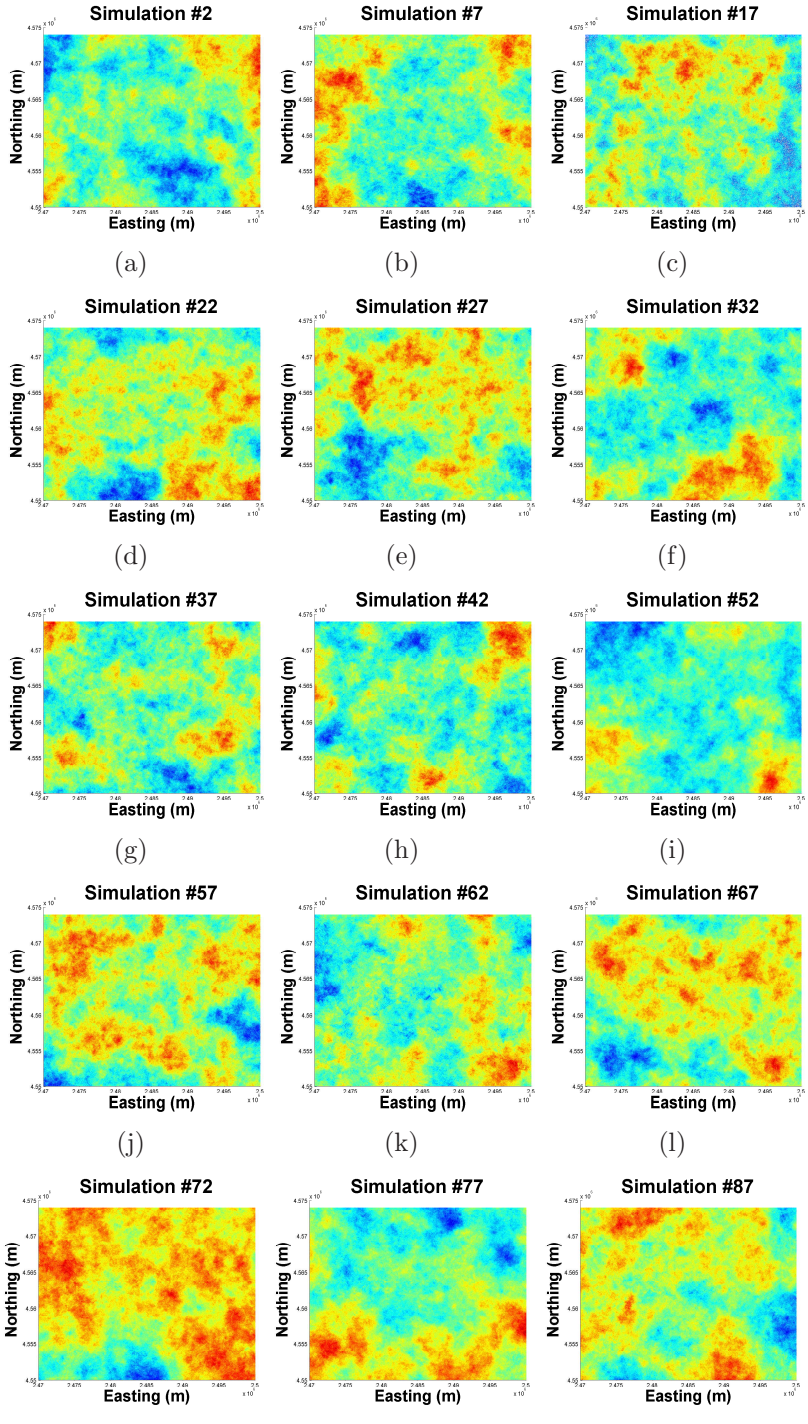
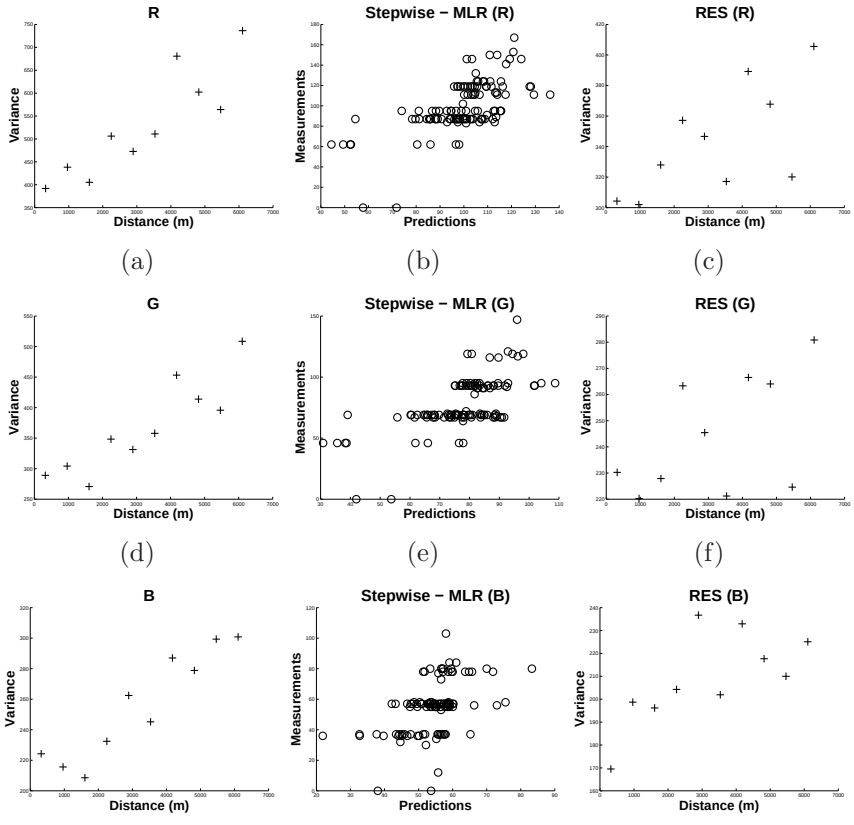


Figure 7.4: Selected outcomes of clay geostatistical simulation



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 7.5: Spatial analysis of RGB colour system

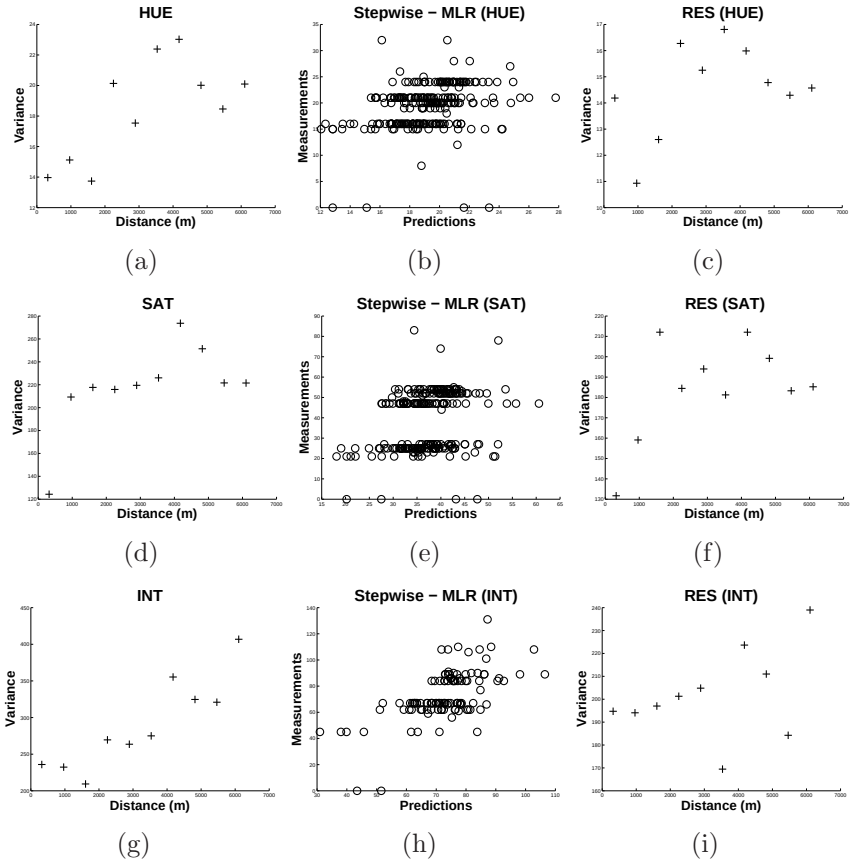


Figure 7.6: Spatial analysis of HSI colour system

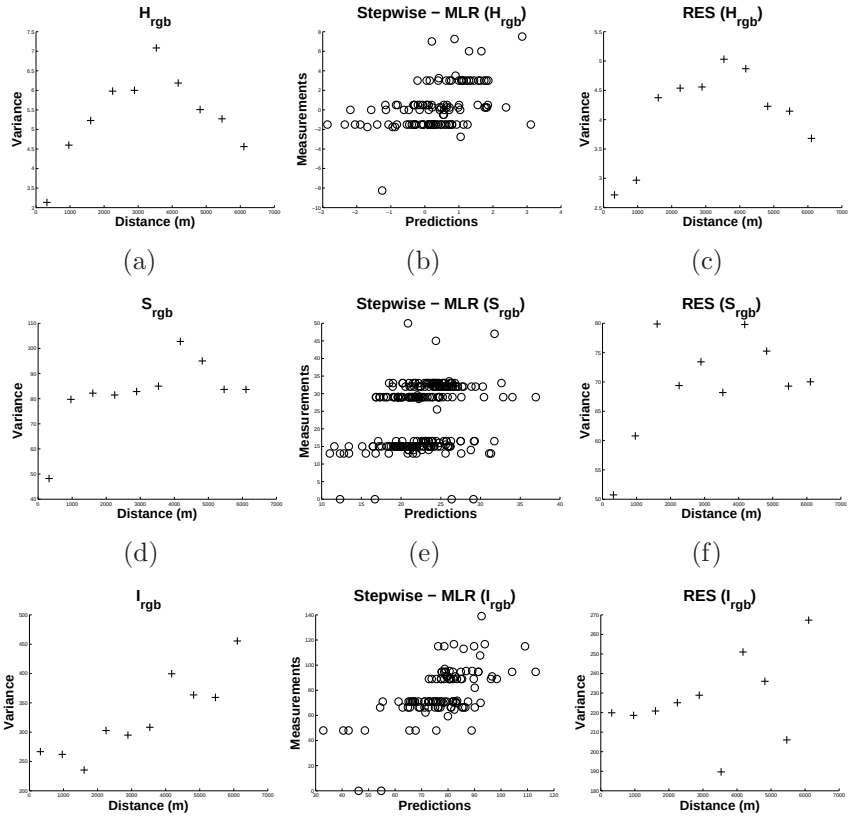


Figure 7.7: Spatial analysis of HSI_{RGB} colour system

Part II

Space time rainfall analysis with ANNs and Geostatistics: a quali-quantitative inference system

Chapter 8

Rainfall Analysis

8.1 Introduction

Water is a prominent environmental factor that regulates natural ecosystems by complex functions operating both within and between biotic and abiotic compartments. Precipitation affects water availability; its geographical distribution is influenced globally by the general circulation of the atmosphere, the proximity to large water bodies like oceans and great lakes and by topographical barriers. Despite the importance of the way in which water distributes on earth surface, gauged measurements of rainfall fields are not adequate to realistically represent such a high space-time stochastic phenomena and do not provide data of good quality (for quality problems/controls of climatic dataset see Peterson et al. [1998]). The outcome of this situation is indeed regrettable when rain information is needed to address environmental vital functions for humans such as watering, agriculture, breeding, forestry and natural hazards among others. On the other side if high resolution precipitation data can be available than many opportunities in taking successfully care of aforecited environmental functions open when using either physically based hydrological models in specific pieces of a landscape or spatially distributed data driven models.

Unfortunately the space-time distribution of rainfall is decidedly nontrivial and consequently it is nigh impossible its description by means of a white box model, that is by numerically solving a complex set of differential equations. On the contrary, statistic dynamical models provide a useful framework within which analyze rain data. A summary of selected studies published from 1990 to nowadays and involved in the prediction (excluding the forecasting) of precipitation in space and/or in time using merely gauged measurements is given in Tab. 8.1.

Table 8.1: Details of selected studies from 1990 to 2008 involved in precipitation prediction from points gauge measurements

No.	Source	Models of inference	Predictors	Rain gauges	Temporal extent	Resolution (time/space)	Study area	Indicators of performance
1	Hevesi and Flint [1992]	OK, OCoK	elevation	42	–	average year	Yucca Mountain, Nevada (4,200,000 km^2)	r
2	Daly et al. [1994]	PRISM, OK, RK, OCoK	elevation	52	1982-1988	month-year/$5'$	Willamette River basin, Oregon (29,000 km^2)	MAE, MBE
3	Bacchi and Kottegoda [1995]	kriging	—	71	1926-1967	—	Lombardia, Italy (3,000 km^2)	r
4	Johnson and Hanson [1995]	regression	elevation, geographic location	46	1968-1975	day-month/$1km^2$	Idaho, USA ($234km^2$)	R^2, RMSE
5	Martnez-Cob [1996]	OK, OCoK, RK	elevation	182	10-20 years	average year/$5km^2$	Aragon, Spain (47,000 km^2)	MAE, MSE
6	Frei and Schar [1998]	advanced IDW	—	6600	1971-1990	day/$25km^2$	European Alps	–
7	Pardo-Igzquiza [1998]	Thiessen poligon, OK, OCoK, KED	terrain parameters	51	20 years	average year/$4km^2$	Guadalhorce river basin, Spain (2864 km^2)	ME, MSE, MSSE
8	Goodale et al. [1998]	regression, modified IDW	elevation, geographical coordinates	618	1951-1980	average month/1 km^2	Ireland	r, MBE, MAE
9	Prudhomme and Reed [1999]	OK, RK	terrain parameters	1003	10 years	median of annual maximum daily rainfall/?	Scotland	ME, RMSE, MBE
10	Xia et al. [1999b]	Barner, Cressman, OI, AA, IDW, regression	elevation	50	1991-1995	month/1km	German	R^2, MAE
11	Goovaerts [2000]	Thiessen poligon, IDW, OK, SKl, KED, Co-CoK	elevation	36	1970-1995	month-year/1 km^2	Region of Portugal (5,000 km^2)	MSE
12	Kyriakidis et al. [2001]	SK, RK	low atmosphere parameters and elevation	77	Nov1981-Jan1982	daily seasonal average/$1km^2$	Northern California (108,000 km^2)	cross-validation error
13	Antonic et al. [2001]	hybrid: ANN and OK	elevation, geographic location, dummy time	127	1956-1995	month/300 m	Croatia (56,538 km^2)	r
14	Brunsdon et al. [2001]	GWR	altitude	10925	1961-1990	average year/$1km^2$	Great Britain	R^2, cross-validation error
15	Shen et al. [2001]	hybrid: IDW and nearest neighbor	—	927	1961-1997	day/polygons	Alberta, Canada	RMSE, MAE, MBE
16	Brown and Comrie [2002]	hybrid: regression and IDW	terrain parameters	572	1961-1990	winter/$1km^2$	Arizona and New Mexico, USA	D, PSE, RMSE, MBE, R^2

17	Drogue et al. [2002]	PLUVIA, KED, Co-CoK	terrain parameters	200	1971-1990	average month-year/100m	Northeast France (30,000 km^2)	adj-R^2, MAE, MBE
18	Vicente-Serrano et al. [2003]	global, local, geostatistical and mixed	geographic and terrain parameters	380	1950-2000	year/—	Middle Ebro Valley (Spain)	MBE, RMSE, MAE, EF, D
19	Marquinez et al. [2003]	regression	elevation, distance from coastline and west, slope	117	1966-1990	month/200m	Cantabrian Coast, Spain (10,590 km^2)	adj-R^2, MAE, MBE
20	Gyalistras [2003]	AURELHY, regression, IDW	elevation	673	1901-2000	month/$5km^2$	Switzerland	RMSE, MAE, MARE, MRE, r, ...
21	Apaydin et al. [2004]	IDW, local/global polynomials, spline, SK, OK, UK, CoK	elevation	117	1971-1999	year/0.01°	Southeastern Anatolia, Turkey	ME, MAE, MRE, RMSE
22	Lloyd [2005]	MWR, IDW, OK, SKl, KED	terrain parameters	3000	1999	month/661.1m	Great Britain	cross-validation error
23	Oettli and Camberlin [2005]	RK	terrain parameters	305	1950-1990	average month/30″	Southern Kenya and NE Tanzania	r, R^2, RMSE, LEPS, SK
24	Pardo-Iguzquiza et al. [2005]	UK	topography, latitude, wind direction	184	July 1999	month/1°	West Africa (22° x 22°)	AIC, ME, RMSE, MSSE, r
25	Diodato and Ceccarelli [2005]	IDW, regression, CoK	elevation	20	40 years	average month-year/0.5 km	Sannio Mountains, Italy (1,400 km^2)	MSE, RMSE
26	Sicard and Sabatier [2006]	local PLS1 regression	sea surface temperature	7	1950-1984	month/—	Nordeste, Brazil	MSE, SSE
27	Celleri et al. [2007]	regression	elevation	23	1975-1989	month/—	Paute Basin, Ecuadorian Andes	—
28	Attorre et al. [2007]	D-IDW, UK, ANN	geographical location, terrain parameters	201	1955-1990	month/200 m	Lazio region, Italy (17,200 km^2)	RMSE
29	Guler et al. [2007]	regression	geographical and climatical variables	11	?	month/250 m	Samsun, Turkey	r^2
30	Ninyerola et al. [2007]	hybrid: regression and splines	terrain parameters	2825	1950-1999	month/200 m	Iberian Peninsula (583,551 km^2)	RMSE, r^2
31	Carrera-Hernandez and Gaskin [2007]	OK, OKl, KED, KEDl, BKED	elevation	200	June-1978, June-1985	day/200 m	Basin of Mexico	cross-validation error
32	Freiwan and Kadioglu [2008]	OK	statistical moments	16	1971-2000	year/contours	Jordan	skewness, CV

It is shown that the rainfall modeling is addressed by means of a plethora of models that can be sorted from statistical viewpoint in four main groups. First, there are the commonly used techniques such as Thiessen polygons [Pardo-Igzquiza, 1998], inverse distance weighting (IDW, Shen et al. [2001]) and polynomials or splines [Apaydin et al., 2004].

