The problem of missing data in hydroclimatic time series. Applicationof spatial interpolation techniques to construct a comprehensive of hydroclimatic data in Sicily, Italy.

by Annalisa Di Piazza

February 15, 2011

Contents

List of Figures

List of Tables

Introduction

Planning, management and effective control of water resource systems, require a considerable amount of hydrological data variables such as rainfall, temperature, streamflow, etc.. Such data are required when a hydrological model has to be developed, as well. Very often hydrological data sequences at a given gauge have gaps or are incomplete, or are not characterized by a good quality or are not sufficiently length. This can severely affect, for example, the reliability of the design of a hydropower plant, the construction of dams, etc. Furthermore, the problem of missing values is a common obstacle in time series analysis and specifically in the context of rainfall, temperature and rainfall–runoff processes modelling.

There may be various reasons for missing values, for instance equipment failure, errors in measurements or faults in data acquisition, and natural hazards such as landslides, or even temporary absence of observers, the cessation of measurement or absence of observations prior to the commencement of measurement or by limited financial resources. Whatever the reasons, missing values produce a significant problem for water resources applications. Consequently, finding efficient methods to deal with the problem of missing values is an important issue in most hydrological analyses. However, hydrological modellers commonly discard the observations with missing values and only use the observations with complete information, which means that a lot of information contained in the dataset is lost. Furthermore, the approach is inadequate for analyses that require serially complete data. On the other hand, the use of the dataset prone to missing data can result in errors that exhibit temporal and spatial patterns (Stooksburry et al., 1999[Stooksbury1999]). As an alternative to this listwise(?) deletion procedure, modellers sometimes replace (or "fill in") a value for the missing values by using, for example, the mean of the observed variables. Such a procedure could, however, seriously distort statistical properties like standard variation, correlations or percentiles. But the best alternative to the above mentioned

approaches consists of filling the gaps in the rainfall, temperature or streamflow time series by estimating the missing values.

In fact, the most common approach followed in technical literature is the application of either deterministic or stocastic methods to estimate the missing data. The described problem is so dramatically important within hydrological research that the scientific community has point out a transnational initiative of international research groups, i.e., the "Decade on Prediction in Ungauged Basins (PUB)", a wide research project promoted by the International Association of Hydrological Sciences (Sivapalan et al., 2003). This effort is particularly focused on the reconstruction of serially incomplete data records in basins with short streamflow records or in ungauged river basins.

In this scenario, the individuation and application of the most suitable methods for the accurate estimation of hydrological variable values, useful to fill in the incomplete time series, is of paramount importance and represents the most promising approach to solve the problem of missing data. In particular, once the hydrological variable is defined together with its specific characteristic, the choice of the estimation methods and their comparison is necessary to carry out the best reconstruction of the considered variable dataset.

The issue of gaps in climatic variables have been the subject of a large number of scientific works where numerous techniques for estimating missing data values have been implemented and compared. Among these methods, the temporal methods, that taking into account the temporal dependence of the considered variables, have been used and among more advanced methods, the space-time models, i.e. models handling dependence the spatial and temporal simultaneously, has been applied. A group of methods that are also widely used in literature for the missing data estimation are the spatial models which represent the spatial distribution of variables over a specific duration. Many papers have been dedicated to the comparison between deterministic and stochastic approaches to reconstruct data records and their results suggest that often the use of geostatistical techniques improve the results since they are able to study the pattern of spatial dependences observed for climatic variables; in the particular case of estimation of runoff, in some works it was highlighted that considering the runoff as an areal process, i.e., considering the strongly dependence of runoff with the basin area, improves by far the estimates obtained. Another consideration is common to many works, i.e., that the use of algorithms that incorporate ancillary informations (geographical and morphological) into the spatial estimation of climatic variables improves the obtained estimates.

The aim of this thesis is to investigate the methods for the optimal estimation of the missing data in time series of hydrological variables with reference to Sicily (Italy). In particular, the following hydrological variables are object of study: precipitation, temperature and runoff.

In this thesis only the spatial structural dependence of rainfall, temperature and runoff data is used to reconstruct missing data, neglecting the spatial-temporal dependence.

On the basis of the variables specified, different estimation methods have been considered, described and applied to solve the problem of missing data. With regard to the variables as precipitation and temperature, that can be represented as point processes, the following algorithms, used for the spatial interpolation, will be applied: inverse distance weighting, radial basis function with thin plate spline, simple linear regression, multiple regression, geographically weighted regression, artificial neural network, ordinary kriging, residual ordinary kriging. With the applications of these methods, serially complete monthly and annual dataset will be obtained.

On the other hand, for the runoff, the proposed investigation stems from the consideration that it can be described as an areal process. With this assumption, a more accurate estimation of the considered variable can be obtained. This approach has very few examples in scientific literature but appears to be very promising in the considered field. For this reason the estimation method, chosen for the runoff, is a stochastic method to derive gridded maps for finer and finer resolution with a geostatistical approach. It is, in particular, a stochastic interpolation system that can be assimilated to kriging system with the explicit consideration of the runoff variable as an areal process. The application of this methods will give the annual runoff estimated data for the stations that have been out of work in the chosen time window of input runoff data and that are characterised by a dataset affected by missing data. Moreover, it will be possible to obtain the annual runoff estimated values also for the areas of the basins not provided with gauge stations. The latter values can be obtained by the gridded map with a certain resolution.

It is important to highlight that the previous applications of such an approach are done in homogeneous climatic contexts with favorable conditions of the flow regime to apply the procedure. On the contrary, here, for the first time, the method is applied in the Sicilian context where both the climatic and morphological profiles are strongly inhomogeneous.

Chapter 1

Missing data in the climatic variables time series

In the scientific community there is great demand of a complete, accurate and reliable source of climate data. This is, as matter of fact, a prerequisite for the efficient modelling of a wide variety of environmental processes. In addition, a reliable estimate and an accurate prediction of weather events, or any environmental phenomenon may aid in decisions and management related to the national economy and security.

A complete climate archive that responds to requests from all over the world, provides historical perspectives on climate which are vital to studies on global climate change, the greenhouse effect, and other environmental issues. A such archive, that stores this kind of information, can be also essential to industry, agriculture, science, hydrology, transportation, recreation, and engineering.

1.1 Climatic Archives

A climate archive is a permanent collection of data or records of climatic variables designed to be maintained for a long time. This is generally a set of climatic informations provided at different time scale. Typical examples of climatic variables dataset are the collections of daily, monthly and annual series of precipitation, temperature, runoff, snowfall, wind, sky cover, weather conditions, relative humidity, evaporation, soil moisture recorded by the gauge station in a state, a nation, a region, etc.

As before said, a reliable archive of data is of crucial importance for the many

sectors, among which: agriculture, civil infrastructure, construction, coastal hazards, energy, eealth, insurance, litigation, marine and coastal ecosystems, national security, tourism, transportation, water resources. Indeed, from a climatic dataset, through using of physically-based, mathematical or stochastic models, it is possible to translate climate data into accessible, useful, and accurate products for each of the above mentioned sectors. For the water resources sector, for example, a complete dataset of some climatic variables can be useful in a variety of considerations relative to action to be taken for better management of resources; for example:

- using short-duration rainfall values and raingauge charts to design retarding basins that will help reduce stormwater-borne pollutants;
- using the amount, location, and duration of rainfall from a heavy precipitation event to define the magnitude of a storm in order to assess and estimate property damage;
- using drought information to determine when water rationing may be required in areas where lake levels are declining;
- using temperature and snowpack trends to determine changes in the timing of runoff. Warmer temperatures cause snowpack to melt earlier in the spring, causing lower streamflow later in the summer.