Second, multilinear regressions [Goodale et al., 1998, Brunsdon et al., 2001] are either implemented stand alone using terrain parameters and geographical coordinates [Marquinez et al., 2003] or integrated in hybrid frameworks as in Brown and Comrie [2002] with IDW.

The third group of models is based on linear geostatistics [Journel and Huijbregts, 1978, Goovaerts, 1997, Wackernagel, 2003] with univariate [Freiwan and Kadioglu, 2008] or multivariate models that address the rainfall space-time pattern analysis by using different algorithms of kriging as the cokriging [Hevesi and Flint, 1992], the kriging with external drift [Lloyd, 2005] and the regression kriging [Prudhomme and Reed, 1999, Kyriakidis et al., 2001].

Finally, the group of soft computing [Tsoukalas and Uhrig, 1997] is merely represented by the neurocomputing technology [Haykin, 1998, Bishop, 1995] which the work of Antonic et al. [2001] belongs to.

The size of space-time support is a key setting for dealing with high resolution precipitation datasets. Disregarding the type of statistical model adopted to make predictions, the temporal units most often used range from instantaneous day [Johnson and Hanson, 1995, Hunter and Meentemeyer, 2005] to year [Daly et al., 1994, Goovaerts, 2000] passing through the more frequent monthly unit [Ninyerola et al., 2007]. Models that make use of temporal units averaged on long time series are also developed such as the works by Martnez-Cob [1996] and Oettli and Camberlin [2005].

The regionalization of points rain signals generally decreases the spatial resolution as the timescale becomes finer. Indeed papers that deal with daily temporal units address the spatial interpolation at a coarser resolution, for instance at 25 km^2 in Frei and Schar [1998] using the raster format or within large polygons [Shen et al., 2001] considering the vector data, with few exceptions as in Carrera-Hernandez and Gaskin [2007] where geostatistical interpolations are made on a 200x200 m^2 grid.

To cope with high spatial resolution Drogue et al. [2002] employed averaged monthly rain data in the PLUVIA framework getting a mean bias error ranging from -34.0 to 3.7 and a mean absolute error between 5.3 and 15.2 with a cell size of 100 meters. Attorre et al. [2007] accomplished as well a neurohydrological model based on the single MLP back propagation neural network template proposed by Antonic et al. [2001], achieving high resolution precipitation data, namely at monthly timescale on a 200 meters

squared grid (RMSE is between 168 and 205).

Artificial neural network modeling is very exiguous when compared with the plenty of the other groups of models. Different justifications could be drawn: one exogenous explanation is that the spatial interpolation at gridded points is expanding the use of continuous covariates obtained by whatever meteorological sensing probe, such as from ground radars [Xiao and Chandrasekar, 1997, Chiang et al., 2007] or from orbital satellites [Tsintikidis et al., 1997, Bellerby et al., 2000], relinquishing or relegating to calibration/validation the use of gauged observations [Joss and Lee, 1995, Rasmusson and Arkin, 1993].

Among intrinsic causes, ANNs require advanced statistical and computational skills to be implemented and evidently localization, density, discontinuity and reliability of measurements at raingauge networks don't offer neither the actual intrinsic spatial variability of meteorological events [Bacchi and Kottegoda, 1995] nor the unbiased huge amount of cases necessary for training a neural network.

Although space-time neurohydrology is rather developed for instance in the modeling of rainfall-runoff processes [Dawson and Wilby, 2001, Kumar et al., 2005, Jeong and Kim, 2005], applications that embed the spatiotemporal domain in an integrated framework are very limited in the case of rainfall prediction. Only the works by Antonic et al. [2001] and Attorre et al. [2007] are involved in the space-time modeling of precipitation from gauged networks by means of neurocomputing.

Nomenclature

AIC	Akaike information criterion
Anchor	anchorage gauges
ANN	artificial neural network
ASP	aspect
BAGNET	bootstrap aggregated neural network
Bootsample	resampling with replacement (bootstrapping)
D	Willmott's agreement index
EAST	easting
ELEV	evelation
FFBP	feedforward back propagation
FFNN	feedforward neural network
I/O	input/output
MPL	multilayer perceptron
MSEPE	percentiles of mean squared error

NNs	neural networks
NORTH	northing
PCR	principal component regression
Randsample	resampling without replacement
RELD	relief distance
sBN	BAGNET selection
SEAD	sea distance
SMAPE	standardized mean absolute percentage error
SN	single network
Ta	rain matrix after imputation
Tb	rain matrix before imputation
Ti	initial rain matrix
Time	time variable
YDP	yearly percentage difference index

8.2 Aim and hypothesis

The primary aims of this paper are addressed in the following § 8.3, that is the creation of a complete rain database by infilling gaps, the predictions at high resolution, the creation of a slim integrated framework for both space and time dimensions, the recognition of the best NN settings to make predictions in each different condition, the calibration of NNs in case of small rain dataset and the accounting for spatial intermittency of rain distribution by means of a geostatistical filter.

Reliability, completeness and representativeness of precipitation time series are essential for any model calibration to give accurate results. This study has the main objective to tackle high resolution precipitation interpolation in space-time domain, but first an automatic procedure is pointed out for infilling missing rain data through a regression imputation technique [Xia et al., 1999a, Ramesh and Chandramoulia, 2005, Coulibaly and Evora, 2007].

From Tab. 8.1 it is clear that most statistical models deals essentially with the spatial dimension and are coerced in reiterating model calibration for each considered time element. Artificial neural network models embed time dimension in an integrated framework with geographical space and therefore exhibit the ability to simultaneously interpolate rainfall variable in space and in time. Further, models that achieve precipitation predictions at finer spatiotemporal scale are very exiguous and accordingly here a fine resolution neurohydrological analysis to calibrate a precipitation model is attempted with a 20x20 m^2 grid at a timescale of 10 days, that is the foundamental

temporal unit has length 10 days (henceforward this unit is called 'decade' without mistaking it with the 10 years length).

The main restriction of neurocomputing is the need for large amount of data for training phase so the bootstrap aggregating technique [Breiman, 1996] is appealing to build a model of inference in case of few calibration data. Cilento subarea is selected from domain wide case study to test the bagging forcefulness in yielding quite the same results as the full extent study area in terms of scale and quality assessment of rainfall predictions.

The regionalization of rainfall values by means of NNs simulate not so likely the spatial intermittency of rain catchments at gauged network, and this fallacy might be greater as ungauged pixels are further away from measured locations. A geostatistical filter based on indicator kriging is fulfilled to cope with the qualitative (binary) occurrence of rainfall in space, which is able to model the spatial distribution of non rainy pixels.

8.3 Methods and Data

8.3.1 The study area

Situated in the Southern Italy, Campania (14000 km^2) is prevalently a hilly region (50%) of complex terrain lying between 13°45′ E and 15°49′ E, 39°59′ N and 41°31′ N.

It can be structurally divided into three clearly defined zones stretching from Northwest to Southeast parallel to the coastline. Inland rise the Campania and Lucanian Apennines with a coverage of about 34% of total area. Along the coast lie the Campano Preapennines, lower in height and of volcanic origin (the extinct Roccamonfina volcano, Campi Flegrei and Vesuvius) or limestone (Lattari mountains). To these two parallel structures a third one can be added: the discontinuous and much less extensive band of offshore islands of volcanic origin (Ischia, Procida, Vivara and Nisida) or limestone (Capri). The Campania Apennines consist of an irregular range of low mountain groups, broken at intervals by intermontane hollows.

Elevation ranges from zero to 1904 meters above mean sea level, whereas gauges are within zero and 1211 meters.

8.3.2 The independent variables

The input space define each space-time element by a 11 dimensional vector in which 6 components are geospatial covariates and 5 components are variables that take into account the temporal variability of precipitation.

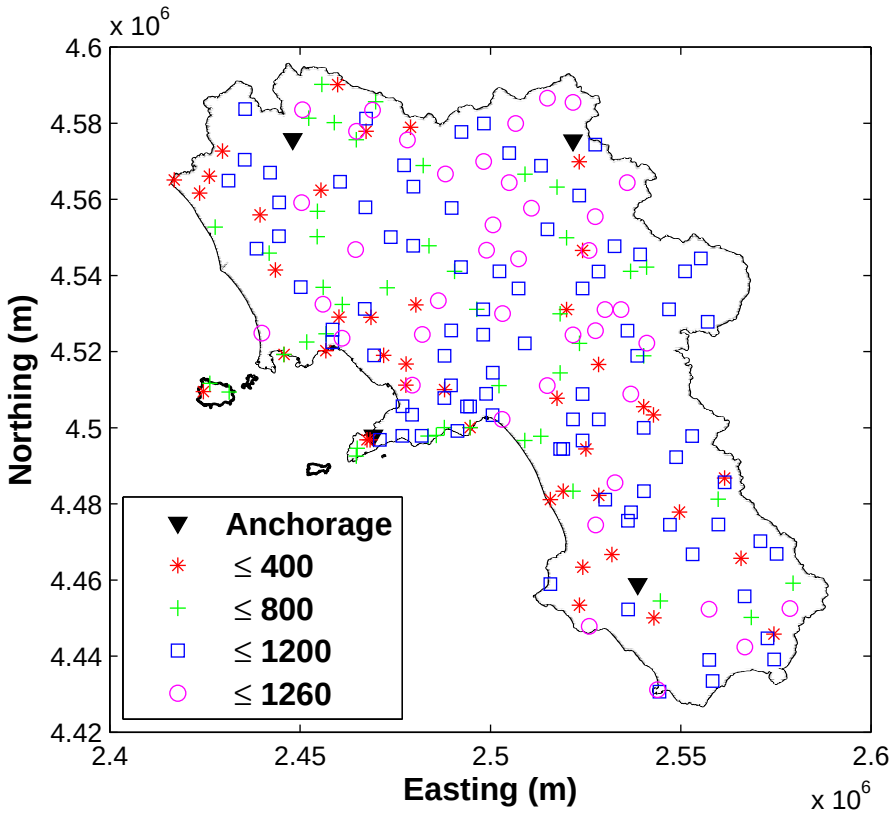


Figure 8.1: Geography of raingauge network. The symbol types, except the downpointing triangle that represents the four anchorage stations, identify the four classes of gauges as they are grouped based upon measurement gaps highlighted in Fig. 8.5.

It was found that using four anchorage stations (**Anchor**) as time dependent variables (see downward pointing triangles of figure 8.1) gave the best result considering the terrain complexity of Campania Region. It is noteworthy for the reason that the anchorage gauges have complete time series and quite cover the outermost of the study area: the inner continental zone, the coastal band and the northern and southern areas.

Time is the fifth time dependent variable and is codified using the ordinal numbers of the 10 days elements in a year. In other words a year is composed of 36 decades; the instance of the 36 number-of-decade in a year is presented to a generic neural network as a particular function f(.) of inverse cosine:

$$time = cos^{-1}(cos(radians(NumDec \cdot 10))) \tag{8.1}$$

where *NumDec* is the ordinal decade number vector of size 1x36 which is multiplied by 10 to enable the complete revolution on the goniometric circle

after one year is elapsed; *deg2rad* is a function converting angles from degrees to radians; *cos* and *arccos* are respectively the cosine and the inverse cosine functions.

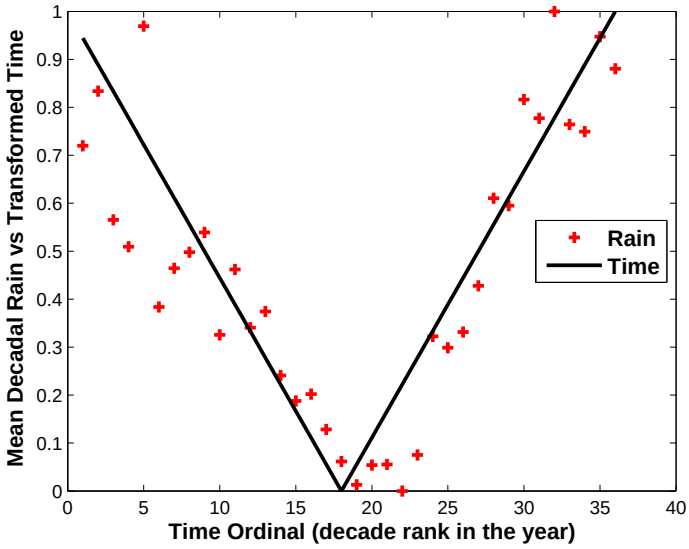


Figure 8.2: Decadal rain measurements are averaged on both the available Campania gauge network and the number of years considered in the analysis. The linear scaling in range [0, 1] of average precipitation and of transformed time variable allows the detection of a straightforward relation between the two variables (Pearson correlation is 0.9).

Antonic et al. [2001] make use of 12 nodes in the input layer just to account for dummy wise monthly time variable; this is debatable according to the acclaimed tendency in reducing the network/layer size through a pruning technique [Haykin, 1998]. Here a transformation function of the *Time* input variable is proposed to take account for the cyclic nature of annual rainfall and devote only one input node to this discrete transformed variable.

Equation 8.1 is preferred to the common sine and cosine functions adopted by other authors [Hsieh and Tang, 1998] for describing the annual cycle of time in ordinal format since the transformed time pattern takes account of the average intra annual variability of precipitation in the whole study area at decadal timescale.

In Fig. 8.2, 36 mean decadal rainfall values, that have been averaged on space (216 stations) and on time (35 years), are plotted with Time transformed by Eq. 8.1.

A digital terrain analysis is carried out in GIS environment on the 20 meters DEM of the study area to derive the set of spatial covariates: elevation (**ELEV**), easting (**EAST**), northing (**NORTH**), aspect (**ASP**), dis-

tance from the sea (**SEAD**) and distance from contiguous orographic barriers (**RELD**).

GIS returns for aspect a raster in degrees from 0 to 359.9 clockwise for those pixels with not null slope, while flat pixels are assigned an aspect of -1. The same criteria is maintained when transforming aspect with the inverse cosine function (Eq. 8.1), in which the argument of f(.) is:

$$argument = \begin{cases} 180^\circ \; and \; f(.) = -f(.), & \text{if } asp = -1 \\ asp, & \text{if } asp = [0,\ 359.9] \end{cases} \tag{8.2}$$

In the first case $-\pi$ is obtained for flat pixels, while in the second case the response belongs to the range $[0,\ \pi]$ for all other sloping pixels. This transformation is useful both to linearize circular data and to put in binary format the presence/absence of sloping for pixels. For instance Brown and Comrie [2002] propose four dummy terms to define the entire compass direction of aspect, contrariwise here a continuous variation is adopted. The SEAD represents

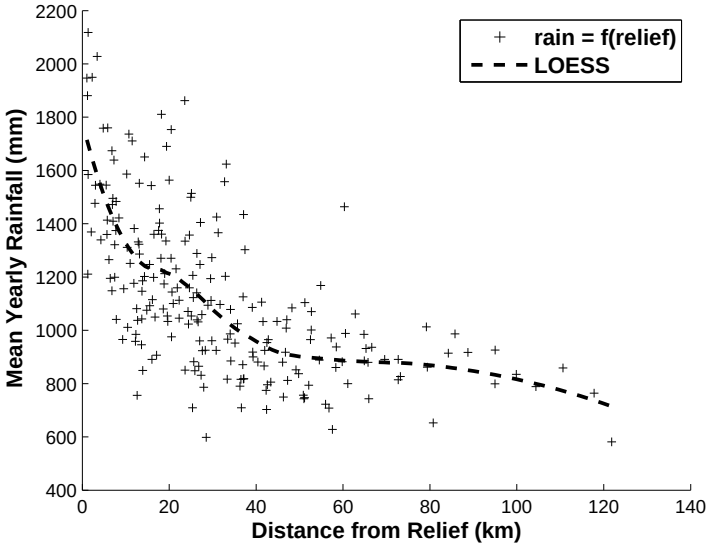


Figure 8.3: The averaged yearly precipitation is plotted as a function of the mountainous degree (RELD) of gauged pixels in Campania region. The trend is fitted by LOESS function to highlight trend in data.

a continentality index and is a buffer computed as the Euclidean distance from coastline. Another Euclidean distance but from adjacent mountains is calculated considering for each pixel the sum of minimum lengths from the three orographic levels above 500mt, 1000mt and 1500mt (8.3). RELD can be viewed as a measure of the mountainous degree of each pixel.

The values of the geospatial variables are extracted in correspondence of the gauged pixels through an ArcGIS tool suitably designed and downloadable for free at http://arcscripts.esri.com/details.asp?dbid=14826. Hence a geospatial matching table is built for the Campania raingauge network to allow the analysis of precipitation through neurocomputing.

The coordinates and the elevation are useful in determining the relative position of points in 3D geographical domain in such a manner to account for distance among measured and/or unsampled geopoints during simulation phase. Elevation, aspect and Euclidean distances are capable to account for the orographic and convectional rains.

8.3.3 The dependent variable: precipitation data

8.3.3.1 Gathering data

Rainfall measurements are administrated by the Italian Environment Protection and Technical Services Agency, APAT. Each station is manually downloaded from the SCIA website of APAT querying the web GIS archive for the stations that worked in the Italian Region of Campania during the period of time from 1860 to 2007 (Fig. 8.4).

A hand made Visual Basic application extracts the compressed files, one for each of the 222 downloaded rain gauges, and creates a unique table with rainfall and frequency values for all the stations in an automated fashion.

The period of time from 1951 to 1987 shows the higher number of working rain gauges; in this time period 6 of the 222 stations are discarded due to complete missing data. Very few stations (< 20) worked during the years 1965 and 1971, so they are removed from the time series.

The space-time rainfall database (Ti, Tab. 8.3) has dimension 35 years per 216 stations or alternatively 1260 decades per 216 stations. In figures 8.1 and 8.5 rain gauge observations are explored both in feature space (consistency) and in geographical domain (jointly localization and consistency).

8.3.3.2 Preprocessing of data: the regression imputation to fill gaps

The histogram of Fig. 8.5 indirectly quantifies the amount of missing data that the 2D target table stores together with the observed values.