For the field of Energy, the climate information can be used in this ways:

- using global surface hourly data for studies of wind energy potential to drive wind turbines for electricity generation;
- using solar radiation data to estimate solar energy potential;
- using temperature information to aid in the assessment of equipment requirements for heavy power line loads during extremely hot weather. NOAA's National Climatic Data Center Sectoral Engagement Fact Sheet ENERGY;
- using hourly temperature, relative humidity (and/or dew point), cloud cover, precipitation, and wind speed and direction data in electric load forecasting models and scenario analysis, for use by utility and power trading companies;
- using heating/cooling degree day data - measures of expected energy usage for heating and cooling, based on cumulative daily average temperature observations below (heating degree days) or above (cooling degree days) a specific threshold, typically 65°F - to help energy regulators determine what rates electric utilities can charge their customers;

Given the importance of having complete and reliable climatic archives in a nation (or in a region), as early as the first of the '800, in Europe and in America, the first agencies dedicated specifically to the collection of data began to be created. For

example, the oldest agenciescreated were the United States Coast and Geodetic Survey formed in 1807, the Weather Bureau formed in 1870, and the Bureau of Commercial Fisheries formed in 1871.

One of the most important scientific agency, focused on the condition of the oceans and atmosphere, is the *National Oceanic and Atmospheric Administration*(NOAA), belonging to the *United States Department of Commerce*, which was formed in 1970. The NOAA services provide to supply the nation with weather forecasts and nautical charts, conserves and manages marine species, restores and enables state and local partners to restore degraded coastal habitats, and conducts the research necessary to improve these and a host of other products and services.

The NOAA climate services provide the climatic data from NOAA's distributed climate service community. These kind of data are subdivided in two macro-sections: *Past & Present Climate*and *Prediction.* For both sections, data of the United States and the entire Globe are presented.

In the *Past & Present Climate*section, for many station across the United States, climate information in daily, monthly and annual scale are available. In particular, different types of reports can be consulted, among these:

- *The Daily and Monthly Climate Report.* This report is issued daily for many stations across the United States by the *National Weather Service*(NWS). This product contains useful climate data including daily temperature, precipitation, snowfall, wind, sky cover, weather conditions, relative humidity.

- *Regional Temperature and Precipitation Summary*(RTPS). The RTPS table issued by NWS is a collection of daily maximum and minimum temperature and precipitation readings from many stations across a region.

- *Soil Moisture Monitoring.* This dataset includes a serie of digital maps displaying Climate Prediction Center (CPC) calculated soil moisture (total, anomalies and percentiles), evaporation and runoff (mean, anomalies and percentiles) for most recent periods.

- *The Preliminary Local Climatological Data*product issued by NWS contains daily updates to the current months climatological data, including temperature and precipitation readings for many station across the United States.

- *State of the Climate- U.S. National Overview.* The *U.S. National Overview*is a framework that summarize U.S. National Observation of surface temperature and precipitation data by placing that data into historical perspective. The

U.S. National Overview provides access to monthly, 3 month/seasonal, 6 months, 12months, 12 month and annual climate summaries by state, division and region.

- Annual Climatological Summary. This is a publication service created and archived at the National Climatic Data center (NCDC). This product contains monthly and annual summaries for over 8000 U.S. locations.

For the entire Globe the available data are:

- The global Analysis Page. It is a framework that summarize global observations of temperature and precipitation data placing the data into historical perspective. It provides monthly, seasonal, and annual climate summaries, including: Global Surface Temperature and Precipitation.

- Global Hazards/Climate Extremes. It highlights significant weather events around the world by month and year as a framework to summarize weather relates hazards and disasters across the world. The information provided is a compilation of media news and other reports accompanied by revelant images, graphs and documents.

- The *Climate Diagnostic Bulletin*(CDB). It is a monthly compilation of meteorological and oceanographic analyses issued by the NOAAs *Climate Prediction Center*(CPC). The CDB is designed for use by scientists and decision makers with a technical background in climate variability.

- The unified global gauge daily precipitation analysis is part of the *Climate Prediction Center*(CPC) Unified Precipitation Products Suite. The daily gauge analysis is created on a 0,5 resolution grid over the global land by interpolating gauge observations from approximately 30000 stations.

This huge amount of data coming from agencies belonging to NOAA are collected at the *National Climatic Data Center*(NCDC). The NCDC is the world's largest active archive of weather data. Its aim is to provide access and stewardship to the U.S. resource of global climate and weather related data and information, and assess and monitor climate variation and change. This effort requires the acquisition, quality control, processing, summarization, dissemination, and preservation of a vast array of climatological data generated by the national and international meteorological services. NCDC's provides the U.S. climate representative to the *World Meteorological*

Organization, the *World Data Center System*, and other international scientific programs.

Climate information is often available only as raw observations or in the form of tables, graphs, or written summaries, which may be difficult for users who are not well-versed in climate science to fully interpret. To bridge this gap, NCDC is partnering with the different sector of interest to translate climate data into accessible, useful, and accurate products; and to leverage NCDC's climate expertise to better understand what the information means and how it can be used most effectively.

In particular, NCDC maintains a repository of weather model output data sets, model input data sets (assimilation), and climate model data sets. These data sets include:

- *Numerical Weather Prediction*(NWP) model input produced by the National Centers for Environmental Prediction (NCEP);
- NWP gridded model output produced by (NCEP);
- NCEP Reanalysis; and
- *Global Climate Model*(GCM) model data from the Geophysical Fluid Dynamics Laboratory (GFDL).

NCDC receives the NWP gridded data directly from NCEP and provides real-time access to the weather model forecast data. In addition to real-time NCEP model data, our repository also includes historical data from May 2002 to present. NCDC serves as a portal that allows user access to numerical weather prediction model data and global climate model data at many other sites on the web in addition to the NCDC site.

Improving quality control and continuity of these new data sets as well as making them available in timely fashion has been paramount. As operator of the *World Data Center for Meteorology*, Asheville, which provides for international data exchange, NCDC also collects data from around the globe. The NCDC has more than 150 years of data on hand with 224 gigabytes of new information added each day that is equivalent to 72 million pages a day.

NCDC archives 99 percent of all NOAA data, including over 320 million paper records; 2.5 million microfiche records; over 1.2 petabytes of digital data residing in a mass storage environment. NCDC has satellite weather images back to 1960. NCDC annually publishes over 1.2 million copies of climate publications that are sent to individual users and 33,000 subscribers. NCDC maintains over 500 digital data sets, receives almost 2,000,000 requests each year, and records over 100 million hits per year on the website. Data are received from a wide variety of sources, including

satellites, radar, remote sensing systems, *National Weather Service*(NWS) cooperative observers, aircraft, ships, radiosonde, wind profiler, rocketsonde and solar radiation networks.

Another important Europe-wide scientific agency is *European Environment Agency*(EEA). Its task is to provide sound, independent information on the environment. It is a major information source for those involved in developing, adopting, implementing and evaluating environmental policy, and also the general public. Currently, the EEA has 32 member countries. The regulation establishing the EEA was adopted by the European Union in 1990. It came into force in late 1993 immediately after the decision was taken to locate the EEA in Copenhagen. Work started in earnest in 1994. The regulation also established the European environment information and observation network (Eionet). EEA was created to help the Community and member countries make informed decisions about improving the environment, integrating environmental considerations into economic policies and moving towards sustainability, to coordinate the European environment information and observation network.

These are just two examples of worldwide scientific agency that provides a considerable climatic dataset.

An important national wide scientific agency is the Italian Air Force Meteorological Service (Servizio Meteorologico dell'Aeronautica Militare). In 1865 a *Central Metereologic Office*(Ufficio Centrale Meteorologico) at the *Navy Government*was constituted and following two Decrees of the President of the Italian Republic The *Air Force*has become the manager of the *Meteorological Service*. The tasks of an organ of this type are those in favor of the civil community, in particular in the field of civil protection and preservation of life at sea and in other important areas of life in the country, such as research, information, environment, transport, agriculture and energy exploitation. In particular, the *National Center for Aeronautical Meteorology and Climatology*, CNMCA, is the operational arm of the *Central Air Force Meteorological Service*. It provides for the receipt, processing and dissemination, national and international data and weather information, conventional and from satellite. It prepares and distributes analysis and forecasts for the basic needs of the Armed Forces and the general users. It produces and disseminates weather warnings for Civil Protection and the Safety of Life at Sea. The CNMCA manages databases and provides weather analysis, design and development of new products for the needs of the Public Service. The CNMCA performs basically three services. The first service is the *Analysis and Forecasting service*: it follows the evolution of weather conditions with absolute continuity of space and time, making very short-term forecasts (up to 24

hours), short (24-48 hours) and medium term (from 48 hours to 5 days), in Italy; it issues warnings of weather forecast risk for the national community and submit them to the Civil Defence; it issues the bulletin of the sea and storm and gale warnings for the protection of life at sea. The second service is about the *Meteorological Applications*: it develops and implements software applications necessary for the operational chain to administer and run properly (integration and conventional and unconventional data analysis, development and implementation of algorithms and numerical prediction models for air, sea and land, computer processing for the realization of weather products for military and civilian , graphical representation of the information and products). This service performs also the quality control of production in real-time weather. The numerical models are widely used in CNMCA, which provides for the development of advanced computing resources.