Almost half the number of stations recorded more than one thousand measurements each; those stations with less than 400 time units are a priori excluded from the analysis with neural networks, and a new target matrix Tb (i.e. target before regression imputation) is created for the purpose (see

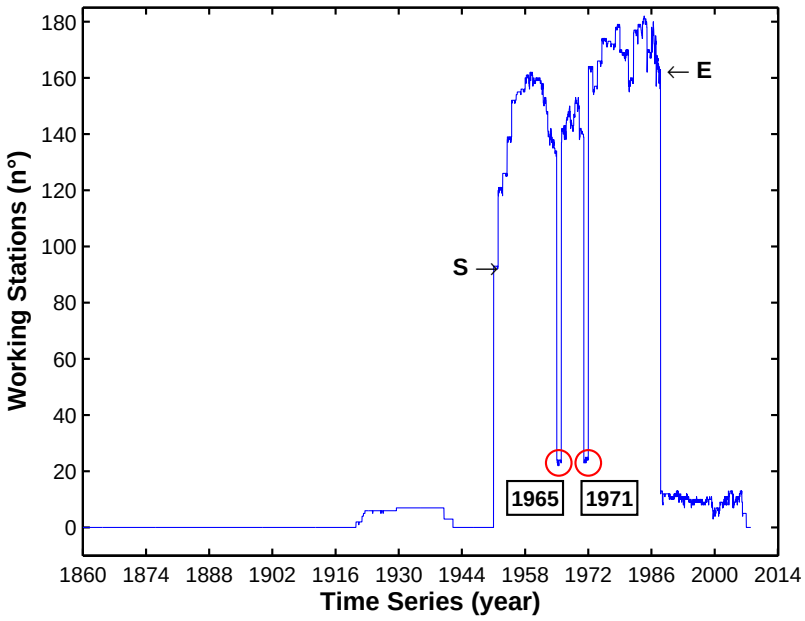


Figure 8.4: Amount of stations that worked during 1860-2007 in Campania region. The two arrows indicate the starting (S, 1951) and ending (E, 1987) date of the time series selected for neurocomputing. Inside S-E time period very few gauges measured precipitation in the years 1965 and 1971, which are removed from database. Note the straightforward reduction of the regional gauged network after 1987 to nowadays.

Tab. 8.3 for details). The pruned rainfall database is filled by means of

Table 8.3: Rainfall matrix size (stations per decades) and consistency during data preprocessing phase. The initial 2-D target array (Ti) includes all downloaded gauges. Adopting the threshold of Fig. 8.5 the Tb array (the target before regression imputation) is outlined. The rainfall matrix after gaps filling (Ta) has 4 unavailable gauges of 162 for neural network analysis because of the input anchorage stations.

Matrix	Stations	Decades	Complete Stations	Total space–time elements	Missing space–time elements	Regression Imputation
Ti	216	1260	8	272160	89433 (33%)	42742 (48%)
Tb	174	1260	8	219240	45333 (20%)	42742 (94%)
Ta	162	1260	162	204120	0	—

regression imputation task. This technique consists of substituting a missing value with a new one predicted by a stepwise multilinear regression method. The multiple linear regression equations are calculated using the significance F test of Fisher Snedecor at 0.5% level for entering (pIN) an independent

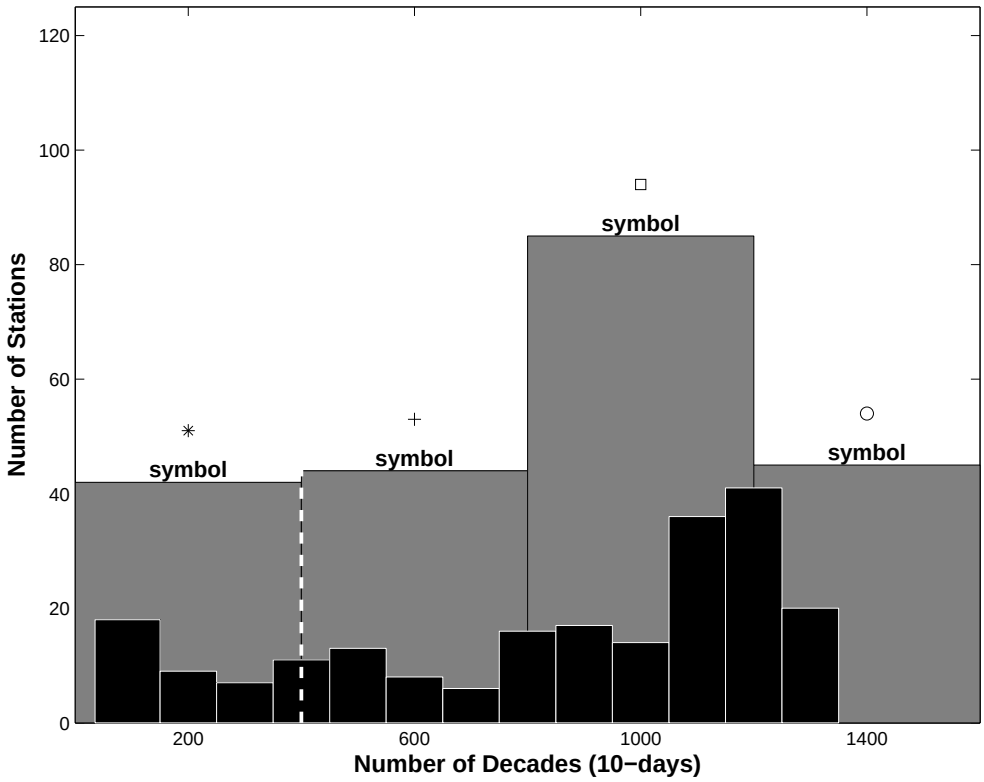


Figure 8.5: Amount of gauges that worked in the selected 35 years length time period (1951-1987). The two histograms, in grey and black color, consider the 400 and 100 decades time lag respectively. The white vertical dashed line indicates the threshold minimum number of singleton decades (=400) for a raingauge to be considered in following analysis.

variable and at 1% for removing (pOUT) a predictor already in equation [Oettli and Camberlin, 2005].

The stepwise method starts entering the most significant independent variable (lowest p-value) if and only if its p-value is less than pIN; each time a new eligible predictor with lowest p-value is entered (if p-value $<$ pIN) the stepwise procedure reexamine all the predictors in equation and the variable with the largest probability of F is removed if the value is larger than pOUT. This process continues until no variables in the equation can be removed and no variables outside the equation are suitable for entry.

An automated algorithm is built around the 'stepwisefit' function of MatLab to run the task of regression imputation, that is the computation of model parameters and the filling of gaps with interpolated precipitaion val-

ues. The code executes the following steps in sequence, looping for each singleton decadal missing value of Tb array:

1. Recognize a missing value for a given incomplete station;
2. Find all the stations that worked in that decade to use as potential predictors;
3. Build sixteen alternative models with MatLab 'stepwisefit' function involving an incremental number of predictors (each optional model is built only if there exist both at least 30 decades throughout and 6 decades per predictor);
4. The quality of each alternative model is tested on a temporal neighborhood astride the missing decade (the neighborhood thickness varies nearby the year length);
5. The model that outperforms the others on the temporal neighborhood is selected, or when a nearby temporal neighborhood cannot be selected (e.g. in case of detached missing time unit in dependent gauge) the model with the highest overall explained variance is selected.
6. Current missing value is substituted by prediction with selected regression model.

The 'best' neighbor station algorithm reported by Gyalistras [2003] for estimating missing values chooses the set of candidate gauges to use as predictors by geographical proximity to the predictand station and among them select those stations with a coefficient of determination (R^2) larger than a preset threshold.

This approach suffer of at least two limitations: first, subdomain proximal gauges might exhibit not the highest correlation coefficient than domain wise gauges as pointed out in Bacchi and Kottegoda [1995], particularly in areas of complex terrain where sharpen changes in slope orientation greatly affects amounts of rain catchments. Second, the use of stations with higher R2 induces a problem of collinearity among predictors.

In current work the statistical distance of stations is preferred to the geographical one [Ahrens, 2006] as it is possible to find very similar temporal precipitation patterns at larger distances and because more robust predictions are made. Furthermore even gauges with low r are presented to the stepwise algorithm such as to account for variance quota unaccounted for by more collinear predictors (this is mostly true in the temporal neighborhood). The complete final target matrix (Ta, 8.3) with no gaps in the data is used to make the space-time inference with artificial neural networks.

8.3.4 Indicators of performance

The performance of a model can be evaluated using one or more statistical indicators. They are based upon the error signal computed as the difference between observations and predictions. Each indicator has a primary purpose to fulfil, and therefore it can account for the model accuracy simply by a strange perspective.

Kumar et al. [2005] warned that there exist *no single definite evaluation test* to use in NN modelling, and here the approach of multi criteria assessment suggested by Willmott [1982] is fulfilled and summarized in Tab. 8.4. The performance of neural networks is evaluated on calibration data during training phase employing global statistics such as the mean squared error (MSE) used in this work.

In general artificial neural networks are just designed to minimize the global error, thus for evaluation and comparison of models of precipitation inference four often used global measures of prediction accuracy are carried out. They are the root mean squared error (RMSE), the mean absolute error (MAE), the mean bias error (MBE) and the Pearson's correlation coefficient (r). It is worth noting that only a very brief description of these statistics are provided here while remanding to other authors such as Brown and Comrie [2002], Gyalistras [2003] and Vicente-Serrano et al. [2003] for applications in precipitation domain and to Willmott [1982] for further detailed explanations about above mentioned statistical measures.

Pearson's correlation has no physical dimension, depicts the accordance of element wise trends between predicted and observed data and do not carry information about precision of fitting. The RMSE is highly influenced by larger errors and outliers, whereas MAE and MBE are less sensitive to extreme values; these three measures of goodness of fit are given in the same unit scale of the predictand variable.

The symmetrical mean absolute percent error (SMAPE) is borrowed from the field of population forecasts [Tayman and Swanson, 1999] where the asymmetrical MAPE is the most often summary measure of error. SMAPE provides a more reliable measure of error because it reduces the influence of outlying observations while using most of information about the error. Together with SMAPE statistic, the agreement index D proposed by Willmott [1981] has the advantage of avoiding the amplification of outliers. The Akaike information criterion (AIC; Akaike [1974], Anders and Korn [1999]) gives a measure of how much a statistical model is overparameterized inasmuch it is a trade off between accuracy and complexity.

Table 8.4: Statistical measures used to assess model accuracy. (O: observed value; P: predicted value; = average symbol; N = number of observations; K = number of model parameters).

Statistic	**Identifier**	**Definition**
Root mean square error	RMSE	$\sqrt[2]{\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)^2}$
Mean absolute error	MAE	$\frac{1}{N}\sum_{i=1}^{N}\lvert P_i - O_i\rvert$
Mean Bias error	MBE	$\frac{1}{N}\sum_{i=1}^{N}(P_i - O_i)$
Correlation coefficient	r	$\frac{\sum_{i=1}^{N}(O_i-\overline{O})(P_i-\overline{P})}{\sqrt{\sum_{i=1}^{N}(O_i-\overline{O})^2}\cdot\sqrt{\sum_{i=1}^{N}(P_i-\overline{P})^2}}$
Symmetrical mean absolute percent error	SMAPE	$\left[\frac{1}{N}\sum_{i=1}^{N}\frac{\lvert P_i-O_i\rvert}{\frac{1}{2}(P_i+O_i)}\right]\cdot 100$
Akaike information criterion	AIC	$-\frac{2}{N}ln\,(RMSE) + \frac{2K}{N}$
Willmott's agreement index	D	$1 - \frac{\sum_{i=1}^{N}(P_i-O_i)^2}{\sum_{i=1}^{N}\left(\lvert P_i-\overline{O}\rvert+\lvert O_i-\overline{O}\rvert\right)^2}$

8.3.5 Topology and functionality of Neural Nets (NN)

The first part of this paragraph deals with the arrangement of a prototype NN with best general performance. It will constitute the building block of three inference systems presented later on: the single network 'SN', the 'BAGNET' ensemble and a particular subset of BAGNET named 'sBN'.

The optional frameworks are calibrated and tested using three subsets of the homogeneous I/O data arrays, without the four anchorage stations just used as predictors. These are the training, validation and testing subsets which are randomly drawn with percentages of about 70, 15 and 15 respectively, using two alternative strategies: (a) the random resampling without replacement (hereafter called randsample) of the spatial elements (rain gauges) and (b) the random and redundant (that is with replacement) resampling of spatial elements through the bootstrap technique (for more details see e.g. Jia and Culver [2006]).

Each b-th bootstrapped resample is hereafter called bootsample to distinguish it from the randsample and particularly to highlight the existence of several resampled replicates in the same subset. The testing and validation

subsets are always randsampled, while the training subset is randsampled for SN systems and bootsampled for ensembles of neural networks in order to accomodate the stacked generalization framework [Wolpert, 1992, Zhang, 1999]. It combines with different techniques (average, regression, PCR, etc.) the outlet of all NN components where to each one a different bootsampled training replicate of the learning set is presented.

As a result of preliminary runs the size of each training bootstrap replicate equals the size of the training subset (i.e. the 70% of the initial learning set) since any improvement in accuracy arises when decreasing the 36.2% of the leaved out instances by enlarging the size of the bootstrap replicates [Breiman, 1996]. Conversely the validation subset is never bootsampled but randsampled for ensuring the validation of the neural network performance on the maximum number of available diverse space points.

The adaptation of the free parameters (weights and thresholds) to address the input/output mapping function might evolve in overtraining, a phenomenon in which a NN discovers features present in training data but that doesn't belong to the underlying function to be modeled (i.e. noise). To avoid the overfitting trouble training and testing signals are jointly submitted through the early stopping technique. Moreover, the validation dataset is submitted in simulation phase to evaluate to which extent the adaptation of the synaptic weights occurred in training phase produces the goodness of fit on the out of sample data (i.e. generalization ability on independent data drawn from the same population of training and testing sets).

The procedure of splitting the complete homogeneous dataset in the three mentioned subsets consists in randsampling the first random 15% to constitute the validation data, then it is the testing set turn with another random 15% quota in randsample strategy, and finally the residual 70% of data is bootsampled or randsampled in consideration of how many NNs are trained in the inference system at hand. The validation subset has the same set of out of sample gauges in every NN based inference system considered to make comparable the goodness of fit of the alternative models. In Fig 8.6 a broad perspective about the main steps involved is given.

8.3.5.1 NN configuration: sensitivity analysis for selecting the proper prototype

To properly configure a multilayer perceptron (MLP) a long trial and error procedure is required. Hence three random subsets (training, validation and testing) occasionally of lesser size are drawn from the Ta matrix (8.5) to speed up the analysis of how different settings of the parameters could affect the performance of an ANN. A feedforward backpropagation (FFBP) neural

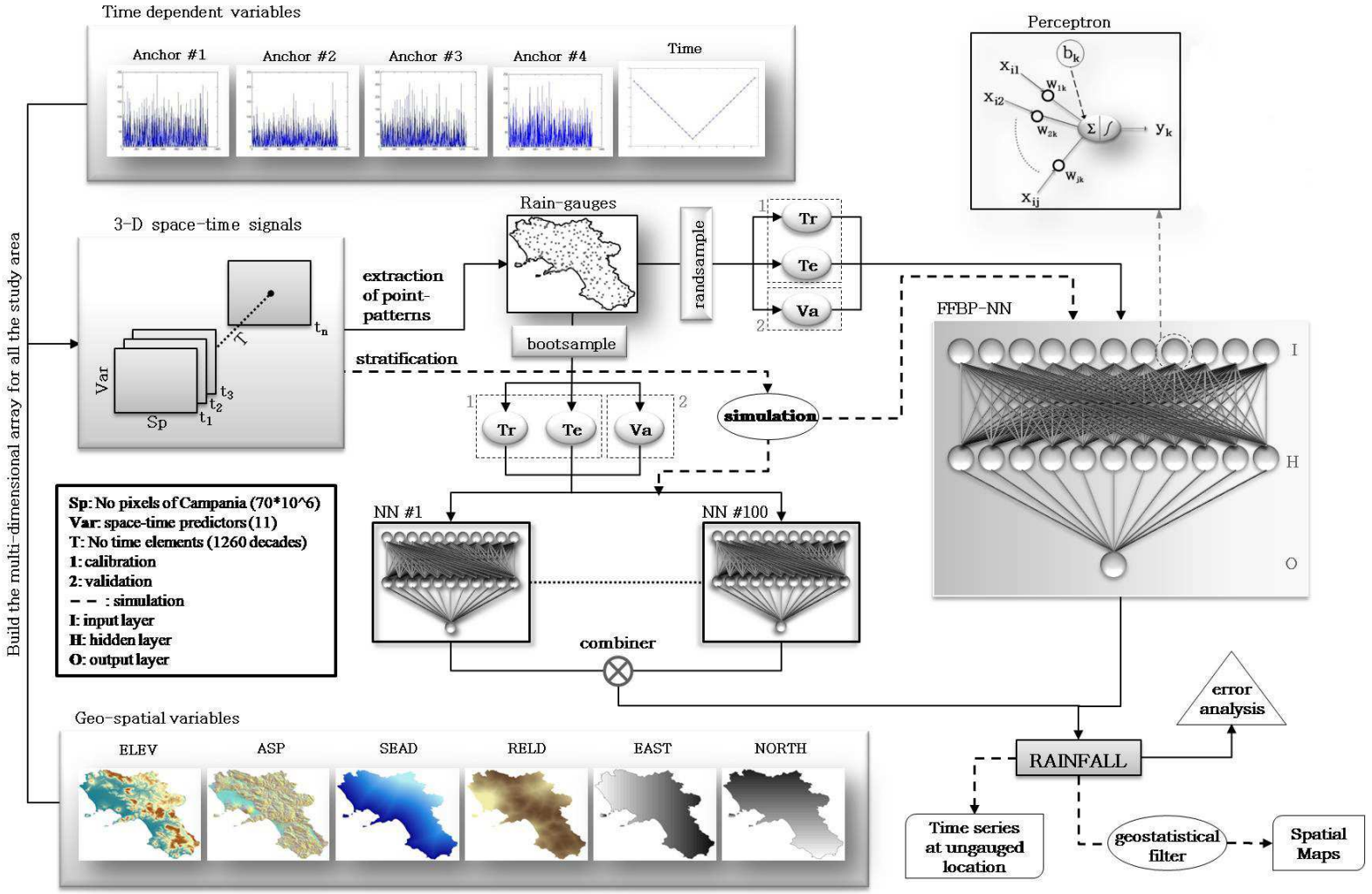


Figure 8.6: Flux of information from predictors to interpolated rainfall. A 3-D matrix with all signals is built for the study area and than from the space domain is extracted the geospatial data relative to gauged locations. Two resampling techniques, randsample and bootsample (see text for details), are pointed out to calibrate/validate the neural networks for the SN, the BAGNET or the sBN inference systems. Multi-temporal spatial maps of rainfall are made stratifying the 3-D array for the interested temporal extent and running the simulation with trained NNs for the domain-wide geographical space. The stratification from the 3-D matrix of one or more ungauged points for the whole time extent allows the simulation of precipitation time series at unobserved points.

network topology with three layers and architecture 11 : 11 : 1 with bias is selected as the initial configuration. A hierarchy between parameters is

Table 8.5: The input subset matrices used in neurocomputing have size GxDxPxB*. The B dimension does not exist in testing and validation subsets. Target arrays have the same size of input arrays, except for predictors dimension which is substituted by the unidimensional dependent variable

Phase	**Training**	**Testing** (r)	**Validation** (r)
Prototype Selection	30x360x11x10 (b)	7x360x11	7x360x11
SN–Campania	118x1260x11 (r)	20x1260x11	20x1260x11
BAGNET–Campania	118x1260x11x100 (b)	20x1260x11	20x1260x11
sBN–Campania	118x1260x11x?? (b)	20x1260x11	20x1260x11
SN–Cilento	28x1260x11 (r)	7x1260x11	7x1260x11
BAGNET–Cilento	28x1260x11x100 (b)	7x1260x11	7x1260x11
sBN–Cilento	28x1260x11x?? (b)	7x1260x11	7x1260x11

pointed out to progressively enable the recognition and the exploitation of those more performing settings. The imprint idea about the sequence and type of parameters that should be investigated is to some degree similar to the neurohydrologic rainfall/runoff modeling template proposed by Dawson and Wilby [2001].