The third service is about the *Climatology*: it provides for the collection, storage and updating of meteorological data; it makes climate studies for the specific needs aeronautical, naval and ground; it manages the permanent data storage system, ensuring their quality; it provides historical data for users of the Armed Forces and external public and private. The fourth service is about the *Data Processing and satellite reception*: it manages operationally, with absolute continuity, the weather data processing center; it manages the centralization and distribution of information and conventional weather satellite, through the use of computers; it performs the function of RTH (*Regional Telecommunication Hub*) in the GTS (*Global Telecommunication System*), the WMO global network for exchanging data. It is responsible for developing and managing the websites of the Meteorological Service and computer networks for data transport.

The CNMCA then maintain and make available, to public and private users, observations, measurements and forecast calculations made in Italy by the various military agencies from setting-up of the Air Force. The electronic archive consists of several databases (DB), in which the data are organized by station from the year 1950. Another relational DB of the stations, shared with other branchs of the Center, contains more detailed historical information on location, equipment, operation, administration and other (metadata) for each observing site. Currently the follow data are stored and managed:

- bulletins of the ground observations (METAR / SPP, SYNOP, syrep);
- bulletins of the elevation observations;
- reports of maritime observations / forecasts (meteomar, gale warnings);

- reports of mountain and land observations / forecasts (Meteomont notices warning);
- reports of special observations (CO2, O3, radiation, sunlight, UV);
- detection of lightning (sfuk, lampinet);
- key images from satellites and weather radar;
- maps, charts and numerical data (grib format) analysis.

The Center also ensures the preservation of original records and microfilm in a paper archive which is an asset of national importance of enormous historical and scientific value. The archive is currently located at the airport area of the airport "*Mario de Bernardi Pratica di Mare*"; an ongoing census activities, accommodation, recovery and preservation of material plays out. It is proceeding in parallel, through tendering and procurement in self-employment, the translation of data to digital media, and then to the complete digitization of paper and microfilmed material and the subsequent insertion into the electronic archives. This activity, designed to fill gaps and extend the time series, is of great importance for establishing new climatological studies. The digitization of data from weather station models and diagrams instrumental, is an activity of renewed national and international interest, as evidenced by the many programs of cooperation on climate data rescue and overall climate change research conducted by national and intergovernmental agencies.

There are other Italian agencies that deal with the climate record stores; for example,the dataset used in this study has been supplied by the OA-ARRA (Osservatorio delle Acque - Agenzia Regionale dei Rifiuti e delle Acque), former Servizio Idrografico Italiano. The latter formed in 1917 by the then Ministry of Public Works in order to standardize, organize and make available the rainfall measurements, hydrometric and tide gauge in Italy. Before then, these measurements were carried out in an uncoordinated way by the facility that had been carried out this task in the pre-unification states of Italy. The Hydrographic Service has also held until his disposal, the publication of so-called Annals of hydrological, relating to different compartments in which the territory was divided. The division compartment roughly traced the river basins of major Italian rivers. The Annal are divided in:

1. Hydrological Annals: I Part:
    - Section A: Thermometric data;

- Section B: pluviometric data;

2. Hydrological Annals: II Part:

   - Sezione A: Precipitation ;
   - Sezione B: hydrometry daily observation;
   - Sezione C: Streamflow and hydrological balance;
   - Sezione D: groundwater observations;
   - Sezione E: surveys, hydrological studies and special events.

1.2 Missing data in precipitation, temperature and runoff time series

For planning, management and effective control of water resource systems, a considerable amount of data on hydrological variables such as rainfall, streamflow, etc. are required. Very often hydrological data sequences at a given network have gaps or are incomplete, or are not of good quality or are not of sufficient length. This can severely affect the reliability of the design of, i.e. a hydropower plant, the construction of dams, etc. Furthermore, the problem of missing values is a common obstacle in time series analysis and specifically in the context of rainfall, temperature and rainfall–runoff processes modelling where it is essential to have serially complete data. There may be various reasons for missing values, for instance equipment failure, errors in measurements or faults in data acquisition, and natural hazards such as landslides, or even temporary absence of observers, the cessation of measurement or absence of observations prior to the commencement of measurement (Makhuvha et al., 1997) or by limited financial resources (Balek, 1992). Whatever the reasons, missing values produce a significant problem for water resources applications, which generally require a continuous database (e.g. Zoppou et al. 2000; Junninen et al. 2004; Ramirez et al. 2005).Consequently, finding efficient and principled methods to deal with the problem of missing values is an important issue in most hydrological analyses.However, hydrological modellers commonly discard the observations with missing values and only use the observations with complete information, which means that a lot of information contained in the dataset is lost. Furthermore, the method is inadequate for analyses that require serially complete data. On the other hand, the use of the dataset prone to missing data can result in errors that exhibit temporal and

spatial patterns (Stooksburry et al., 1999).As an alternative to this listwise deletion procedure, modellers sometimes replace ("or fill in") a value for the missing values by using, for example, the mean of the observed variables. Such a procedure will, however, seriously distort statistical properties like standard variation, correlations or percentiles. But the best alternative to the above mentioned approaches consists of filling the gaps in the rainfall, temperature or streamflow time series by estimating the missing values. In fact, the most common approach followed in technical literature is the application of either deterministic or stocastic methods to estimate the missing data. In this way the filling gaps in the considered time series is possible.

Moreover, for the streamflow time series, it is possible reconstruct serially incomplete data records in basins with short streamflow records or in ungauged river basins by some interpolation methods. This is a very vital issue in hydrology research, as pointed out by trans-national initiatives of international research groups such as the "Decade on Prediction in Ungauged Basins (PUB)", a wide research project promoted by the International Association of Hydrological Sciences (Sivapalan et al., 2003). So the problem of missing data in streamflow dataset and then the reconstruction of serially incomplete data records, in order to obtain complete and reliable time series, are within the scope of the PUB.

Before presenting a wide and comprehensive overview of the methods used in literature to face this problem, it is appropriate to describe the variables under consideration and to highlight differences on their characteristics.

1.2.1 Characteristics of the considered variables

The variables such as precipitation and temperature show a strong dependence on space and time. This variables are continuous in space and evolve in a continuous time scale. For instance, a recording rain gauging station provides a continuous record of rainfall and rainfall depth $y(t)$through time. A plot of the rainfall depth $y(t)$versus time $t$constitutes a rainfall time series in continuous time. However, most hydrologic processes of practical interest are defined in a discrete time scale. A discrete time series may be derived by sampling the continuous process $y(t)$at discrete points in time, or by integrating the continuous time series over successive time intervals. Moreover, also in terms of the spatial these variables (precipitation and temperature) can be conceptualized and analyzed as a point process.

In particular, the temperature of the atmosphere represents the average kinetic energy of the molecular motion in a small region, defined in terms of a standard or

calibrated thermometer in thermal equilibrium with the air. The temperature of a given small mass of air varies with time because of heat added or subtracted from it, and also because of work done during changes of volume. For its own nature, assessments of the temporal variability of air temperature provide fundamental information on how the climate system responds to a variety of forcings. In evaluating the temporal variability of air temperature, it also is useful to compare the magnitude of temporal changes to unresolved spatial variability. One reason for comparing spatial and temporal variability is fundamental to evaluate climatic change. Observed temporal variability in air temperature at a particular location, for instance, might be the result of unresolved (i.e. aliased) changes in spatial patterns that are not the result of a spatially uniform climatic change, but of local-scale climatic variability. Another reason for comparing spatial and temporal variability is more practical and is related to the compilation of climatological means or *normals*(WMO, 1983) of air temperature. Within such climatologies, station data usually are analyzed only within a standard base period (e.g. 1961 to 1990) and station records that do not contain sufficient data during the base period are removed (e.g. Hulme et al., 1995). If between-station variability is of interest (e.g. a map or gridded field of climatological mean air temperature is needed), then removing such stations assumes that spatial interpolation between stations is more reliable than using a temporal mean from a shorter or different averaging period (Robeson and Janis, 1998). Temperature typically is less spatially variable than precipitation but it is important, for estimation of this variable on sites where there is no measurement, to take into account the spatial dependence structure of temperature.