The template developed and suggested in the present application is articulated in succeeding steps where the best parameter settings acknowledged in the previous stage is embodied in the stage in progress (except for the first one). The MatLab nomenclature and functions reported below are explained in more detail in Demuth et al. [2008].

1. Selection of the Training function among five inductive backpropagation learning algorithms:

   a. Gradient descent with momentum and adaptive learning rate (**traingdx**);

   b. Resilient (**trainrp**);

   c. BFGS (Broyden-Fletcher-Goldfarb-Shanno) quasi-Newton (**trainbfg**);

   d. One step secant (**trainoss**);

   e. Levenberg-Marquardt (**trainlm**).

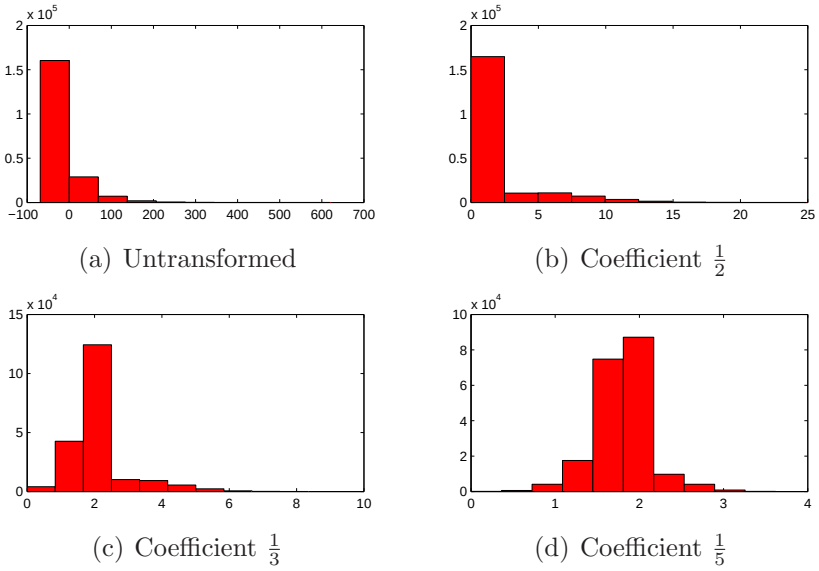
   The best performing learning functions are the trainrp and the trainlm. The latter algorithm [Hagan and Menhaj, 1994] is faster and more performant in regression problems.

2. Selection of the Transfer functions for the artificial neurons. Since the nodes are arranged in three layers (input, hidden and output) the identification of the best activation function involve each layer as a unique block. Two types of transfer functions are reviewed, the linear transfer function (p) and the hyperbolic tangent sigmoid transfer function (t). The eight possible combinations of 't' and 'p' units are considered:

   'p:p:p', 't:t:t', 'p:p:t', 'p:t:p', 't:p:p', 'p:t:t', 't:p:t', 't:t:p',

   with t=tansig and p=purelin. The best combination of transfer functions for training and validation subsets is constantly 't:t:t' for both 'trainrp' and 'trainlm' algorithms, while for testing it is erratic with a major redundancy of the 't' element in more than one casual positions. Hence the 't:t:t' combination is selected, with the 'trainlm' training algorithm which outperformes the 'trainrp' in whatever transfer function setting.

3. Data cleansing: transformation of precipitation data with a power function using a coefficient of 0.2. The frequency distribution of untransformed rainfall has a main vertical pattern (Fig. 8.7a). The monolith of data, that also accounts for the non rainy decadal units, assumes a wider distribution as coefficient of power function decrease.



(a) Untransformed (b) Coefficient $\frac{1}{2}$

(c) Coefficient $\frac{1}{3}$ (d) Coefficient $\frac{1}{5}$

Figure 8.7: Rain histograms for varying coefficients of power transformation. Untransformed has coefficient equal to 1.

4. To squash the magnitude effect of the input/target variables data standardization is adopted comparing three alternative ranges: [-1, 1], [-0.9, 0.9], [-0.8, 0.8]. The limits of these ranges are chosen to accommodate the hyperbolic tangent sigmoid activation function which is bounded between -1 and 1. The output signal of an artificial neuron never reach the theoretical minimum or maximum of its sigmoid function [Masters, 1993, Maier and Dandy, 2001]; indeed a neuron equipped with sigmoid transfer function should be considered fully activated at around 0.9 and turned off at -0.9. The performance of a NN is evaluated on six cases using the three linear scaling ranges both with row and power transformed rain data. The ± 0.8 range on row rain data and the ± 1.0 range on power transformed rain data highlight the higher accuracy. A decomposition of the rain intensity signal into three ranges (peak, middle and bottom) is carried out on the validation subset in order to explore the likelihood of the two best cited cases with respect to this signal magnitudes. The decomposition is based on the analysis of the box and whisker plot, from which two thresholds, the lower quartile and 1.5 the interquartile range from upper quartile, are selected to split the rain data in feature space. The peak signal with the ouliers, the middle signal groups the higher amount of values that are comprised between the two mentioned thresholds and finally the bottom signal with values ranging from the non rainy elements (zero values) to the first quartile. This procedure is useful because breaking up the performance enables a deeper insight on how a basic NN generalize on different signal magnitudes. Results are reported in Tab. 8.6. As confirmed by most indicators the ± 0.8 scaling range on row rain data outperforms the ± 1.0 range on power transformed rain data in peak and conversely the latter case gives better predictions in middle and bottom signals. Accordingly to the frequencies 5%, 70% and 25% of the precipitation values within the intensity ranges peak, middle and bottom, it is selected the ± 1.0 linear scaling range with 0.2 power transformed rain data as the best I/O data arrangement to train a neural network.

5. Selection of the proper dimensionality of the hidden layer. The opportunity of higher dimensionality of the hidden layer discloses a critical trade off between a global perspective of the trained NN and the undesired storing in synaptic weights of MLP of noise present in the input space. Four alternative sizes of the hidden layer with 11, 16, 22 and 33 artificial hidden neurons are investigated. The hidden layer topologies with 11 and 33 artificial neurons give quite the same goodness of fit on validation data, outperforming all other topologies; therefore ac-

Range	Intensity	**RMSE**	**MAE**	**r**	**SMAPE**	**AIC**
0.8	peak	63.63	51.04	0.4593	32.79	5880.4
	middle	19.14	13.43	0.7326	56.76	52750.9
	bottom	10.55	5.79	0.1421	804.41	15238.3
1.0	peak	67.61	55.29	0.4585	35.46	5958.3
	middle	19.07	13.48	0.7355	53.01	52678.9
	bottom	8.39	2.82	0.1149	302.31	13812.8

Table 8.6: Effect of normalization range and intensity decomposition on performance

cording to Ockham's razor the cheaper model with 11 hidden nodes is selected (from the Latin statement of the principle: *entities should not be multiplied beyond necessity*).

6. Selection of the range of weights and biases (the so called free parameters) for initialization: $\pm 10^{-x}$, for $x \in [0, 4]$.

The finest configuration is a multilayer perceptron backpropagation neural network with three layers and architecture 11:11:1, training algorithm Levenberg-Marquardt, activation function of type tansig for all nodes and custom initialization of thresholds and weights inside range ± 100. This will be the neural network prototype used throughout the paper.

8.3.5.2 Description of learning paradigms

The training paradigms are composed of the multilayer perceptron block properly tuned and of the input/output (I/O) pair of signals. In the previous paragraph the behavior of the explored learning parameters led to the identification of the most successful neural network arrangement able to map the rainfall pattern from the proposed set of predictors.

This well tuned NN block is embodied in three alternative inference systems whose architecture and functionality are unfolded afterwards. It might be of interest to compare the ability of ANNs in addressing the predictors/rainfall mapping using very small I/O pairs such as in the Cilento case study for a temporal extent from 1973 to 1987. Accordingly, the three inference systems are developed for both the Campania Region and for the Cilento subarea (Tab. 8.5).

SN The Single Network (**SN**) system is composed of only one 11:11:1 three layer MLP BP prototype component. It is trained and then validated with training, testing and validation subsets drawn in randsample strategy.

BAGNET The parallel connection of more prototype outlets for composing an ensemble prediction 8.6 should improve the generalization ability of a NN based inference system [Hansen and P., 1990].

The bootstrap aggregated neural network framework, called BAGNET in Zhang [1999], is a stacked ensemble with at most 100 NN prototypes. Each b-th network component has a unique starting condition defined by a particular configuration of the initial free parameters (weights and thresholds) and a peculiar pair of the I/O bootstrap training replicate.

After the training phase with early stopping technique each coached FFNN becomes an *expert* of at least a specific region of the potential input/output space to be mapped. The aggregated response is computed comparing the average (i.e. with all weights being equal) and the principal component regression (PCR) methods. The former treats all the neural network components with the same importance careless of the more or less performance of a single network. Conversely it introduces any hazard in generalization ability which occurs when ranking the NN performance based upon the training data.

This drawback partly accounts for the use of the PCR method proposed by Zhang [1999], where the PC (principal component) weights are computed on the training set and the selection of the proper number of weights used to build the BAGNET response is made on the testing set. The use of larger number of PC weights insinuate the overfitting on calibration data, this way the shortest number of PC weights with higher accuracy on testing set is selected.

One limitation is that testing set is considered as the only available source of *real world* patterns with which measure the generalization capability of the PC scores. The number of NN components influence the ability of the stacked inference system in producing an acceptable goodness of fit, thus a growing number (1, 5, 10, 25, 50, 100) of NN components is used to investigate as well the accuracy trend. Starting from an available set of 100 trained neural networks each BAGNET variant except the *BAGNET100*, is built 50 times with diverse and random combinations of NN components to investigate both accuracy and precision of the nonlinear fitting.

The single component inside BAGNET named *Best BAGNET 1* is characterized by the higher accuracy on training set among available NN components. This inference system is also fulfilled to ascertain to usefulness of the more cumbersome BAGNET variants in gaining good results.

The BAGNET system is calibrated with bootsampled training replicates and randsampled testing subset, whilst it is validated with randsampled validation subset.

sBN Accuracy of rainfall inference in space-time is highly influenced by how many NN components are assembled in calculating the BAGNET response, but also by which NN components are used. Zhou et al. [2002] states that *it may be better to ensemble many instead of all of the neural networks at hand*, and employs a genetic algorithm (GA) to detect the appropriate neural networks for composing an ensemble.

Here genetic computing is even run in order to select the proper neural network components for aggregating the *sBN-GA* variant. Moreover aside the GA technology a simpler selection method named *MSEPE* (percentiles of mean squared error) is proposed to generate an ensemble with smaller size than full BAGNET with 100 NN.

It is based on the performance of the training set, that is only those trained neural networks that behave well in terms of overall MSE on the calibration data are retained by keeping a threshold value. Instead of selecting a subjective threshold, five percentiles of MSE values (5, 10, 25, 50 and 75) are adopted in order to detect differences and recognize the best suitable percentile level for components selection. The NN experts selected with whatsoever method are connected in parallel to compose the sBN ensemble response. The training phase with early stopping criteria is fulfilled with the same resampling scheme adopted in BAGNET for drawing out the training, testing and validation subsets.

8.3.6 Boolean occurrence of rainy space-time elements with Geostatistics

The spatial intermittency of a rainfall field at level of a single decade segments the geographical space into two possible states, the rainy and non rainy state. Indicator kriging is the discrete binary random function [Goovaerts, 1997] used here to describe such a Boolean condition.

A mobile threshold depth is applied to the precipitation at sampling locations for each time element and is assigned a value of 0 to locations where rain catchment is less than or equal to the threshold depth and a value of 1 to locations where rainfall is greater than the fixed threshold. Thus, indicator values are calculated for each sampled location $\mathbf{u}$ as

$$I(\mathbf{u}|\mathbf{z}) = \begin{cases} 0, & \text{if } Z(\mathbf{u}) \leq z \\ 1, & \text{otherwise} \end{cases} \tag{8.3}$$

where $I(\mathbf{u}|\mathbf{z})$ is the indicator value at location u and for threshold depth $\mathbf{z}$, Z($\mathbf{u}$) is the measured rainfall at location u, and $\mathbf{z}$ is the threshold depth.

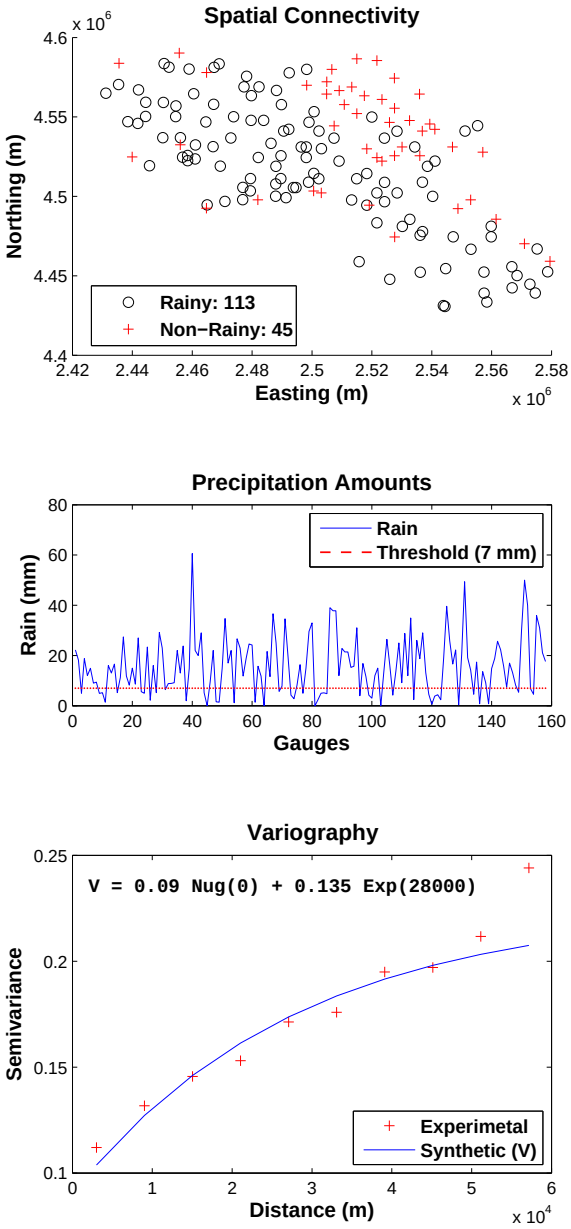


Figure 8.8: The selection of threshold depth (7mm, middle) influence the spatial connectivity of 158 gauges (top), the amount of rainy and non rainy gauges and the variographic analysis (bottom). Equation of synthetic semivariogram is printed for the first decade of May 1987

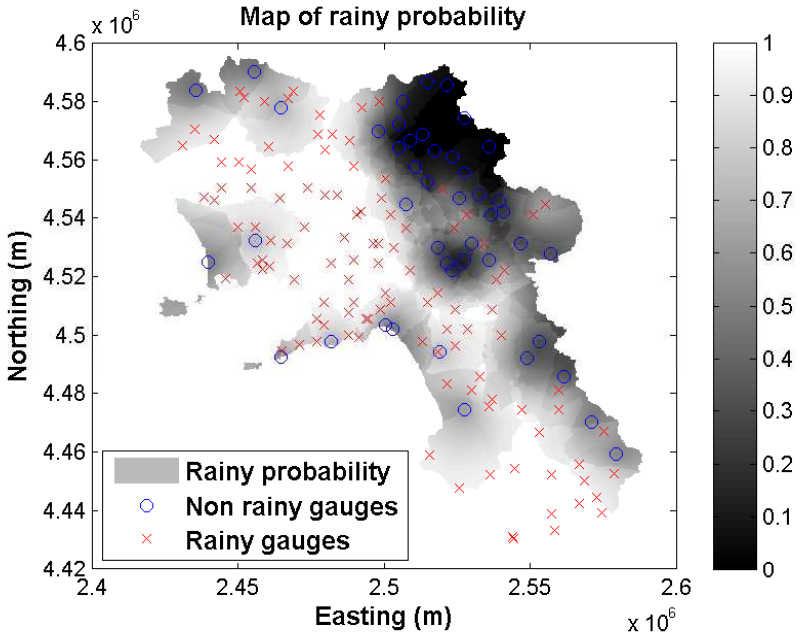


Figure 8.9: Indicator kriging of discrete binary precipitation. (Year: 1987; Decade: 13; Threshold: 7mm)

The threshold is mobile in the sense that it is not fixed for all time units but each decade is characterized by its own z value in order to put in evidence the spatial structure.

To be more rigorous the two possible states should not be called rainy and non rainy. In fact a year is composed of 36 decades and only few of them are non rainy at location u or conversely only few gauges are totally empty per time unit. Hence instead of considering the rainy quality one should consider the spatial connectivity of a sort of twofold (low and high) density cloud probability for that temporal element, interpreted as the spatial arrangement and time frequency of wet and dry periods within the considered decade. Variography (Fig. 8.8) and kriging (Fig. 8.9) are carried out in MatLab with mGstat.

8.4 Results

8.4.1 The regression imputation

The Figure 8.10 gives an idea of the number and performance of the multilinear regressions involved for the prediction of a single missing data; in

particular is evidenced the selection of the model believed to be the best one.

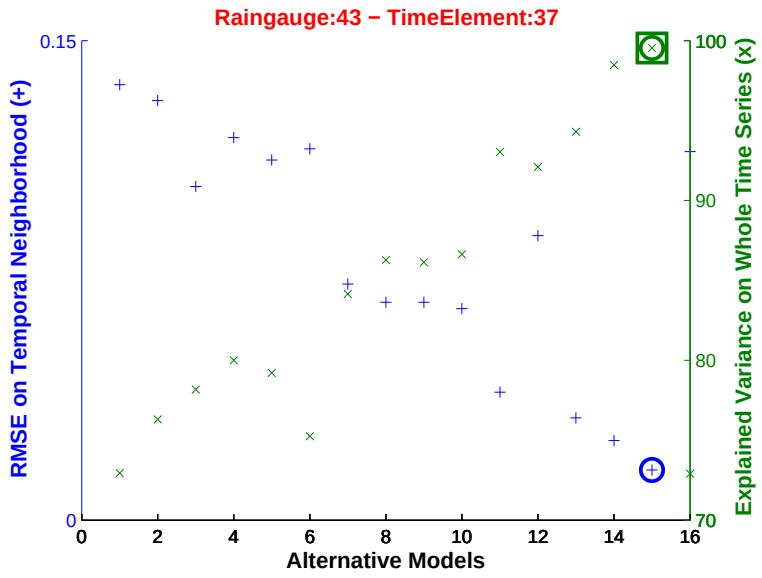


Figure 8.10: Performance on temporal neighbourhood (left ordinate) and on whole available time series (right ordinate) of the sixteen alternative models built to fill gap of 43rd incomplete station at 37th missing time unit. Circles select the best performance on both time lengths and square highlights the model selected to predict current missing value

Each model is characterized by a particular configuration of predictors both in geographical domain and from statistical viewpoint, in terms of correlation coefficients and p-values (Fig. 8.11). The algorithm is able to predict the 94% of the total missing values occurring in the Tb target matrix after running nonstop for about 5 days (Tab. 8.3).

The algorithm exhibits (Fig. 8.12) an overall good performance as more than 70% of total missing data is predicted by means of multilinear regression models with at least 90% of variance explained. Bad predictions (less than 70% of variance explained) are very exiguous (about 1%) and consequently are considered as noise when training a NN.

An important portion of models with less than 90% of explained variance are the result of how the regression imputation algorithm works when selecting the model with the best performance on neighborhood from the set of possible models.

Model selection based on overall/neighborhood performance is a strategic trade off: the algorithm put priority on the neighborhood performance as the indicator of the best possible prediction of a given decadal rainfall value, even if the overall performance of the model is penalized. This choice is supported

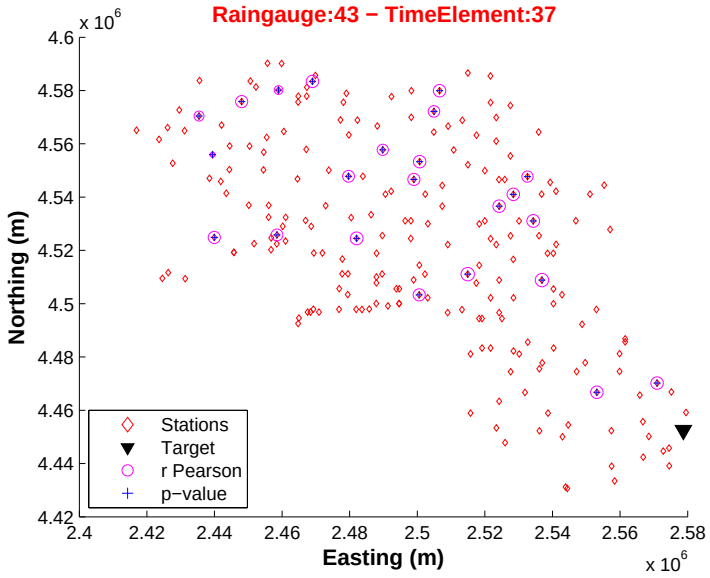


Figure 8.11: Geographical and statistical configuration of gauges from network population used to fulfill the 15th multilinear regression model showed in (A). Variability in Pearsons r and p-value domains is evidenced by symbol size which gives an idea of potential contribution in terms of correlation and significance. In current case correlation coefficient range from 0.102 to 0.755 and p-value range from 1.678e-232 to 2.867e-004. Note that similar size of circles doesnt account necessary for collinearity degree among predictors

by trials conducted on non missing data and obviously only account for a part of not so good models depicted in Fig. 8.12.

8.4.2 The ANN inference systems

The performance of the three inference systems is evaluated on the same validation data and is summarized in Tab. 8.7. There are the single network (SN), the BAGNET ensemble in seven variants and the sBN framework with two selection methods (GA and MSEPE) in six variants.

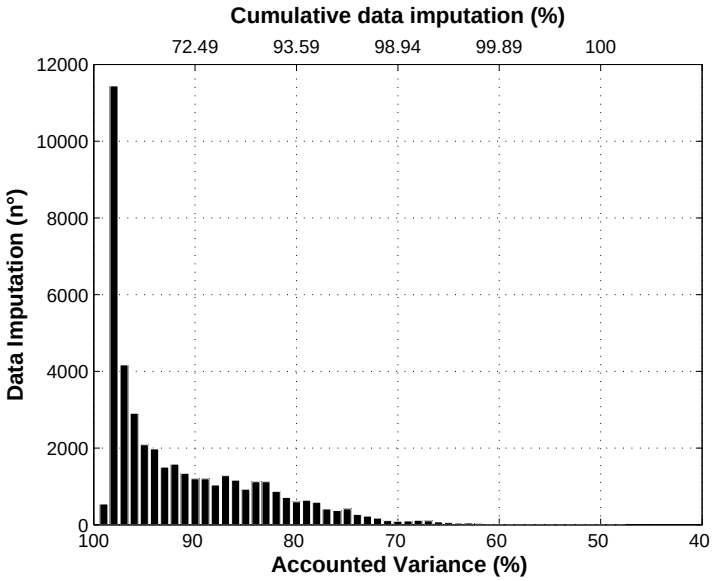


Figure 8.12: Gaps filled by regression imputation. Ordinate indicates the absolute number of predictions made in correspondence of percentage variance accounted for. Upper abscissa shows the percentage of data predicted at corresponding levels of goodness of fit

Table 8.7: Indicators are computed on the out-of-sample data (validation subset) with 25200 cases (20 gauges * 1260 decades) for Campania area and 3870 cases (7 gauges * 540 decades) for Cilento subarea. The 'Best BAGNET 1' is the single NN within BAGNET that performs at best on training subset. The 'sBN-Tr' series is based on the MSEPE method on training data (see text for details).

		RMSE		**MAE**		**MBE**		**r**	
		Average	PCR	Average	PCR	Average	PCR	Average	PCR
Cilento	SN	23.56		13.70		-3.03		0.844	
	BAGNET-1[1]	54.03±94.6		36.11±78.5		17.99±80.2		0.781±0.15	
	BAGNET-5[1]	46.86±32	22.47±0.8	33.28±29.8	13.49±0.5	20.1±34.9	0.72±1.1	0.74±0.13	0.859±0.01
	BAGNET-10[1]	38.49±18.3	22.44±0.8	27.79±17.9	13.51±0.5	16.8±23	1.08±0.8	0.79±0.06	0.86±0.01
	BAGNET-25[1]	30.57±7.4	22.14±0.4	21.8±7.7	13.43±0.3	12.62±12.1	1.34±0.5	0.83±0.02	0.863±0
	BAGNET-50[1]	29.39±4.7	22.21±0.3	21.64±5	13.63±0.2	14.93±7.4	1.87±0.4	0.84±0.01	0.862±0
	BAGNET-100	28.14	22.29	20.82	13.75	15.07	2.24	0.849	0.862
	Best-BAGNET-1	23.03		13.74		0.19		0.852	
	sBN-GA	22.30	22.47	12.97	13.90	-3.07	2.41	0.863	0.860
	sBN-Tr-5prct	22.98	23.75	13.56	14.15	-0.91	1.79	0.851	0.849
	sBN-Tr-10prct	25.53	22.83	15.09	13.68	-0.38	1.54	0.812	0.858
	sBN-Tr-25prct	23.23	22.77	13.94	14.02	-1.10	1.89	0.846	0.856
	sBN-Tr-50prct	26.86	22.30	19.47	13.73	12.50	2.17	0.846	0.862
	sBN-Tr-75prct	24.86	22.25	17.30	13.68	9.24	1.94	0.852	0.862
Campania	SN	22.48		12.97		-4.56		0.826	
	BAGNET-1[1]	23.85±1.3		13.69±0.5		-5.01±1.3		0.804±0.02	
	BAGNET-5[1]	24.08±4.8	22.46±0.2	13.79±2.7	13.26±0.2	-4.57±2.1	-1.1±0.5	0.806±0.06	0.822±0
	BAGNET-10[1]	23.28±1.8	22.34±0.1	13.41±1.3	13.19±0.1	-4.73±1.2	-1.08±0.3	0.815±0.03	0.824±0
	BAGNET-25[1]	23.08±0.6	22.22±0.1	13.39±0.6	13.13±0	-4.49±0.8	-1.02±0.2	0.816±0.01	0.826±0
	BAGNET-50[1]	22.82±0.2	22.17±0.1	13.19±0.2	13.11±0	-4.61±0.4	-0.95±0.1	0.821±0.01	0.827±0
	BAGNET-100	22.73	22.14	13.12	13.09	-4.6	-0.97	0.823	0.827
	Best-BAGNET-1	22.79		13.16		-4.55		0.821	
	sBN-GA	22.67	22.25	12.98	13.14	-4.99	-0.96	0.826	0.826
	sBN-Tr-5prct	23.85	23.17	13.73	13.97	-5.33	-0.40	0.807	0.808
	sBN-Tr-10prct	23.06	22.43	13.23	13.40	-5.31	-0.65	0.821	0.822
	sBN-Tr-25prct	22.73	22.20	13.03	13.20	-5.08	-0.83	0.826	0.826
	sBN-Tr-50prct	22.99	22.12	13.43	13.10	-4.17	-0.96	0.816	0.827
	sBN-Tr-75prct	22.83	22.14	13.22	13.10	-4.50	-0.99	0.820	0.827

(continued on next page)

(continued)

		SMAPE		AIC		Willmott's D	
		Average	PCR	Average	PCR	Average	PCR
Cilento	SN	82.07		0.144		0.907	
	BAGNET-1[1]	93.85±24.4		0.144±0		0.828±0.2	
	BAGNET-5[1]	94.93±19.9	79.17±1.6	0.728±0	0.729±0	0.784±0.16	0.923±0.01
	BAGNET-10[1]	94.38±15.1	77.62±2.5	1.458±0	1.459±0	0.83±0.1	0.924±0.01
	BAGNET-25[1]	90.5±9	76.48±2.9	3.649±0	3.649±0	0.875±0.04	0.925±0
	BAGNET-50[1]	89.97±6	77.11±11.5	7.3±0	7.3±0	0.883±0.03	0.924±0
	BAGNET-100	88.22	74.84	14.601	14.602	0.891	0.924
	Best-BAGNET-1	80.14		0.144		0.919	
	sBN-GA	78.63	75.28	9.490	9.783	0.916	0.923
	sBN-Tr-5prct	80.15	80.27	0.729	0.729	0.918	0.919
	sBN-Tr-10prct	80.90	78.84	1.459	1.459	0.894	0.924
	sBN-Tr-25prct	80.38	77.73	3.649	3.649	0.911	0.922
	sBN-Tr-50prct	86.79	72.74	7.300	7.300	0.898	0.924
	sBN-Tr-75prct	83.42	73.07	10.951	10.951	0.909	0.924
Campania	SN	81.41		0.022		0.892	
	BAGNET-1[1]	83.92±1.6		0.022±0		0.876±0.02	
	BAGNET-5[1]	82±3.2	79.91±0.9	0.109±0	0.109±0	0.874±0.04	0.901±0
	BAGNET-10[1]	81.53±2.1	79.7±0.6	0.219±0	0.219±0	0.882±0.01	0.902±0
	BAGNET-25[1]	81.37±1.2	78.86±1	0.547±0	0.547±0	0.883±0	0.903±0
	BAGNET-50[1]	80.84±0.7	75.99±9.7	1.095±0	1.095±0	0.886±0	0.904±0
	BAGNET-100	80.56	74.98	2.19	2.19	0.886	0.904
	Best-BAGNET-1	82.23		0.022		0.893	
	sBN-GA	80.44	79.30	1.467	1.467	0.887	0.903
	sBN-Tr-5prct	83.33	82.01	0.109	0.109	0.867	0.892
	sBN-Tr-10prct	81.32	79.98	0.219	0.219	0.879	0.901
	sBN-Tr-25prct	80.73	79.58	0.547	0.547	0.884	0.903
	sBN-Tr-50prct	81.17	77.45	1.095	1.095	0.884	0.904
	sBN-Tr-75prct	80.87	76.94	1.643	1.643	0.885	0.904

[1] The value of each indicator is computed on 50 repetitions with diverse and random combinations of the bootsampled training subset replicates; the aggregation methods 'Average' and 'PCR' are evaluated on the same 50 random compositions of replicates. Standard deviation is reported too.

Several considerations might be elaborated, therefore I will focus only on those aspects relevant for the pursuits of the paper.

The use of a single network inference system (SN) is not the best choice considering the overall trend of indicators. Enlarging the validation subset size (from 7 to 20 gauges respectively for Cilento and Campania) decreases the performance of the SN, notwithstanding Campania case study has a larger and more variegate learning set in terms of representativeness of both realty and feature space. This suggests that a more complex inference system represented by ensembles of NN should be selected in order to get better results in making spatial maps of precipitation, where the number of simulated ungauged pixels could be very high.

The way of BAGNET is worthwhile. The stacked generalization shows less precision in case of smaller number of neural network components, and the truthfulness of this statement is higher for the Cilento subarea and for the average method (see standard deviation).

The numerousness of space-time elements used in Campania case study for calibration ensure high accuracy even at lower BAGNET size. The PCR method quite always outperforms the average aggregation type which is also outperformed by the 'Best BAGNET 1' inference system; more probably the prediction of the time series at one or very few ungauged locations could be addressed with good likelihood by using the BAGNET best single NN component.

Comparing the PCR aggregation method in BAGNET learning paradigm for Cilento and Campania it is remarkable how the performance in the former case is higher in the 25 components BAGNET model while in Campania case study the full components model gives the best result. The validation subset of smaller size in Cilento is maybe more flexible in converging towards target signals.

Contrariwise the AIC statistic, SMAPE decreases as the learning paradigm at hand became more cumbersome, highlighting a higher accuracy and a smaller precision such as in the BAGNET50 of Campania case study where a standard deviation of about 9.7 units is revealed (the larger value for Campania BAGNET variants aggregated by PCR). The Pearson coefficient r and the Willmott's D don't represent a good discriminating within the same learning paradigms as in order of magnitude the AIC, MBE and SMAPE do.

8.4.2.1 Building a time-series at ungauged location

It is used the *Best BAGNET 1* variant on validation subset to evaluate the ability of the best single trained prototype inference system in giving good results for predicting precipitation values at few ungauged locations.

Six gauges of one year length are drawn out from the validation set at random in order to graphically evaluate performance during the 36 decades (Fig. 8.13). Dashed line put in evidence the underestimation of relative peaks as pointed out by the Yearly Difference in Percentage (YDP) index too. It is computed as follows:

$$YDP = \frac{\sum_{i=1}^{N} P_i - \sum_{i=1}^{N} O_i}{\sum_{i=1}^{N} O_i} \cdot 100 \tag{8.4}$$

where P_i and O_i are respectively the predicted and the observed values on decade i, and N=36. It forthwith provides a yearly percentage distance of predictions from observations.

8.4.2.2 Multitemporal maps of rainfall fields

It is pointed out that a stacked model of inference with more neural network components is required to draw up a spatial map of precipitation for the Campania domain wide area at level of a single temporal unit.

In Tab. 8.7 statistical measures are computed for the whole time series, but here the task consists in building the decadal precipitation field of a year as an example of possible application. Statistical indicators derived on validation set and within this restricted temporal window are different with respect to the values obtained for the entire time series.

An additional statistic for a year selected at random, say the 1979, is run to identify the more performing model variants. Two comparative models of inference, the 5^{th} and the 50^{th} percentiles variants of sBN-Tr, are employed to make the maps reported in figures 8.4.2.2 and 8.4.2.2 for visualizing the spatial pattern of synthetic rainfall fields.

8.4.3 Spatial intermittency of rainy occurrence

Calibration of inference systems is carried out using large multidimensional matrices. To make a spatial map of precipitation the sBN-Tr framework is trained with matrices 118x1260x11x5 or 118x1260x11x50 for respectively the 5^{th} and the 50^{th} percentiles variants.

There exist a great unbalance between the spatial (118) and temporal (1260) singletons, which is responsible for a better prediction in time (Fig. 8.13) than in space domain (red dotted line, Fig. 8.16). This should justify the use of further computation to filter the predicted precipitation map.

The indicator kriged map (Fig. 8.9) is multiplied by the map of predictions and then values are extracted from gauged locations to compare

patterns. To highlight fitting performance of filtered and unfiltered predictions the network rainfall signature (solid line, Fig. 8.16) is depicted as sorted in ascending order by rain catchments. The filtered overcome the unfiltered predictions as correlation coefficient and YDP put in evidence.

8.5 Conclusions

In this chapter the space-time analysis of gauged precipitation is addressed with artificial neural networks. It was demonstrated that:

Low cost geospatial and temporal covariates

High space-time resolution

Good accuracy (compared with other models in literature at lower resolution e.g. Drogue et al. [2002])

Inference with ensemble of NN in case of few data

Several indicators of performance (to be sure of goodness of fit)

Difference for year 1979 for map and time series (show that inference is better in time than in space because of the major number of cases in that dimension)

Machine time consuming for the different elaborations (imputation, calibration, simulation,)

Speak about the possibility to apply a PLS algorithm which takes into account both the explanatory variables and the response variable (rainfall) see Sicard and Sabatier [2006] page 1396.