Rainfall, or precipitation in a more general sense, is generated by rising and cooling air masses in the atmosphere. Precipitation affects surface runoff, infiltration, groundwater, seepage, percolation, and evaporation etc. The effects change in time and space. This temporal and spatial variability is due to temporal (seasonal) atmospheric phenomena and spatial geographical (topographical) factors, respectively. Precipitation records in a geographical region show a complex structure. Basic statistical information such as average, standard deviation, skewness, correlation structure, median value, and range can be computed in a simple way from the precipitation record. Additionally, structural characteristics of the time series can be derived. These characteristics include consistency, independence, randomness, and homogeneity as well as the presence of trends and jumps. A bestfit probability distribution function can also be applied to the time series. Precipitation data analysis is usually based on a network of point meteorological stations. A well-distributed network of

stations is required to extrapolate point-scale results to the area-scale. The spatial distribution may not necessarily be homogeneous across a study area, and this heterogeneous scatter might result in significant differences to results obtained through the analysis of individual stations. Structural characteristics of hydrometeorological variables (precipitation in this study) are important in modeling studies. Another obvious property of rainfall time series is that they are intermittent, displaying long periods of inactivity punctuated by relatively short periods of activity (i.e. storms). This property is one of the primary sources of uncertainty in hydrologic forecasting and remains difficult to model at many scales. It is assumed in the literature, albeit often implicitly, that the source of this intermittent behavior is fundamentally dynamical, meaning that the governing equations of atmospheric phenomena generate intermittent precipitation intrinsically. However, this assumption has not been employed actively in characterizing rainfall (Rigby and Porporato, 2010). As previously mentioned, precipitation is strictly related to local runoff generation. So, in this context, since for estimating missing data into runoff series is very important to understanding the mechanism of rainfall-runoff transformation, runoff generation can be conceptualized as a point process, i.e. runoff generation is assumed to exist at any point in the landscape.

In reality, considering runoff as a punctual variables is a simplified assumption, since it is heavily influenced by the hierarchical structure of the basin drainage system. Also the runoff has a spatial and temporal structure. But as before mentioned, with regard to the spatial dependence, specific comments have to be made on the characteristics of this variable which differ substantially from those of precipitation and temperature. Runoff characteristics are, by definition, derived as integrated values over a basin, i.e. representing a generalized random space-time process with local support equal to the basin area. In order to highlight the complexity in the study of runoff, it can consider the following classification of some of the variables that influence the runoff generation on the basin: there are two main groups of variables that control streamflow (Fig. 1.1). The first group consists of variables that are continuous in space. These variables, including rainfall, evapotranspiration and soil characteristics, are related to local runoff generation. In this context, runoff generation is conceptualized as a point process, i.e. runoff generation is assumed to exist at any point in the landscape. This concept is discussed in Woods and Sivapalan (1999). In a similar way, other streamflow-related variables can be conceptualized locally as a point process. The second group of variables is related to routing in the stream network. These variables are affected by the catchment organization of nested

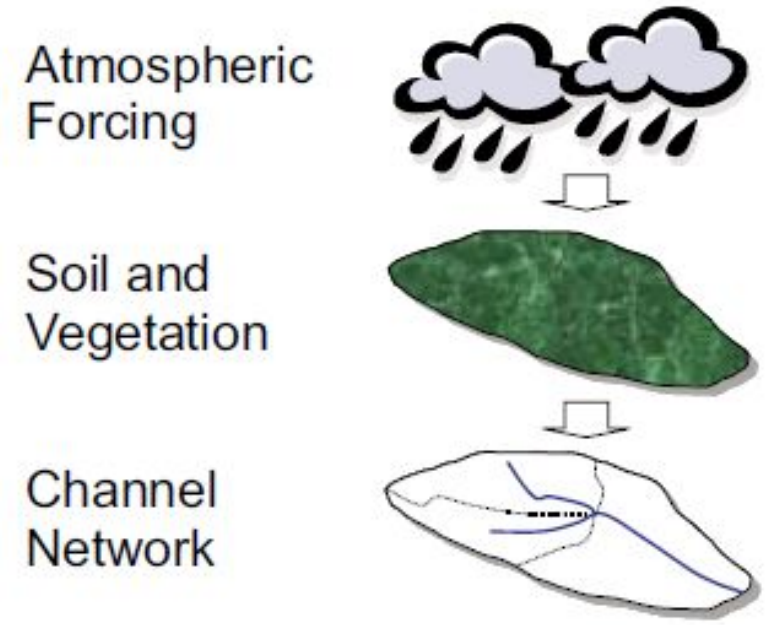

Figure 1.1: Atmospheric forcing and soil and vegetation contribute to the runoff generation process locally and can be represented by point process. The channel network organizes runoff into streams, which can be represented by the catchment boundaries.

catchments where runoff accumulates along the stream network. Variables of this type include mean annual discharge, flood characteristics, low flow characteristics, concentrations, turbidity and stream temperature. These variables are only defined for points on the stream network while has to be considered an areal variables taking into account the tree structure of the entire stream network.

It is necessary, moreover, to highlight the difference on the methods of measurament of these variables. A variety of measurement devices for precipitation have been developed. Rain gauges are of two types: recording and nonrecording. The more widespread rain gauges, now, are the former. In particular recording rain gauges are those used from Servizio Idrografico Italiano (SII), from which rainfall data used in

study are recorded. A recording rain gauge automatically records rainfall accumulation at temporal resolution down to 1 minute o less. Three major types of recording rain gauges are *weighing*type, the *float and syphon*type and the *tipping-bucket*type. For example, a *tipping-bucket*rain gauge consists of a funnel that collects and channels the precipitation into a small seesaw-like container. After an amount of precipitation equal to 0.2 mm (0.007 in) falls, the lever tips, dumping the collected water and sending an electrical signal. The recorder consists of a pen mounted on an arm attached to a geared wheel that moves once with each signal sent from the collector. When the wheel turns the pen arm moves either up or down leaving a trace on the graph and at the same time making a loud click. Each jump of the arm is sometimes referred to as a 'click' in reference to the noise. The chart is measured in 10 minute periods (vertical lines) and 0.4 mm (0.015 in) (horizontal lines) and rotates once every 24 hours and is powered by a clockwork motor that must be manually wound. The temperature is measured by temperature gauge. The rain gauge is usually equipped by the thermometers. Lately rain and themperature gauge amounts are read by by AWS (*Automatic Weather Station*). An automatic weather station (AWS) is an automated version of the traditional weather station, either to save human labour or to enable measurements from remote areas. An AWS will typically consist of a weather-proof enclosure containing the data logger, rechargeable battery, telemetry (optional) and the meteorological sensors with an attached Solar panel or wind turbine and mounted upon a mast. The specific configuration may vary due to the purpose of the system. The system may report in near real time via the *Argos System*(satellite-based system which collects, processes and disseminates environmental data from fixed and mobile platforms worldwide) and the *Global Telecommunications System*(a global network for the transmission of meteorological data), or save the data for later recovery. In the past automatic weather stations were often placed where electricity and communication lines were available. Nowadays, the solar panel, wind turbine and mobile phone technology have made it possible to have wireless AWSs that are not connected to the electrical grid or telecommunications network.

Virtually all raingauge suffer from errors due to modification of the wind field. The magnitude of errors depends heavily on wind speed, siting characteristics, type of precipitation (rain or snow) and temperature. Rain gauges measuraments is difficult in a variety of setting, including mountain ridges, forests and water bodies.