Speak about the possibility to applay a geostatistical filter also to the top part of network rainfall signature (and not only to the bottom).

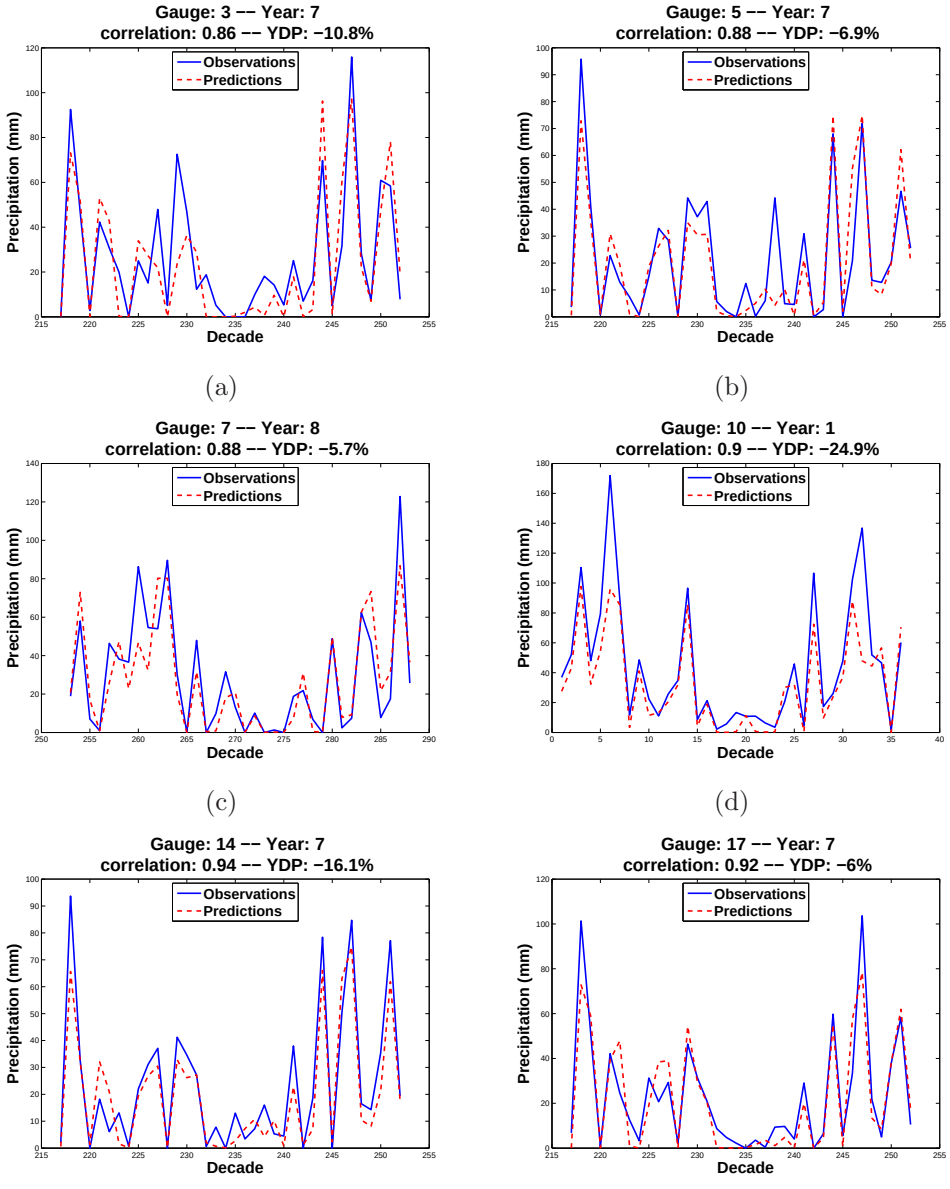


Figure 8.13: Rainfall prediction at six different locations (a-f) belonging to the validation set and during different years. Note the constant underestimation of higher peaks. YDP: Yearly Difference in Percentage

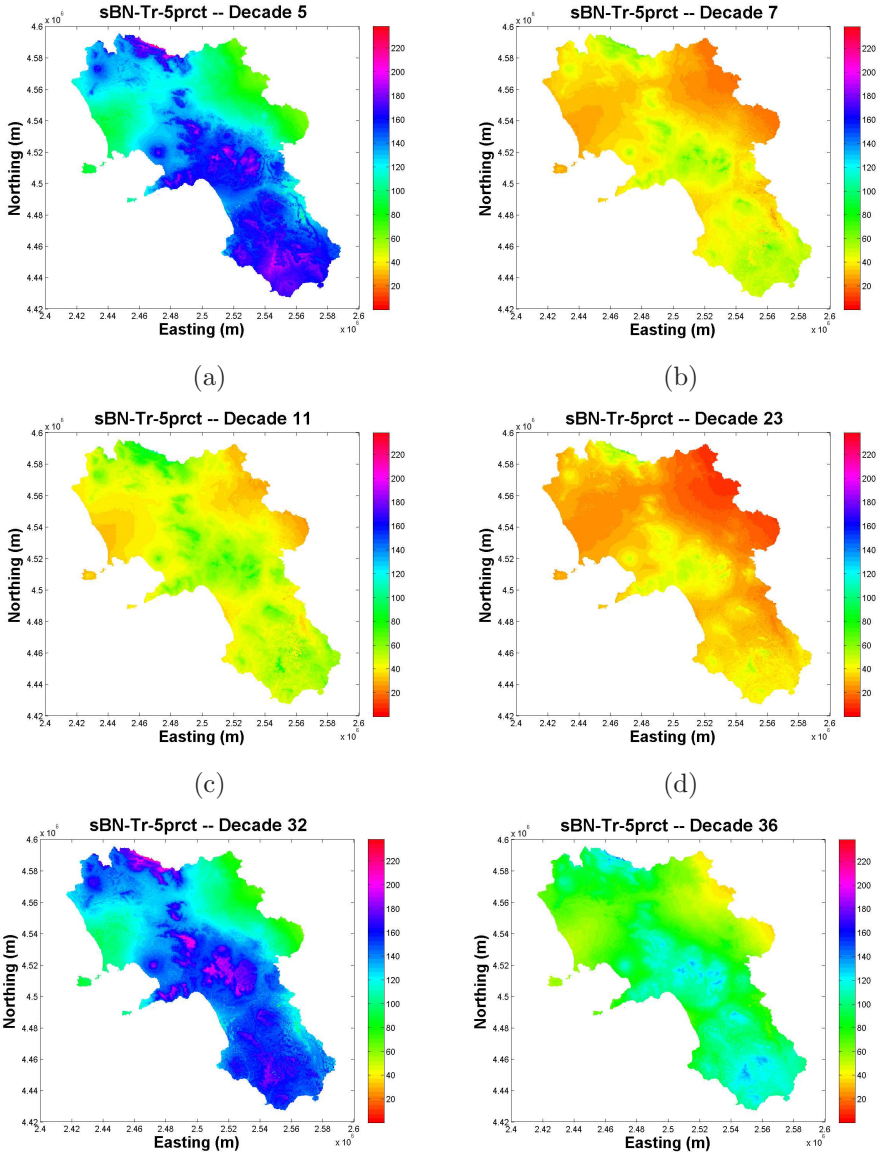


Figure 8.14: Precipitation maps for Campania region study area. Predictions are made using the *sBN-Tr-5prct* variant.

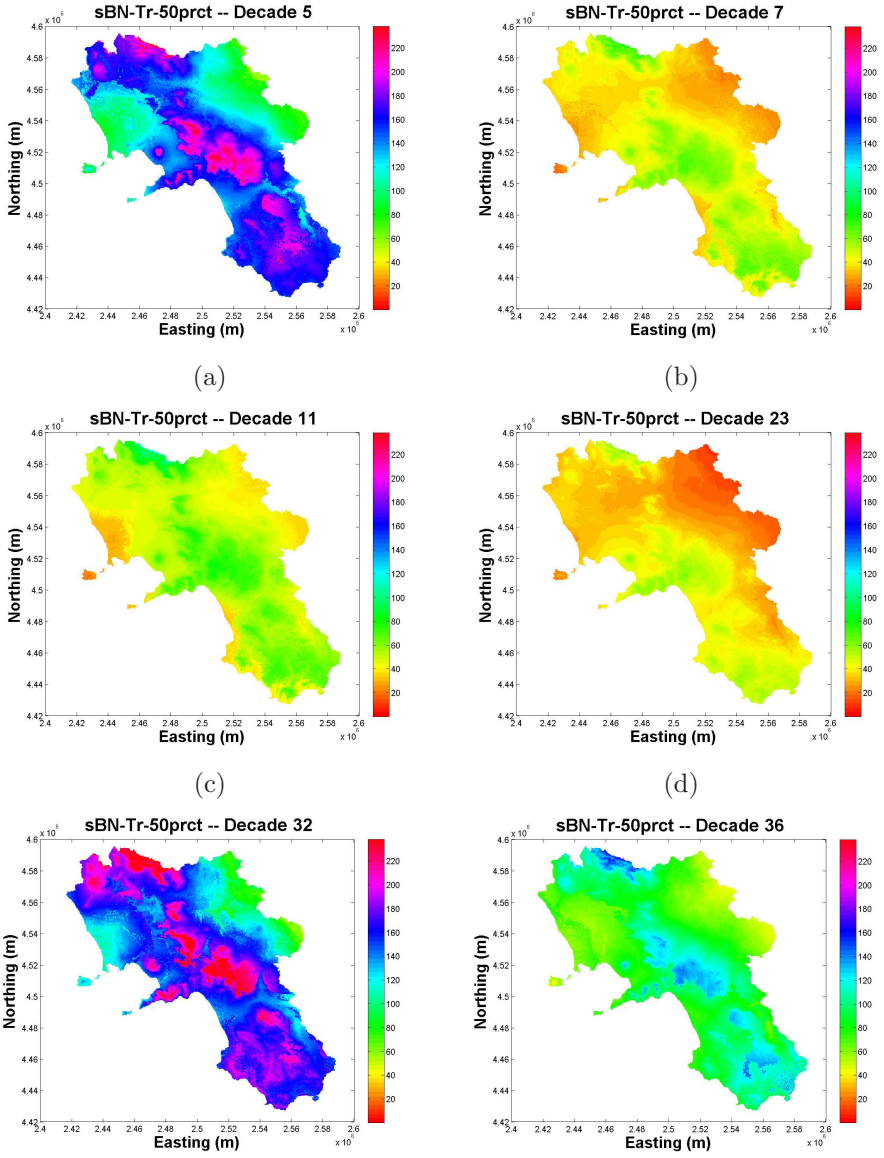


Figure 8.15: Precipitation maps for Campania region study area. Predictions are made using the *sBN-Tr-50prct* variant.

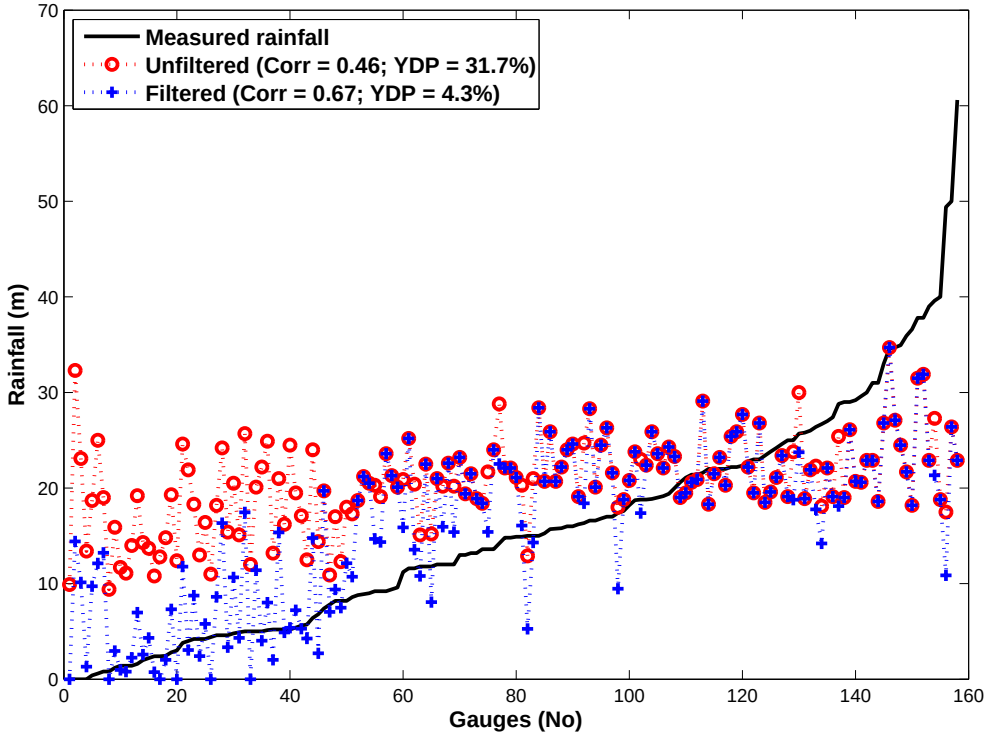


Figure 8.16: Raingauges are sorted in ascending order based on measured rainfall. Note the incompetence of neural networks to satisfy target signals at low intensities (red dotted line, sBN-Tr-5prct variant). The filtered predictions provide higher correlation coefficient and yearly difference in percentage (YDP)

Part III

Conclusions and Addendum

Chapter 9

Conclusions

The major agricultural and environmental problems indeed require a detailed knowledge of agricultural and forest ecosystems. This knowledge demands a detailed information about the spatial distribution of soils, climate and plants.

The major technological innovations produced in the last decades by the use of satellites enable to obtain detailed land use information at high spatial and temporal resolution. This unfortunately is not the case for the pedoclimatic data which by their nature are much more complex to be determined. Moreover, in the case of climate, while temperature (eg. daily mean temperature) is generally well correlated to some physical land parameters such as altitude and aspect, this is not true for the rainfall parameter because of its implicit complexity (non linearity). These difficult problems have been the focus of this thesis which aimed to approach the two major issues of spatial inference concerning soils and rainfall comparing different methodologies.

In the case of soils, and only for selected investigated variables such as the clay content and the soil colour, it was shown that data obtained from the standard soil map did not well perform in differentiating landscape classes according to their clay content. More specifically, while some landscape units resulted well differentiated (mountain relieves against alluvial terraces), other landscape units (hilly environments) did not show such significant differences in the comparison with other landscape units.

The fuzzy analysis applied to the digital terrain model for obtaining discrete units of landforms made it possible to clearly differentiate some morphological landscapes with very significant differences in their clay content but, again, this approach has highlighted that some other morphological landscape did not significantly differ from others in their mean clay content.

Geostatistical techniques applied by means of Universal Kriging have not always provided interesting data. In fact many of the obtained variograms

were not structured and did not enable to obtain a proper spatialisation of the information.

The multiple regression techniques and the application of neural networks were much more promising. In particular, the ANN has certainly produced the best results despite that I did not have an extremely high number of basic information.

In general, these results clearly show how the most complex techniques of spatial inference are more promising and more efficient but they are also, unfortunately, those that require more data and thus are more costly.

Among the investigated variables it is important to quote the example of the soil colour. We know that this morphological parameter described using the Munsel soil charts is generally associated with the content of organic carbon, the presence of Fe and Mn oxides, it depends by the degree of weathering, by the content of calcium carbonate, etc. The quantitative analysis of this attribute made after colour transformation has produced very poor results highlighting the absence of structured variograms and of good independent variables for a suitable regression analysis. On the other side, an alternative approach based on a generic interpretation of the soil colour across the study area made it possible to produce an empirical index of pedogenesis (soil redding and darkening). The PDI showed to have a good variogram structure and enabled the spatial inference of this powerful soil information.

The space-time analysis of gauged precipitation, which is addressed with artificial neural networks, demonstrates that low cost geospatial and temporal covariates can account with relative very high performance for the high space-time resolution of precipitation data. It is showed how neurocomputing yield good accuracy compared with other models in literature at coarser resolution (e.g. Drogue et al. [2002]).

Inference with ensemble of neural networks in case of few and sparse data (Cilento subcase study) indicates that bootstrapping neural networks for building a stack inference system can help in set up satisfactory models for studying non linear processes such as precipitation distribution.

Also the use of several indicators for evaluating the performance of models can support a straightforward decision about the best approach.

Predictions for the year 1979, exhibits a less performance in the case of map productions compared with the building of a time series at a single ungauged location, due to the use of more temporal singletons then spatial ones. The application of a geostatistical filter to a rainfall map produced by neurocompting displays good results, as evaluated on the out-of-sample data. Maybe an indicator kriging should also be applied to the higher magnitude area of rainfall spatial signature (Fig. 8.16), since predictions are inaccurate

both at the bottom (where nonrainy gauged elements verify too) and at the top of rain spatial signature.

There exist the possibility to apply a PLS algorithm Sicard and Sabatier [2006] in combining neural networks components. Indeed it takes into account both the explanatory variables and the response variable.

Appendix A

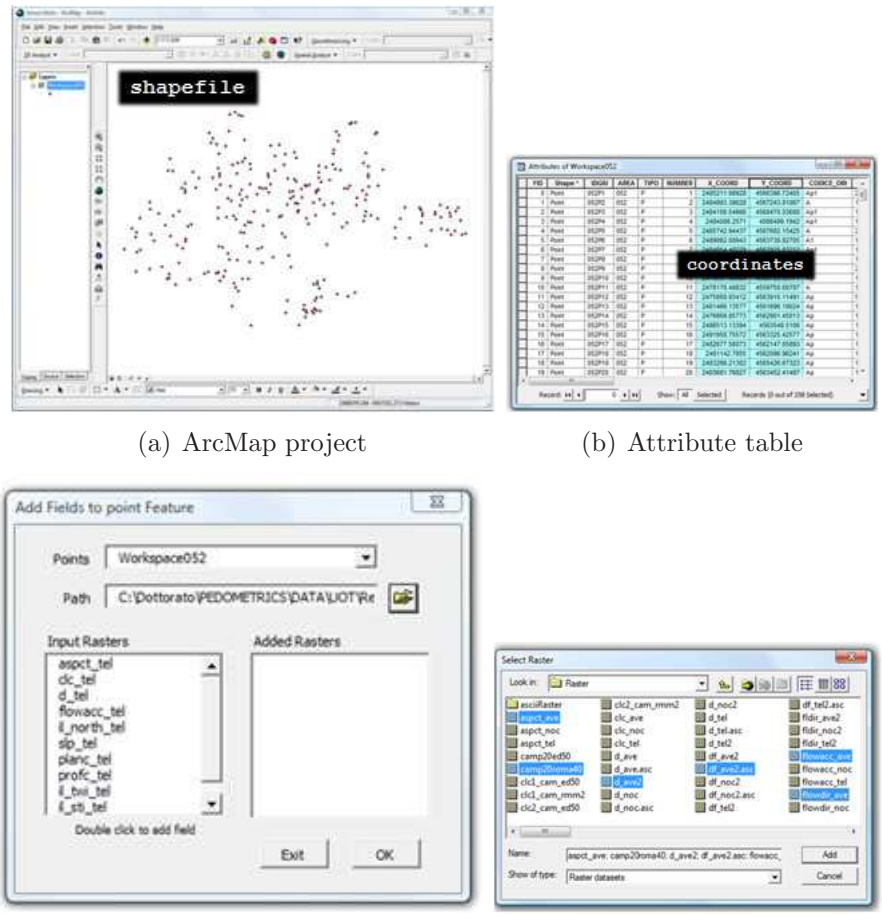
The MultiFieldAdder tool

The MultiFieldAdder is a very useful ArcGIS tool for building a matching table to be used in the spatial analysis of landscape attributes.

It adds to a point feature (shapefile) loaded in the current ArcMap project a new field with label equal to the string name of the selected raster layer. Then the new empty field of the attribute table is compiled with values picked from the lattice grid where the point feature has sample locations (points with coordinates).

The MultiFieldAdder is able to load several raster layers at the same time (multi-selection ability), in order to add many new fields in few seconds. This tool is very useful if considering the number of available auxiliary maps that otherwise should be elaborated in a manual fashion so that a complete matching table is built.

From ESRI website at URL `http://arcscripts.esri.com/details.asp?dbid=14826` you can download the zipped file with tool and detailed instructions on how to install it. In Fig. A it is showed how the user interface works.



(a) ArcMap project (b) Attribute table

(c) MutiFieldAdder tool (d) Multi-selector

Figure A.1: How the MultiFieldAdder tool works

Appendix B

The EDASS tool

The most important stuff is explaining how to get started with the EDASS tool. Firstly one should have Microsoft Access installed on workstation, then should own the database file with EDASS implemented in. Afterwards user have to:

(i.) install R;

(ii.) install R/Scilab (D)COM Server V2.50;

(iii.) open Access soil database within which the EDASS tool is implemented and run the EDASS mask (it is developed on a VB form). During working session a hidden connection is established with R software across the (D)COM Server platform, and R instructions are conveyed to run designed operations.

EDASS cannot be released as independent tool, since it is designed in VBA (Visual Basic for Applications) to specifically work under the database at hand. This means that it can works within whatsoever Access database, but is not able to perform tasks independently. Therefore it is my intention to translate the VBA application into a Visual Basic 6 format in order to allow anybody enjoying it outside Microsoft Access environment.

The EDASS tool is a user friendly interface that links in the background the power of SQL statements possible in Access with graphical and statistical capabilities of R statistical software.

Appendix C

ANNvsREGR MatLab script

Here is printed the script used to quickly implement a comparison between multi linear regression and artificial neural networks in explaining the variance in a target attribute using environmental covariates.

Steps are enumerated from zero to six, and settings to be manually performed are highlighted as **bold** text:

-0. The matching **table** is loaded in workspace.

-1. Select the **column number** of target variable, and the **threshold** for Pearson correlation. One can graphically recognize the good predictors at different levels of thresholds. Then for a selected threshold, the set of predictors are fixed and used later on; otherwise one manually select/deselect predictors in `pred_sel` and `head_pred_sel` MatLab variables.

-2. Split the total number of cases in the calibration and validation subsets, according to a manual selected **threshold**. Here the input/output couples are also normalised in range $[-1, +1]$.

-3. The neural network is initialised through a personal function (one can personalize parameter settings), trained with calibration data, and simulated with validation subset. In this step following ANN settings are fulfilled: **range of input**, **topology of network**, **activation functions**, **range** within which **initialization** is made for weights and biases.

-4. Stepwise regression analysis is fulfilled with the same calibration subset used for ANN, and then model performance is evaluated on validation set.

-5. Save the workspace in a .mat file for future visualization through step 6.

-6. Load a saved analysis made with the ANNvsREGR function through steps 0 to 5, and show graphics and correlations for performance comparison of the two statistical methods on the sub case study at hand.

```
%% ANNvsREGR
%% -0.\ LOAD TABLE
clear
cd('C:\Dottorato\PEDOMETRICS\DATA\UOT\MatLab')
load pick_terrain.mat merge

%% -1.\ SELECT COVARIATES

%--------------------
col_target = 15;    % clay:15;
Threshold = 0.2;    %of correlation coefficient
%--------------------

%-TARGET
TARGET = cell2mat(merge(2:end,col_target));
head_target = merge(1,col_target)

%-PREDICTORS
pred = cell2mat(merge(2:end,[39:74 91:end]));
head_pred = merge(1,[39:74 91:end]);

%-CORRELATION TASK:
c = corr([TARGET pred], 'rows','pairwise');
c_pred = c(2:end,1);
pred_good_c = find(c_pred > +Threshold | c_pred < -Threshold);
%-select good predictors
pred_sel = pred(:,pred_good_c);
head_pred_sel = head_pred(:,pred_good_c);

%--plot
plot(c_pred,':k');
hold on; scatter(pred_good_c, c_pred(pred_good_c), 'r', 'LineWidth',2); hold off;
title('Recognize Good Predictors', 'FontWeight','b', 'FontSize',14)
xlabel('# Covariates', 'FontWeight','b', 'FontSize',12);
ylabel('Pearson Correlation', 'FontWeight','b', 'FontSize',12);
for i = 1:size(pred_sel,2)
    text(pred_good_c(i),c_pred(pred_good_c(i)),['\',head_pred_sel(i)])
end
%-clear
clear head_pred pred c pred_good_c ans i Threshold c_pred col_target

%% -2.\ PREPARE IN/OUT ARRAYS
%---------------------------------
%split amount between datasets
subsets = [0.70 0.30];
%---------------------------------

% find nans in predictors/target:
[r1 c1] = find(not(isnan(pred_sel)));
[r2 c2] = find(not(isnan(TARGET)));
```

```
% take all not NaNs values
r = intersect(r1,r2);
t = TARGET(r)';
p = pred_sel(r,:)';
xy = cell2mat(merge(1+r,[25 26]))';
clear c* r* ans

% Normalize
mm = minmax(t);
tn = normalize(t, [mm(:,2) mm(:,1)],-1,+1);
mm = [min(p,[],2) max(p,[],2)];
pn = normalize(p, [mm(:,2) mm(:,1)],-1,+1);
clear mm

% create N random numbers from 1 to N
N = size(tn,2);
F = ceil(N.*rand(1,N));
%data partition
ptr = pn(:,F(1:ceil(N*subsets(1))));
ttr = tn(:,F(1:ceil(N*subsets(1))));
pte = pn(:,F(ceil(N*subsets(1))+1:end));
tte = tn(:,F(ceil(N*subsets(1))+1:end));
% coordinates
xy_p = xy(:,F(1:ceil(N*subsets(1))));
xy_t = xy(:,F(ceil(N*subsets(1))+1:end));
figure(gcf+1); scatter(xy_p(1,:), xy_p(2,:), 'ko')
hold on; scatter(xy_t(1,:), xy_t(2,:), 'rx'); hold off;
%clear and save
clear N ans subsets

%% -3.\ TRAINING and SIMULATION      [WITH LM]

%----------------------------
% The neff_stack function creates five ANN with training algorithms: 'gdx',
% 'rp', 'bfg', 'oss', 'lm'.
% net = newff_stack(input_ranges, topology, activation_functions,initial_weights_range)
net = newff_stack([-1 +1], [size(p,1) 1], 'tt', 1);
%----------------------------
% The  'lm' method is selcted because more performant in preliminary
% analysis:
n = net.lm; clear net
TV.P = pte; TV.T = tte;
[n trainRec] = train(n, ptr,ttr,[],[],[],TV);
pred_ann = sim(n,pte);
c_ann = corr([tte' pred_ann']);
figure(2); scatter(pred_ann,tte);
hold on; line([-1 1],[-1 1]); hold off;
title(strcat(head_target(1), ' (corr=', num2str(c_ann(2,1)), ')'),'FontWeight','b', 'FontSize',14);
xlabel('ANN Inference', 'FontWeight','b', 'FontSize',12);
ylabel('Measured', 'FontWeight','b', 'FontSize',12)

%% -4.\ STEPWISE FIT: REGRESSION MODEL
% stepwisefit
[nan,nan,nan,inmodel,stats,nan,nan] = stepwisefit(ptr',ttr');
inmodel = find(inmodel==1);
stats.inmodel = inmodel;
clear nan inmodel
% initialize vector of predictions on testing subset PLUS intercept
pred_regr = zeros(size(tte,2),1) + stats.intercept;
for pred = 1:size(stats.inmodel,2)
    pred_regr = pred_regr + stats.B(stats.inmodel(pred))*pte(stats.inmodel(pred),:)';
end
```

```
pred_regr=pred_regr';

% plot
c_regr = corr([tte' pred_regr']);
figure(3); scatter(pred_regr,tte);
hold on; line([-1 0],[-1 0]); hold off;
title(strcat(head_target(1), ' (corr=', num2str(c_regr(2,1)), ')'), 'FontWeight','b', 'FontSize',14);
xlabel('Regression Inference', 'FontWeight','b', 'FontSize',12);
ylabel('Measured', 'FontWeight','b', 'FontSize',12)

%% -5.\ SAVE
cd('C:\Dottorato\PEDOMETRICS\DATA\UOT\MatLab')
eval(['save ANN_' head_target{1} '.mat'])
clear
%% -6.\ COMPARE MODELS: ANN vs MULTIPLE REGRESSION
clear
%-load
cd('C:\Dottorato\PEDOMETRICS\DATA\UOT\MatLab')
uiload;
%-correlation
c = corr([tte' pred_regr' pred_ann']);
%-VALUE PLOT
figure(4);
subplot(2,1,1);plot(1:size(tte,2),tte,':k','LineWidth', 2)
hold on;
subplot(2,1,1);plot(1:size(tte,2),pred_regr,'-r','LineWidth', 1)
subplot(2,1,1);plot(1:size(tte,2),pred_ann,'-b','LineWidth', 1)
hold off;
title('TESTING DATASET', 'FontWeight','b', 'FontSize',14)
%xlabel('Horizons', 'FontWeight','b', 'FontSize',12)
ylabel(head_target(1), 'FontWeight','b', 'FontSize',12)
L1 = legend('TARGET', ['REGR (corr=' num2str(c(2,1)) ')'],
['ANN (corr=' num2str(c(3,1)) ')']);
set(L1, 'FontWeight','b', 'FontSize',11)

%-error
e_reg = (pred_regr-tte);
e_ann = (pred_ann-tte);
%-ERROR PLOT
subplot(2,1,2);subplot(2,1,2);plot(1:size(tte,2),e_reg,':r','LineWidth', 2)
hold on;
subplot(2,1,2);plot(1:size(tte,2),e_ann,':b','LineWidth', 2)
line([0 size(pred_regr,2)],[0 0], 'Color','k', 'LineStyle','--')
hold off;
%title('TESTING DATASET')
xlabel('Horizons', 'FontWeight','b', 'FontSize',12)
ylabel(strcat('ERROR (', head_target(1), ')'), 'FontWeight','b', 'FontSize',12)
L2 = legend('Err\_Regr', 'Err\_ANN', 'Location','SE');
set(L2, 'FontWeight','b', 'FontSize',11)
clear
```

Index

Bibliography

B. Ahrens. Distance in spatial interpolation of daily rain gauge data. *Hydrology and Earth System Sciences*, 10:197–208, 2006. 80

H. Akaike. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723, 1974. 81

F.K. Aljibury and D. D. Evans. Soil sampling for moisture retention and bulk density measurements. *Proceedings - Soil Science Society of America*, 25: 180–182, 1961. 3

U. Anders and O. Korn. Models selections in neural netoworks. *Neural Networks*, 12:309–323, 1999. 81

K. E. Anderson and P. A. Furley. An assessment of the relationship between surface properties of chalk soils and slope form using principal component analysis. *Journal of Soil Science*, 26:130–143, 1975. 4

L. E. Andrew and F. W. Stearns. Soil sampling for moisture retention and bulk density measurements. *Proceedings - Soil Science Society of America*, 27:693–697, 1963. 3

O. Antonic, J. Krizan, A. Marki, and D. Bukovec. Spatio-temporal interpolation of climatic variables over large region of complex terrain using neural networks. *Ecological Modelling*, 138:255–263, 2001. 68, 70, 71, 75

H. Apaydin, F. K. Sonmez, and Y. E. Yildirim. Spatial interpolation techniques for climate data in the gap region in turkey. *Climate Research*, 28: 31–40, 2004. 69, 70

M. A. Arbib. Levels of modeling of mechanisms of visually guided behavior. *The Behavioral and Brain Sciences*, 10:407–465, 1987. 26

M. Armstrong. *Basic Linear Geostatistics*. Springer-Verlag, Berlin Heidelberg, Germany, 1998. 24

F. Attorre, M. Alf, M. De Sanctis, F. Francesconi, and F. Bruno. Comparison of interpolation methods for mapping climatic and bioclimatic variables at regional scale. *International Journal of Climatology*, 27:1825–1843, 2007. 69, 70, 71

B. Bacchi and N. Kottegoda. Identification and calibration of spatial correlation patterns of rainfall. *Journal of Hydrology*, 165:311–348, 1995. 68, 71, 80

F. G. Baker. Variability of hydraulic conductivity within and between nine wisconsin soil series. *Water Resources Research*, 14:103–108, 1978. 3

L. E. Band and I. D. Moore. Scale: Landscape attributes and geographical information systems. *Hydrological Processes*, 9:401–422, 1995. 2

C. L. Bascomb and M. G. Jarvis. Variability in three areas of the denchworth soil map unit. i. purity of the map unit and property variability within it. *J Soil Sci*, 27:420–437, 1976. 3

P. H. T. Beckett and R. Webster. Soil variability: A review. *Soils and Fertilisers*, 34(1):1–14, 1971. 3

T. Bellerby, M. Todd, D. Kniveton, and C. Kidd. Rainfall estimation from a combination of trmm precipitation radar and goes multispectral satellite imagery through the use of an artificial neural network. *Journal of Applied Meteorology*, 39(12):2115–2128, 2000. 71

C.M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, Oxford, 1995. 70

L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996. 73, 83

D. P Brown and A. C. Comrie. Spatial modeling of winter temperature and precipitation in arizona and new mexico, usa. *Climate Research*, 22: 115–128, 2002. 68, 70, 76, 81

C. Brunsdon, J. McClatchey, and D. J. Unwin. Spatial variations in the average rainfall-altitude relationship in great britain: an approach using geographically weighted regression. *International Journal of Climatology*, 21:455–466, 2001. 68, 70

P. A. Burrough and R. A. McDonell. *Principles of Geographical Information Systems*. Oxford University Press, New York, 1998. 32

J. J. Carrera-Hernandez and S. J. Gaskin. Spatio temporal analysis of daily precipitation and temperature in the basin of mexico. *Journal of Hydrology*, 336:231–249, 2007. 69, 70

D. K. Cassel and A. Bauer. Spatial variability in soils below depth of tillage: bulk density and fifteen atmosphere percentage. *Proceedings - Soil Science Society of America*, 39:247–250, 1975. 3

R. Celleri, P. Willems, W. Buytaert, and J. Feyen. Spacetime rainfall variability in the paute basin, ecuadorian andes. *Hydrological Processes*, 21: 3316–3327, 2007. 69

Y. M. Chiang, F. J. Chang, B. J. D. Jou, and P. F. Lin. Dynamic ann for precipitation estimation and forecasting from radar observations. *Journal of Hydrology*, 334:250–261, 2007. 71

P. Coulibaly and N. D. Evora. Comparison of neural network methods for infilling missing daily weather records. *Journal of Hydrology*, 341:27–41, 2007. 72

C. Daly, R. P. Neilson, and D. L. Phillips. A statistical-topographical model for mapping climatological precipitation over mountainous terrain. *Journal of Applied Meteorology*, 33:140–158, 1994. 68, 70

C. W. Dawson and R. L. Wilby. Hydrological modelling using artificial neural networks. *Progress in Physical Geography*, 25(1):80–108, 2001. 71, 85

F. M. Dekking, C. Kraaikamp, H. P. Lopuhaa, and L. E. Meester. *A Modern Introduction to Probability and Statistics*. Springer Texts in Statistics, Springer-Verlag, New York, 2005. 23

H. Demuth, M. Beale, and M. Hagan. *Neural Network Toolbox Users Guide*. Natick, MA, 2008. 85

A. Di Gennaro, A. D'Antonio, M.R. Ingenito, L. Lulli, G. Marseglia, F. Terribile, and L. Toderico. *I suoli della provincia di Napoli*. CUEN, Napoli, 1995. 9

N. Diodato and M. Ceccarelli. Interpolation processes using multivariate geostatistics for mapping of climatological precipitation mean in the sannio mountains (southern italy). *Earth Surface Processes and Landforms*, 30: 259–268, 2005. 69

G. Drogue, J. Humbert, J. Deraisme, N. Mahr, and N. Freslon. A statistical-topographic model using an omnidirectional parameterization of the relief for mapping orographic rainfall. *International Journal of Climatology*, 22: 599–613, 2002. 69, 70, 100, 108

W. J. Edmonds, J. C. Baker, and T. W. Simpson. Variance and scale influences on classifying and interpreting soil map units. *Soil Science Society of America Journal*, 49:957–961, 1985a. 2

W.J. Edmonds, J.B. Campbell, and M. Lentner. Taxonomic variation within three soil mapping units in virginia. *Soil Science Society of America Journal*, 49:394–401, 1985b. 3

C. Frei and C. Schar. A precipitation climatology of the alps from high-resolution rain-gauge observations. *International Journal of Climatology*, 18:873–900, 1998. 68, 70