Instead, the procedure to trace the streamflow (flow rate of water in cubic meters per second (m^3/s) along a defined natural channel) in natural channel is more complicated measurement. This is a variable that changes over time and its mea-

surement may not be easy because it is impossible measure directly the volume that crosses a section of a stream per unit time. It is therefore necessary measure streamflow variables-related. Streamflow or discharge measurement normally involves (1) obtaining a continuous record of water levels, or stage above a datum; (2) establishing the relationship between water level and discharge (the stage-discharge relation); and (3) transforming the record of stage into a record of discharge. A streamflow measurement station is commonly called gauging station. In general, measurements of streamflow are less precise than measurements of precipitation both because it is not a direct measurement and because the river always carries the solid material which, once deposited on the stream bed, also induces significant changes in the measurement section. Another reason that makes the measurements not reliable is due to flood events. During this event can happen that the free surface of the stream exceeds the height crowning of the banks and the measured flow rate does not match the one that actually river delivers. Sometimes just because of flood event, the measuring instrument is overwhelmed and the measured data is lost. Even the measurement of streamflow are published on the SII. In Sicily there are a few tens of hydrometric stations and less than ten working more than half a century. The streamflow values are measured in m^3/s, they represent an instantaneous value or an average value on whole period of time. Sometimes, in particular in this study, need to know the volume flowing in a water stream in large time intervals such as months or years. In these cases, the variable taking into account must be the runoff. It is defined as the volume of water that in a certain period of time has passed in the measuring section of the basin, divided by the area S of the same basin.

From the description of the sampling methods of the three variables taken into consideration, one can conclude that time series with reliable and continuous records are not easily obtainable.

1.3 Methods to infilling missing data in precipitation, temperature and runoff time series

As before said, finding efficient and principled methods to deal with the problem of missing values in a climatic time series is an important issue in most hydrological analyses. It should be noted that, according to the previously described variables characteristics, two different classifications of methods are possible. In particular a first classification is done for variables such as precipitation and temperature (point

processes), whereas second classification is done with reference to variables such as runoff (either point or areal processes).

Generally the methods used for obtain the estimated values for the variables such as precipitation and temperature in literature, can be grouped into:

1. *weighting methods*:
   (a) *deterministic interpolation methods* (inverse-distance weighting, non-linear interpolation as spline techniques, etc..)
   (b) *stochastic interpolation methods* (kriging, residual kriging, etc..)
2. *data-driven methods* (regression, artificial neural networks, time series analysis, etc..)

Concerning to the runoff variable, the following the methods used in letterature for obtain the estimated values, can be grouped into:

1. *rainfall-runoff models*(Masky, 2004):
   (a) *physically-based;*
   (b) *conceptual and empirical;*
   (c) *data-driven or black box.*
2. *methods for the construction of runoff maps:*
   (a) *Subjective methods;*
   (b) *Stochastic methods:*
      i. *Classic stochastic approach*;
      ii. *Geostatistical approach* (runoff as areal process).

These models can can be distinguished by their representation of climatic variables in space and time into the following three general classes models:

1. *Spatial models*, which represent the spatial distribution of variables over a specific duration.
2. *Temporal models*, which represent the variables at a point over time.
3. *Space-time models*, which represent both the spatial and temporal evolution of the variables.

In the following sections (1.4 and 1.5), a comprehensive overview of the papers where these methods are applied has been carried out. Concerningthe methods used in literature to obtain the estimated values of runoff variable, the papers in which physically-basedand empiricalrainfall-runoff models are used, have been neglected. These methodologies are, in fact, quite different than the ones that it decided to take into consideration for this thesis.

The methods, among these above mentioned, that will be used in the case of study of this thesis, will be described in detail in Chapter 2.

1.4 State of the art concerning the precipitation and temperature time series filling

As said before, the reconstruction of serially incomplete data records has been the subject of a large number of scientific works where numerous techniques for estimating missing data values have been implemented and compared.

Some types of methods, widely used in letterature for estimating missing data in dataset that suffer from gaps, are that of spatial methods. A detailed study of the literature is here done.

Among the works aimed to create a complete annual and monthly precipitation data set, Tang et al. (1996) used various methods, as the Bruce and Clark (1969) method, the autoregressive order-1 model, the arithmetic mean, the normal ratio, the modified normal ratio, the inverse distance, the quadrant, the isohyetal, the rank matching. These methods were compared to select the best one to make up the missing records of rainfall data in Malaysia; the authors, on the basis of analysis of the errors between the observed and estimated missing values, found that the modified normal ratio method performs satisfactory for the monthly, annual and annual maximum rainfall data. In the Eischeid et al. (2000) work, six different methods of spatial interpolation have been used to create the serially complete daily temperature and precipitation dataset for the United States. Among these methods there were optimal interpolation (OI), multiple regression using the least absolute deviation criterion (MLAD), the single best estimator, and median (MED) of the all the six methods (Eischeid et al., 1995). Since the reliability of the estimation procedure is strongly dependent on the appropriate selection of the validation set, the authors have selected appropriate stations as "target stations" for estimation and have applied the spatial interpolation methods using neighboring stations which had sufficient record (greater

than 10 yr) to provide stable estimation statistics with the target stations.

Many papers have been dedicated to the comparison between deterministic and stochastic approaches to reconstruct daily records using spatial interpolation algorithms to estimate missing data. For example, in Jeffrey et al. (2001), a comprehensive archive of Australian rainfall and climate data (maximum and minimum temperatures, evaporation, solar radiation and vapor pressure) has been derived using a thin plate smoothing spline to interpolate daily climate variables, and ordinary kriging to interpolate daily and monthly rainfall.

Among all different spatial interpolation methods, the inverse distance weighting (IDW) method is, probably, the most commonly used for estimation of missing data in hydrology and geographical sciences. But several variants of IDW are derived and adopted by researchers with a main focus on the weighting schemes. In fact the success of IDW method depends primarily on the existence of positive spatial autocorrelation (Griffith, 1987; Vasiliev, 1996), because data from locations near one another in space are more likely to be similar than data from location remote from one another (Tobler, 1970). Unfortunately this condition is not always true and then insert an arbitrariness in the choice of weighting parameters.

Another significant issue is the arbitrary selection of neighborhood points of observations for the estimation of missing data at a point of interest. Beginning from this limitations, Teegavarapu and Chandramouli (2005) introduced several conceptual improvements to the traditional inverse distance weighting method. They also used a data-driven approach to estimate missing precipitation data and suggested different types of algorithms including modified inverse distance weighting method (MIDWM), coefficient of correlation weighting method (CCWM), inverse exponential weighting method (IEWM), nearest neighbor distance weighting method (NNWM) along with revised nearest neighbor weighting method (RNNWM) and artificial neural network (ANN) estimation. The results obtained by the same authors, suggested that the conceptual revisions can improve estimation of missing precipitation records by improving the procedure to estimate the weighting parameters and surrogating measures for distances used in the IDW method.

Coulibaly et al. (2007) performed a comparison of six different types of ANN approaches for infilling of missing daily total precipitation and daily extreme temperature series in study. Daily precipitation from 15 weather validation stations, are used to evaluate the accuracy of the different models for infilling data gaps. The results highlighted the Multi Layer Perceptron (MLP) as the most effective for infilling missing daily precipitation values.

An interesting application is that of Claps et al. (2008) where the annual and monthly average temperatures are analyzed with statistical methods to characterize the temperature regime in Italy. Data from 738 weather stations, with homogeneous spatial cover throughout Italy, are used to estimate the annual and monthly air temperature normals. Geographic and morphologic parameters, computed around the points of measure, are considered as explicative variables within a georegression model. On the basis of a stepwise regression analysis, the variables pointed out as significant were elevation, latitude, distance from the sea, and a measure of terrain concavity. The relationship between the average annual air temperature and the mentioned variables explains 92% of the variance and produces a standard error of 0.89°C. The temperature regime (normalized mean temperature for each of the 12 months) is reproduced with a two-harmonic Fourier series, with parameters estimated using stepwise regression. Analyses of the reconstruction errors demonstrate that the results are quite satisfactory for many technical purposes, particularly for large-scale climatic characterization.