M. Freiwan and M. Kadioglu. Spatial and temporal analysis of climatological data in jordan. *International Journal of Climatology*, 28:521–535, 2008. 69, 70

L. Gardin, R. Napoli, F. Primavera, E. Gregori, and E. Costantini. *Guida al Rilevamento dei Suoli, Version II.* UOT Project, November 1995. 12

C. L. Goodale, J. D. Aber, and S. V. Ollinger. Mapping monthly precipitation temperature and solar radiation for ireland with polynomial regression and digital elevation model. *Climate Research*, 10:35–49, 1998. 68, 70

P. Goovaerts. *Geostatistics For Natural Resources Evaluation.* Oxford University Press, New York, 1997. 25, 50, 70, 90

P. Goovaerts. Geostatistical approaches for incorporating elevation into the spatial interpolation of rainfall. *Journal of Hydrology*, 228:113–129, 2000. 68, 70

M. Guler, B. Cemek, and H. Gunal. Assessment of some spatial climatic layers through gis and statistical analysis techniques in samsun turkey. *Meteorological Applications*, 14:163–169, 2007. 69

D. Gyalistras. Development and validation of a high-resolution monthly gridded temperature and precipitation data set for switzerland (19512000). *Climate Research*, 25:55–83, 2003. 69, 80, 81

M. T. Hagan and M. Menhaj. Training feed-forward networks with the marquardt algorithm. *IEEE Transactions on Neural Networks*, 5(6):989–993, 1994. 85

L. C. Hammond, W. L. Pritchett, and V. Chew. Soil sampling in relation to soil heterogeneity. *Soil Science Society of America Journal*, 22:548–552, 1958. 3

L. K. Hansen and Salamon P. Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 12(10):993–1001, 1990. 89

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction*. Springer Series in Statistics, Springer-Verlag, New York, 2001. 23

R. L. Hay. Rate of clay formation and mineral alteration in a 4000-years-old volcanic ash soil on st. vincent, b.w.i. *American Journal of Science*, 258: 354–368, 1960. 4

S. Haykin. *Neural Networks: A Comprehensive Foundation*. Prentice-Hall, Inc., second edition, July 1998. 26, 70, 75

T. Hengl, S. Gruber, and D. P. Shrestha. Digital terrain analysis in ilwis, August 2003. URL http://www.itc.nl/personal/shrestha/DTA/. 32

T. Hengl, G. B. M. Heuvelink, and D. G. Rossiter. About regression-kriging: From equations to case studies. *Computers & Geosciences*, 33(10):1301 – 1315, 2007. 25, 46

T. Hengl and G. Langella. Mapping soil colour from munsell colour chart codes. In *Pedometrics 2007*, page 11, University of Tuebingen, Institute of Geography, 27-30 August 2007. Pedometrics. 50

J. A. Hevesi and J. D. Flint, A. L. ans Istok. Precipitation estimation in mountainous terrain using multivariate geostatistics. part ii. isohyetal maps. *Journal of Applied Climatology*, 31:667–688, 1992. 68, 70

W. W. Hsieh and B. Y. Tang. Applying neural network models to prediction and data analysis in meteorology and oceanography. *Bulletin of the American Meteorological Society*, 79(9):1855–1870, 1998. 75

B.D. Hudson. The soil survey as paradigm-based science. *Soil Sci. Soc. Am. J.*, 56:836–841, 1992. 2

R. D. Hunter and R. K. Meentemeyer. Climatologically aided mapping of daily precipitation and temperature. *Journal of Applied Meteorology*, 44 (10):1501–1510, 2005. 70

W.C. Jacob and A. Klute. Sampling soils for physical and chemical properties. *Proceedings - Soil Science Society of America*, 20:170–172, 1956. 3

H. Jenny. *Factors of Soil Formation - A System of Quantitative Pedology*. McGraw-Hill, New York, 1941. 2, 4

D. I. Jeong and Y. O. Kim. Rainfall-runoff models using artificial neural networks for ensemble streamflow prediction. *Hydrological Processes*, 19: 3819–3835, 2005. 71

Y. Jia and T. B. Culver. Bootstrapped artificial neural networks for synthetic flow generation with a small data sample. *Journal of Hydrology*, 331:580–590, 2006. 82

G. L. Johnson and C. L. Hanson. Topographic and atmospheric influences on precipitation variability over a mountainous watershed. *Journal of Applied Meteorology*, 34(1):68–87, 1995. 68, 70

M. J. Jones. The organic matter content of the savanna soils of west africa. *Journal of Soil Science*, 24:42–53, 1973. 4

J. Joss and R. Lee. The application of radar-gauge comparisons to operational precipitation profile corrections. *Journal of Applied Meteorology*, 34:2612–2630, 1995. 71

A. G. Journel and Ch. J. Huijbregts. *Mining Geostatistics*. Academic Press, New York, 1978. 25, 70

S. A. R. Kumar, K. P. Sudheer, S. K. Jain, and P. K. Agarwal. Rainfall-runoff modelling using artificial neural networks: comparison of network types. *Hydrological Processes*, 19:1277–1291, 2005. 71, 81

P. C. Kyriakidis, J. Kim, and N. L. Miller. Geostatistical mapping of precipitation from rain gauge data using atmospheric and terrain haracteristics. *Journal of Applied Meteorology*, 40:1855–1877, 2001. 68, 70

P. W. Lane. Generalized linear models in soil science. *European Journal of Soil Science*, 53:241–251, 2002. 5, 23

C. D. Lloyd. Assessing the effect of integrating elevation data into the estimation of monthly precipitation in great britain. *Journal of Hydrology*, 308:128–150, 2005. 69, 70

L. Lulli. *I suoli caposaldo dell'apparato vulcanico di Vico.* Ministero dell'Agricoltura e delle Foreste, Istituto Sperimentale per lo Studio e la Difesa del Suolo, Firenze, 1990. 9

R. A. MacMillan. *LandMapR Software Toolkit C++ Version: User Manual.* Edmonton, Alberta, 2003. 21, 22

R. A. MacMillan, W. W. Pettapiece, S. C. Nolan, and T. W. Goddard. A generic procedure for automatically segmenting landforms into landform elements using dems, heuristic rules and fuzzy logic. *Fuzzy Sets and Systems*, 113:81–109, 2000. 21, 22, 32

D. L. Mader. Soil variabilitya serious problem in soil-site studies in the northeast. *Soil Science Society of America Journal*, 27:707–709, 1963. 3

T. Maeda, H. Takenaka, and B.P. Warkentin. Physical properties of allofane soils. *Advances Agronomy*, 29:229–264, 1977. 9

H. R. Maier and G. C. Dandy. Neural network based modelling of environmental variables: A systematic approach. *Mathematical and Computer Modelling*, 33(2661):669–682, 2001. 87

J. Marquinez, J. Lastra, and P. Garcia. Estimation models for rainfall in mountainous regions: the use of gis and multivariate analysis. *Journal of Hydrology*, 270:1–11, 2003. 69, 70

W. L. Martinez and A. R. Martinez. *Explorative data analysis with MatLab®.* Computer Science and Data Analysis. Chapman & Hall/CRC Press, U.K., 2005. 32

A. Martnez-Cob. Multivariate geostatistical analysis of evapotranspiration and precipitation in mountainous terrain. *Journal of Hydrology*, 174:19–35, 1996. 68, 70

T. Masters. *Practical Neural Network Recipes in C++.* Academic Press, San Diego, 1993. 87

G. Matheron. The intrinsic random functions and their applications. *Adv. Appl. Prob.*, 5:239–465, 1973. 24

A. B. McBratney, M. L. Mendonca Santos, and B. Minasny. On digital soil mapping. *Geoderma*, 117:3–52, 2003. 4, 24

A. B. McBratney, I. O. A. Odeh, T. F. A. Bishop, M. S. Dunbar, and T. M. Shatar. A review of pedometric techniques for use in soil survey. *Geoderma*, 97(3-4):293–327, 2000. 4

D. E. McCormack and L. P. Wilding. Variation of soil properties within mapping units of soils with contrasting substrata in northwestern ohio. *Proceedings - Soil Science Society of America*, 33:587–593, 1969. 3

N. J. McKenzie and P. J. Ryan. Spatial prediction of soil properties using environmental correlation. *Geoderma*, 89(1-2):67–94, 1999. 4, 24, 45, 53

J. Meersmans, F. De Ridder, F. Canters, S. De Baets, and M. Van Molle. A multiple regression approach to assess the spatial distribution of soil organic carbon (soc) at the regional scale (flanders, belgium). *Geoderma*, 143:1–13, 2008. 33

G. Milne. Some suggested units of classification and mapping particularly for east african soils. *Soil Research*, 4:183–198, 1935. 20

MIPAF, Ministero delle Politiche Agricole e Forestali, and Osservatorio Nazionale Pedologico e per la Qualit del Suolo. *Metodi di analisi chimica del suolo*, 2000. 13

C. Mizota and L. P. Van Reeuwijk. Clay mineralogy and chemistry of soils formed in volcanic material in diverse climatic regions, 1989. 13

I. D. Moore, R. B. Grayson, and A. R. Landson. Digital Terrain Modelling: A Review of Hydrological, Geomorphological, and Biological Applications. *Hydrological Processes*, 5:3–30, 1991. 32

I.D. Moore, P.E. Gessler, G.A. Nielsen, and G.A. Peterson. Soil attribute prediction using terrain analysis. *Soil Science Society of America Journal*, 57(2):443–452, 1993. 20

L. A. Nelson and R. J. McCracken. Properties of norfolk and portsmouth soils: Statistical summarization and influence on corn yields. *Proceedings - Soil Science Society of America*, 26:497–502, 1962. 3

W.D. Nettleton, B.R. Brasher, and G. Borst. The taxadjunct problem. *Soil Science Society of America Journal*, 55:421–427, 1991. 3

M. Ninyerola, X. Pons, and J. M. Roure. Monthly precipitation mapping of the iberian peninsula using spatial interpolation tools implemented in a geographic information system. *Theoretical and Applied Climatology*, 89: 195–209, 2007. 69, 70

I. Noy-Meir. Multivariate analysis of the semiarid vegetation in south-eastern australia: Ii. vegetation catena and environmental gradients. *Australian Journal of Botany*, 22:115–140, 1974. 4

I. O. A. Odeh and A. B. McBratney. Using avhrr images for spatial prediction of clay content in the lower namoi valley of eastern australia. *Geoderma*, 97:237–254, 2000. 46

I. O. A. Odeh, A. B. McBratney, and D. J. Chittleborough. Spatial prediction of soil properties from landform attributes derived from a digital elevation model. *Geoderma*, 63(3-4):197–214, 1994. 46

P. Oettli and P. Camberlin. Influence of topography on monthly rainfall distribution over east africa. *Climate Research*, 28:199–212, 2005. 69, 70, 79

E. Pardo-Iguzquiza, P. A. Dowd, and D. I. F. Grimes. An automatic moving window approach for mapping meteorological data. *International Journal of Climatology*, 25:665–678, 2005. 69

E. Pardo-Igzquiza. Comparison of geostatistical methods for estimating the areal average climatological rainfall mean using data on precipitation and topography. *International Journal of Climatology*, 18(9):1031–1047, 1998. 68, 70

S. J. Park, K. McSweeney, and B. Lowery. Identification of the spatial distribution of soils using a process-based terrain characterization. *Geoderma*, 103:249–272, 2001. 54

S. J. Park and P. L. G. Vlek. Environmental correlation of three-dimensional soil spatial variability: a comparison of three adaptive techniques. *Geoderma*, 109:117–140, 2002. 5

E. J. Pebesma and C. G. Wesseling. Gstat, a program for geostatistical modelling, prediction and simulation. *Computers & Geosciences*, 24(1): 17–31, 1998. 33

T. C. Peterson, R. Vose, R. Schmoyer, and V. Razuvav. Global historical climatology network (ghcn) quality control of monthly temperature data. *International Journal of Climatology*, 18:1169–1179, 1998. 67

J. C. Powel and M. E. Springer. Composition and precision of classification of several mapping units of the appling, cecil, and lloyd series in walton county, georgia. *Proceedings - Soil Science Society of America*, 29:454–458, 1965. 3

C. Prudhomme and D. W. Reed. Mapping extreme rainfall in a mountainous region using geostatistical techniques: a case study in scotland. *International Journal of Climatology*, 19:1337–1356, 1999. 68, 70

P. Quantin. Andisols, 1990. 9

P. Quinn, K. Beven, P. Chevallier, and O. Planchon. The prediction of hillslope flow paths for distributed hydrological modelling using digital terrain models. *Hydrological Processes*, 5:59–79, 1991. 22

S. V. T. Ramesh and V. Chandramoulia. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. *Journal of Hydrology*, 312:191–206, 2005. 72

E. M. Rasmusson and P. A. Arkin. A global view of large-scale precipitation variability. *Journal of Climate*, 6:1495–1522, 1993. 71

R. A. V. Rossel, B. Minasny, P. Roudier, and A. B. McBratney. Colour space models for soil science. *Geoderma*, 133:320–337, 2006. 35, 38

U. Schwertmann. Differenzierung der eisenoxide des boden 5 durch extraktion mit ammoniumoxalat-l6sung. *Z Pflanzenernahr Dang Bodenk*, 105:194–202, 1964. 13

P. Scull, J. Franklin, O. A. Chadwick, and D. McArthur. Predictive soil mapping: a review. *Progress in Physical Geography*, 27(2):171–197, 2003. 5

P. A. Shary, L. S. Sharaya, and A. V. Mitusov. Fundamental quantitative methods of land surface analysis. *Geoderma*, 107:1–32, 2002. 20, 32

S. S. P. Shen, P. Dzikowski, G. Li, and D. Griffith. Interpolation of 196197 daily temperature and precipitation data onto alberta polygons of ecodistrict and soil landscapes of canada. *Journal of Applied Meteorology*, 40 (12):2162–2177, 2001. 68, 70

E. Sicard and R. Sabatier. Theoretical framework for local pls1 regression, and application to a rainfall data set. *Computational Statistics and Data Analysis*, 51:1393–1410, 2006. 69, 100, 109

R.W. Simonson. Outline of a generalized theory of soil genesis. *Soil Sci. Soc. Am. Proc.*, 23:152–156, 1959. 20

A. Skidmore. *Environmental modelling with GIS and Remote Sensing*. Taylor & Francis, London, 2002. 6

Soil Survey Division Staff. *Soil Survey Manual*. U.S. Gov. Print. Office, Washington, D.C., 1993. 1

J. Tayman and D. A. Swanson. On the validity of mape as a measure of population forecast accuracy. *Population Research and Policy Review*, 18: 299–322, 1999. 81

T. H. Thornburn and W. R. Larsen. A statistical study of soil sampling. *Proc. ASCE Jour. Soil Mech. And Found. Div*, 85(SM5):1–13, 1959. 3

D. Tsintikidis, J. L. Haferman, E. N. Anagnostou, W. F. Krajewski, and T. F. Smith. A neural network approach to estimating rainfall from spaceborne microwave data. *IEEE Transactions on Geoscience and Remote Sensing*, 35(5):1079–1093, 1997. 71

L. H. Tsoukalas and R. E. Uhrig. *Fuzzy and Neural Approaches in Engineering*. John Wiley and Sons, Inc., New York, 1997. 26, 70

S. M. Vicente-Serrano, M. A. Saz-Sanchez, and J. M. Cuadrat. Comparative analysis of interpolation methods in the middle ebro valley (spain) application to annual precipitation and temperature. *Climate Research*, 24:161–180, 2003. 69, 81

H. Wackernagel. *Multivariate geostatistics: An introduction with applications*. Springer, New York, 2003. 70

R. E. White. *Principles and Practice of Soil Science: The Soil as a Natural Resource*. Wiley-Blackwell, 2005. 1

L. P. Wilding, R. B. Jones, and G. M. Schafer. Variation of soil morphological properties within miami, celina, and crosby mapping units in west-central ohio. *Proceedings - Soil Science Society of America*, 29:711–717, 1965. 3

L. P. Wilding, G. M. Schafer, and R. B. Jones. Morley and blount soils: A statistical summary of certain physical and chemical properties of some selected profiles from ohio. *Proceedings - Soil Science Society of America*, 28:674–679, 1964. 3

C.J. Willmott. On the validation of models. *Physical Geography*, 2:184–194, 1981. 81

C.J. Willmott. Some comments on the evaluation of model performance. *Bulletin of the American Meteorological Society*, 63:1309–1313, 1982. 81

J. P. Wilson and J. C. Gallant. *Terrain analysis: principle and application*. John Wiley & Sons, New York, 2000. 20

D.H. Wolpert. Stacked generalization. *Neural Networks*, 5:241–259, 1992. 83

G. Wyszecki and W. S. Stiles. *Color Science: Concepts and Methods, Quantitative Data and Formulae*. Wiley, New York, second edition, 1982. 35

Y. Xia, P. Fabian, A. Stohl, and M. Winterhalter. Forest climatology: estimation of missing values for bavaria, germany. *Agricultural and Forest Meteorology*, 96:131–144, 1999a. 72

Y. Xia, M. Winterhalter, and P. Fabian. A model to interpolate monthly mean climatological data at bavarian forest climate stations. *Theoretical and Applied Climatology*, 64:27–38, 1999b. 68

R. Xiao and V. Chandrasekar. Development of a neural network based algorithm for rainfall estimation from radar observations. *IEEE Transactions on Geoscience and Remote Sensing*, 35(1):160–171, 1997. 71

Yaalon. Conceptual models in pedogenesis: can soil-forming functions be solved? *Geoderma*, 14:189–205, 1975. 4

L. A. Zadeh. Fuzzy logic: Issues, contentions and perspectives (abstract). In *ACM Conference on Computer Science*, page 407, 1994. 26

L. W. Zeverbergen and C. R. Thorne. Quantitative analysis of land surface topography. *Earth Surface Processes and Landforms*, 12:47–56, 1987. 32

J. Zhang. Developing robust non-linear models through bootstrap aggregated neural networks. *Neurocomputing*, 25:93–113, 1999. 83, 89

Q. Zhou, B. Lees, and G. Tang. *Advances in digital terrain analysis*. Springer-Verlag, Berlin Heidelberg, 2008. 19

Z. H. Zhou, J. Wu, and W. Tang. Ensembling neural networks: Many could be better than all. *Artificial Intelligence*, 137:239–263, 2002. 90