Other interesting papers present in literature have focused the attention on the interpolation problem addressing through a variety of a spatial methods especially geostatistical. Demyanov et al. (1998) proposed a two step spatial interpolation method named direct neural network residual kriging (DNNRK): the first step is a data-driven approach which includes estimating large scale spatial structure by using an artificial neural network while the second step is the analysis of residuals when a geostatistical method, like ordinary kriging, is applied to model the local spatial correlation. Final estimates are produced as a sum of ANN estimates and ordinary kriging estimates of residuals. Another interesting approach, used to derive the precipitation spatial distribution especially in mountainous regions, is the use of algorithms that incorporate elevation into the spatial prediction of rainfall. Martinez-Cob (1996) used three different geostatistical methods (ordinary kriging, co-kriging with elevation and modified residual kriging) to interpolate precipitation and reference evapotranspiration at annual scale. The author found that co-kriging was superior for precipitation interpolation reducing estimation uncertainty by 18.7% and 24.3% compared with ordinary kriging and modified residual kriging, respectively. Goovaerts (2000) applied spatial interpolation methods to annual and monthly rainfall observations measured at available raingauges using two different groups of algorithms: three multivariate geostatistical algorithms that incorporate a digital elevation model into the spatial prediction of rainfall (simple kriging with varying local means, kriging with an external drift, colocated cokriging) and three univariate techniques (the Thiessen polygon,

inverse square distance, ordinary kriging) which do not take into account the elevation. The comparison among these methods pointed out that the three multivariate geostatistical algorithms gave the lowest errors in rainfall prediction. Lin and Chen (2004) proposed a spatial interpolation method, based on the the classical radial basis function network (RBFN), which incorporates a semivariogram model. From the structure of the standard RBFN the authors took into account the activation function form, in particular the gaussian function, and they looked for similarities between the purely deterministic meaning of the terms it contains and the stochastic mean that one should consider when dealing with problems of spatial interpolation.

The Hierarchical Bayesan models (Banerjee et al., 2004), included among the most promising stochastic spatial interpolation methods and widely used for the estimation and modelling of climatic spatial data, can estimate precipitation at ungauged site taking into account the dependence of the elevation or other variables.

The use of elevation can improve, according to Diodato and Ceccarelli (2005), the spatial interpolation of mean annual precipitation. The authors compared three interpolation methods (IDW, linear regression and co-kriging) to the rainfall recorded in a region of 1,400 km^2 in Southern Italy with elevation ranging from 400 m to 1100 m, concluding that the best method is co-kriging since it is able to take into account several properties of the landscape.

As mentioned above, one of the areas of greatest interest to the scientific community of hydrologists has been the estimation of climatic variables (especially precipitation) through the use of space-time models, i.e. models handling dependence the spatial and temporal simultaneously. Spatio-temporal estimation of precipitation over a region is essential to the modeling of hydrologic processes for water resources management. The changes of magnitude and space–time heterogeneity of rainfall observations make space–time estimation of precipitation a challenging task. An interesting work about spatio-temporal interpolation of climatic variables over large region of complex terrain has been made by Antonic et al. (2001). The aim of this study was to introduce a spatial interpolation of climatic variables obtained at the numerous meteorological stations over the large region of complex terrain, simultaneously with temporal interpolation during the long period covered only by several stations. Empirical models for seven climatic variables (monthly mean air temperature, monthly mean daily minimum and maximum air temperature, monthly mean relative humidity, monthly precipitation, monthly mean global solar irradiation and monthly potential evapotranspiration) were built using neural networks. Independent estimators were elevation, latitude, longitude, month and time series of respective cli-

matic variable observed at two weather stations (coastal and inland), which have long time-series of climatic variables (from mid last century). Differences in residual error around model were insignificant between months, but significant between weather stations, both for all climatic variables. This was the reason for calculation of mean residual error for all stations, which were spatially interpolated by kriging and used as a model correction. Goodness of fit after the averaging of monthly values between years was very high for all climatic variables, which enables construction of spatial distributions of average climate (climatic atlas) for a given period. Presented interpolation models provide reliable, both spatial and temporal estimations of climatic variables, especially useful for dendroecological analysis.

An important contribution to the study of the interpolation methods of climatic variables is given from the analysis of nonseparable spatio-temporal models. Gneiting (2002) proposed a general class of nonseparable, stationary covariance functions for spatio-temporal random processes directly in the space-time domain (i.e., it is a construction not based on the inversion of a Fourier transformation). By studying the parameter values characterizing this class of covariance functions it could be possible to detect if there is separability or not. A common topic of these papers the space-time processes consists of assuming stationary in time and space. The author used a covariance model with a readily interpretable space–time interaction parameter to analyze wind data from Ireland. An other interesting work that used the spatio-temporal modelling by adopting Bayesan inference has been carried out by Gelfand et al. (2004). The authors viewed climatic data as a time series of spatial processes and worked in the setting of dynamic models, achieving a class of dynamic models for such data (precipitation and temperature coming from monitoring stations in Colorado).Gneiting et al.(2006) review recent advances in the literature of space-time covariance functions by focusing on assumptions like separability, fully symmetry, stationarity, etc.

Sang and Gelfand (2009) propose a hierarchical modeling approach for explaining a collection of point referenced extreme values. In particular, annual maxima over space and time are assumed to follow generalized extreme value (GEV) distributions, with parameters μ, σ, and ξ specified in the latent stage to reflect underlying spatio-temporal structure. In this study the authors have relaxed the conditionally independence assumption in the first stage of the hierarchical model. This assumption implies that realizations of the the surface of spatial maxima will be everywhere discontinuous. For many phenomena including, e.g., temperature and precipitation, this behavior is inappropriate. Instead, they have offered a spatial process model for

extreme values that provides mean square continuous realizations, where the behavior of the surface is driven by the spatial dependence which is unexplained under the latent spatio-temporal specification for the GEV parameters. In this sense, the first stage smoothing is viewed as fine scale or short range smoothing while the larger scale smoothing will be captured in the second stage of the modeling. In addition, as would be desired, we are able to implement spatial interpolation for extreme values based on this model.

Hussain et al. (2010) have instead proposed a Box–Cox transformed hierarchical Bayesian multivariate spatio-temporal interpolation method for the skewed response variable. The proposed method is applied to estimate space–time monthly precipitation in the monsoon periods during 1974–2000, and 27-year monthly average precipitation data are obtained from 51 stations in Pakistan. The results of transformed hierarchical Bayesian multivariate spatio-temporal interpolation are compared to those of non-transformed hierarchical Bayesian interpolation by using crossvalidation. It is observed that the transformed hierarchical Bayesian method provides more accuracy than the non-transformed hierarchical Bayesian method.

1.5 State of the art concerning the runoff time series filling

The reconstruction of serially incomplete data records has been the subject of a large number of scientific works where numerous techniques for estimating missing data values have been implemented and compared. Among the works aimed to create a complete monthly streamflow data set, Raman et al. (1995) proposed models to extend the monthly streamflow data at a site where the available historic rainfall and streamflow data are too short for adequate systems study subject to the condition that there are no gauging sites in the basin or adjacent basins with a longer period of streamflow data. In fact, the water resource system chosen for this study was the Kudhiraiyar Basin, a sub-basin of the Amaravathy River located in Tamil Nadu State, India, where a raingauge and a flow measuring device were installed in 1984. Gauged daily flow and rainfall data are available from 1984-1991, a period of eight years. Since the available inflow data was of shorter duration and there were no adjacent basins with longer flow data, extension of the flow record was envisaged. Monthly rainfall data for a nearby raingauge station, Palani, 15 km from the reservoir site, were used for this purpose. The monthly rainfall data were available for a period of 35 years

from 1957-1991. The value of the correlation coefficient between monthly rainfalls at Palani and Kudhiraiyar for the period 1984-1991 was found to be 0.837. Five regression models, namely, runoff coefficient model, single linear regression, monthly linear regression, monthly linear regression with stochastic description for residuals, and a double regressed model are used. These five models were applied to the data pertaining to the Kudhiraiyar basin. Seven years were used for model application and calibration and the remaining one year of data used for validation. The results show that the monthly linear regression model with stochastic description for the residuals is the method that give the best estimates. Therefore the inflows into the Kudhiraiyar reservoir for 27 years from 1957 to 1983 were computed using the monthly linear model with stochastic description for the residual errors.

Many hydrological researchers have adopted and developed various models and techniques to deal with the problem of estimating missing data. The efforts not only are devoted to extending short records by adding lengthy segments of estimated data, but also attention is given to the gaps of short duration. One can find in the literature cases in which sophisticated techniques are used for estimation of a single missing observation (e.g. Griffith et ah, 1985). In water resources, the commonly used techniques for estimation of missing data are based on regression analysis, time series analysis, artificial neural networks and interpolation techniques. A major commonality exists in most of the applications of these techniques; that is, any hydrological time series record is perceived as a sequence of single-valued observations irrespective of the time scale of the data. In hydrological data (e.g. streamflows), it can be noted that annual or seasonal data might be independent, while monthly or weekly data of the same river have significant levels of autocorrelations. Different techniques are employed for modelling annual and monthly data; however, in both cases (year 1, year 2, ... or month 1, month 2, ...) observations are treated, in the literature, as single-valued entities that have interrelations modelled at one stage. This does not happen in the work of Khalil et al., (2000), when it is argued that hydrological data can have a hierarchical type of structure that needs to be modelled in more than one stage. For example, when different months and different seasons are recognized in the same time series, months within each season can be modelled in the first stage and inter-season relationships can be modelled in a second stage. This approach is identified here as "group approach". Based on concepts and properties of groups and artificial neural networks (ANN), this paper develops a segment estimation model for infilling of missing hydrologic records. Efficacy of the proposed model is demonstrated through applications to a number of natural watersheds. ANN, coupled with

seasonal grouping, appear to be one such alternative to the linear statistical methods in identifying and studying the intricate nature of such time series. The ANN-based models are applied on several rivers to evaluate their data infilling efficacy in terms of statistical and graphical assessment. Based on the consideration of the group-valued data approach and the relevant structural composition of ANN (in particular *Multi Layer Perceptron*with *Backpropagation*) two types of models are evaluated in this paper namely the *Multi-layer-feed-forward Autovariate lag-one Series Model*(MASM) and the *Multi-layer-feed-forward Bivariate Series Model*(MBSM). The MASM and MBSM models are, respectively, for the cases involving only one data series with data gaps (autovariate), and for the data series with data gaps in which vicinity one or more concurrent but complete data series are available (bivariate). The group-based neural network models are compared with multi-dimensional regression (MR) and the pattern recognition based methods (PR) models, by some statistical indexes. This comparison shown that the ANN-based methods returned the best performance and retained relevant properties of the historical streamflows both at the auto and cross-variate series levels in the estimates. Further, the group-based neural network models are found to closely infill the missing peak flows and also the moderate flows. The results suggest that infilling of data gaps of streamflows based on the concept of neural networks and group-valued data approach is a good approach .

Another study that uses feed-forward artificial neural networks (ANNs) techniques for streamflow data infilling is that of Ilunga and Stephenson (2005). The standard back-propagation (BP) technique with a sigmoid activation function is used. Besides this technique, the BP technique with an approximation of the sigmoid function by pseudo Mac Laurin power series Order 1 and Order 2 derivatives, as introduced in this paper, is also used. Empirical comparisons of the predictive accuracy, in terms of root mean square error of predictions (RMSEp), are then made. A preliminary case study in South Africa (i.e. using the Diepkloof (control) gauge on the Wonderboomspruit River and the Molteno (target) gauge on Stormbergspruit River in the River summer rainfall catchment) was then done. The available data in this gauges was mean montly and mean annual runoff and two seasons of a 6-month period each were assumed. Generally, it is demonstrated that the standard BP technique performed just slightly better than the pseudo BP Mac Laurin Orders 1 and 2 techniques when mean values of seasonal data are used. However, the pseudo Mac Laurin approximation power series of the sigmoid function did not show any substantial impact on the accuracy of the estimated missing values at the Molteno gauge. Thus, all three the standard BP and pseudo BP Mac Laurin orders 1 and 2 techniques could be used to fill in the

missing values at the Molteno gauge. It was also observed that a linear regression could describe a strong relationship between the gap size (0 to 30%) and the expected RMSEp (thus accuracy) for the three techniques used here.

An intersting procedure to estimate monthly streamflow series in ungauged basins and filling the gap in short and intermittent series is the regionalization procedure for estimating the parameters of simple rainfall–runoff models. Many studies can be mentioned. Although the importance of the use of complex hydrological models for water resources planning and management is widely recognized, experience has often shown that simple models can be usefully adopted for the needs of the water agencies in the assessment of the available water resources in a region. Assessment of surface water resources basically requires the knowledge of streamflow data at specific time scales (daily, monthly, yearly), as suggested by Alley (1984) and Xu and Singh (1998), and space scales (river basin, regional, national and international). These data are often scarce both in time and space, thus rainfall–runoff models are usually adopted in order to estimate streamflows as a function of available hydro-meteorological information. Regionalization procedures, able to express the model parameters on the basis of physical characteristics of the basin, can be effectively adopted to estimate streamflows at ungauged sites or to extend short series. The simplest regionalization procedures are based on the direct transfer of model parameters to ungauged basins from nearby hydrologically similar basins. These approaches have included proxy-basin method (Klemes 1986; Xu 1999), linear interpolation methods (Guo et al. 2001), Kriging interpolation methods (Vandewiele and Elias 1995). Other regionalization procedures are based on a "two-step" approach. In the first step, different models are calibrated separately for each gauged basin of the region; in the second step, the parameters of each model are expressed, usually by means of multiple regressions, as a function of the geomorphological characteristics of the examined basins (Weeks and Ashkanasy 1985; Braun and Renner 1992; Franchini et al. 1996; Adbulla and Lettenmaier 1997; Tung et al. 1997). Multiple linear regressions are also used for the estimation of regional relationships of conceptual model parameters. In a comparative study, Peel et al. (2000) found that some parameters of the SYMHID model were significantly correlated to the basin attributes. Seibert (1999) related the model parameters of the HBV model to the basin attributes. Beldring et al. (2003) used various basins in Norway for calibrating a version of the HBV model. They regionalized the model parameters as a function of land use classes obtaining high values for the performance indexes. Kokkonen et al. (2003) used the IHACRES model with six parameters and found regressive regional relationships which did not guarantee a

good predictive capability. Recently Fernandez et al. (2000) have adopted a different regionalization procedure based on a "one-step" approach. This approach is based on the development of a single regional model calibrated using hydrological, climatical and geomorphological data derived from all the gauged basins of the region. The result of the calibration, is a model that can be applied directly to ungauged basins within the region. A similar approach was applied in Szolgay et al. (2003), to find regionally valid parameters of a monthly water balance model. They jointly calibrated a model using multiobjective calibration, where the basins were pooled together using cluster analysis of selected physiographic basin attributes. The paper of Cutore et al. (2006) aims to analyze applicability and limitations of two regionalization procedures based on a "two-step" and on a "one-step" approach, respectively, for the estimation of monthly streamflow series in ungauged basins. In particular the two-step approach requires a first step consisting in the preliminary definition of a simple regression-based rainfall–streamflow models for all the gauged basins of the region (Cutore et al. 2005). The second step of the regionalization procedure consists in the determination of n regional regression equations between the parameters and the geomorphological characteristics of the gauged basins (average altitude, soil permeability, stream length, etc). In the "one-step" approach a regional rainfall–runoff model is calibrated making use of all the information at the $k$gauged basins, considering also geomorphological characteristics as independent variables, in addition to hydrological and climatological data, using a neural network as a MLP. The comparison of the two regionalization procedures seems to indicate (at least for the investigated sub-basins) that the neural network "one-step" approach should to be preferred. The improvement shown by such models can also depend on the difficulties of defining an appropriate prior non-linear structure for regression model to approximate better the nonlinearity of the investigated physical process.

Still within regionalization procedures, very interesting is the work of Castiglioni et al. (2009). This study investigates the applicability of physiographical space-based interpolation techniques for the prediction of low-flow indices in ungauged basins (basins for which discharge observations are sparse or unavailable). The study considers 51 catchments located in a wide region of central Italy, for which several geomorphological and climatic descriptors are available (drainage area; main channel length, percentage of permeable area, maximum, mean and minimum elevations, average elevation relative, concentration time, mean annual precipitation and average annual temperature). The analysis applies both deterministic and geostatistical techniques for interpolating low-flow indices (the discharge associated with a duration of 355 days, Q355 (l/s); the

daily discharge equalled or exceeded 95% of the time, Q95% (l/s); the 7- day discharge with a recurrence interval of 10 years, 7Q10 (l/s). in the physiographical space. The size of the geomorphoclimatic space n by performing a Principal Components Analysis (Basilevsky, 1994; Chokmani and Ouarda, 2004). Initial applications of PCA, highlighted the lack of descriptive power of TAM (temperature annual mean), whose empirical values are very similar for all considered basins and virtually independent of any of the three low-flow indices. Therefore, the authors dropped TAM obtaining and retained n = 9 physical descriptors in the remainder of the analysis (i.e., A, L, P, Hmax, Hmean, H0, DH, sc, MAP). A jack-knife cross-validation procedure is applied in order to quantify the accuracy of each technique when it is applied to ungauged basins. The results of the study show that physiographical space-based interpolation is a viable approach for estimating low-flow indices in ungauged basins and geostatistical techniques outperform deterministic techniques.

An interesting area to deal with the problems of missing data filling is that of the fuzzy algorithms applications. More recently, neurofuzzy systems have gained attention. They are a composition of artificial neural networks (ANN) and fuzzy logic (FL) approaches. ANNs reconstruct links between input–output pairs for the system being modeled. ANNs have to be trained in order to generate the desired output (Chen et al. 2006; Tingsanchali and Gautam 2000). Fuzzy logic and fuzzy set theory, founded by Zadeh 1965, are used to identify the characteristics of decision making through a set of logical rules. Qualitative modeling techniques can cope without crisp numbers representing variables. They may use imprecise statements made in natural language (Sugeno and Yasukawa, 1993). Fuzzy logic approaches have been applied in the estimation of water resources for more than 10 years Zhu and Fujita 1994; (Zhu et al., 1994; Şen, 1998; Stuber et al., 2000; See and Openshaw, 2000; Hundecha et al., 2001; Xiong et al., 2001; Keskin et al., 2004, 2006; Terzi et al., 2006). Recently, neurofuzzy systems have been introduced in hydrology, taking the advantage of both FL and ANN, i.e., benefiting from the training ability of the ANN and the fuzzy IF–THEN rule generation and parameter optimization. The adaptive neural-based fuzzy inference system ANFIS model and its principles, proposed by Jang 1992, have been applied to study many problems. The model identifies a set of parameters through a hybrid learning rule combining the backpropagation gradient descent and a least-squares method. It can be used as a basis for constructing a set of fuzzy IF–THEN rules with appropriate membership functions in order to generate the preliminary stipulated input–output pairs. Some researchers have applied ANFIS in hydrological modelling. Chang and Chang (2001) studied the intelligent control of a real-time

reservoir operation model and found that, given sufficient information to construct the fuzzy rules, the ANFIS helps to ensure more efficient reservoir operation than the classical models based on rule curve. The aim of the study of Keskin and Taylan (2009) was to develop an optimum flow prediction method, based on the adaptive neural-based fuzzy inference system ANFIS and artificial neural network ANN; each methodology was applied to river flow predicting in Manavgat Stream in the southern part of Turkey. In application, Manavgat Stream flows were predicted from Dalaman Stream, Alara Stream, and Göksu Stream flows. Each stream is located in different catchments. For monthly streamflow predictions, data were taken from the *General Directorate of Electrical Power Resources Survey and Development Administration.* Used data covered a 35- year period 1969–2003 for monthly streamflows. The ANFIS and ANN models had only one output with three input variables. In particular, in this part of the study, the ANFIS model was trained using the observed monthly mean flow data. First, 336 monthly mean values the period between October 1969 and September 1996 constituted the training data set, whereas the last 84 ones the period between October 1996 and September 2003 were available for the testing stage. The input layer consisted three streamflows at time t in different catchments: Dalamant, Alarat, and Göksut Streams. The output layer contained a single flow value at time *t*for Manavgat Streamt. In the ANFIS structure, each variable may have several values in terms of rules, and each rule includes several parameters of membership functions. Comparison of the ANFIS and ANN models showed an agreement between the ANFIS model estimations and measurements of monthly flows better than ANN. With the help of the ANFIS model for interbasin flow prediction, it was possible to estimate missing data.

Observed data in spatial (e.g., topography and land cover), temporal (e.g., streamflow and groundwater levels), and spatiotemporal domains (e.g., rainfall) impact streamflow.For this reason, even in the field of estimate of the streamflows space-time models are applied in literature. An interesting example is the Amisigo and Giesen (2005) work. In this study a spatio-temporal linear dynamic model has been developed for patching short gaps in daily river runoff series. The model was cast in a state-space form in which the state variable was estimated using the Kalman smoother (RTS smoother). The EM (expectation-maximization) algorithm was used to concurrently estimate both parameters and missing runoff values. Application of the model to daily runoff series in the Volta Basin of West Africa showed that the model was capable of providing good estimates of missing runoff values at a gauging station from the remaining time series at the station and at spatially correlated

stations in the same sub-basin.

Geostatistical interpolation approaches can be used to explore the whole spatial–temporal correlation structure of the runoff field (e.g. Gottschalk et al. 2006; Skøien and Bloschl, 2006) as well. Gottschalk et al. (2006) presented an approach to depict the two first order moments of runoff as a function of area (and thus on a map). The focal point is the mapping of the statistical properties of runoff $q = q(A, D)$ in space (area A) and time (time interval D). The problem is divided into two steps. Firstly the first order moment (the long term mean value) is analyzed and mapped applying an interpolation procedure for river runoff. In a second step a simple random model for the river runoff process is proposed for the instantaneous point runoff normalized with respect to the long term mean. From this model, theoretical expressions for the time-space variance-covariance of the inflow to the river network are developed, which then is used to predict how the second order moment vary along rivers from headwaters to the mouth. The observation data are handled in the frame of a hydrological information system HydroDem, which allows displaying the results either in the form of area dependence of moments along the river branches to the basin outlet or as a map of the variation of the moments across the basin space. The findings are demonstrated on the example of the Moselle drainage basin (French part). In the Skøien and Bloschl (2006) paper catchments are conceptualized as linear space-time filters. Catchment area A is interpreted as the spatial support and the catchment response time T is interpreted as the temporal support of the runoff measurements. These two supports are related by $T \backsim A^k$ which embodies the space-time connections of the rainfall-runoff process from a geostatistical perspective. To test the framework, spatiotemporal variograms are estimated from about 30 years of quarter hourly precipitation and runoff data from about 500 catchments in Austria. In a first step, spatio-temporal variogram models are fitted to the sample variograms for three catchment size classes independently. In a second step, variograms are fitted to all three catchment size classes jointly by estimating the parameters of a point/instantaneous spatiotemporal variogram model and aggregating (regularizing) it to the spatial and temporal scales of the catchments. The exponential, Cressie-Huang and product-sum variogram models give good fits to the sample variograms of runoff with dimensionless errors ranging from 0.02 to 0.03, and the model parameters are plausible. This indicates that the first order effects of the spatio-temporal variability of runoff are indeed captured by conceptualizing catchments as linear space-time filters.

In the Nagarajan et al. (2010) work starting from the considerations that: (1)

physically based hydrologic models have been used to predict streamflow but often with significant uncertainty because numerous assumptions are made for many missing data in the input and parameter values and (2) traditional Bayesian inference approaches suffer from superlinear increases in computational complexity as the number of data sets to be fused grows, a scalable spatio-temporal approach based on Bayesian networks (BNs) is presented for estimating streamflow. An information-theoretic methodology based on conditional entropy is employed to quantify the impact of adding nodes in the BN in terms of information gained. The framework offers the flexibility of embedding knowledge from hydrologic models calibrated for the study area by introducing them as additional nodes in the network, thereby improving prediction accuracy. Posterior probabilities of estimates and the associated entropy provide valuable information on the quality of predictions and also offer directions for future watershed instrumentation.

Another interesting approach to address the problem of Prediction in Ungauged Basins (PUB) (Sivapalan et al., 2003), i.e. to estimate streamflow-related variables at locations where no measurements are available, is that of the production of runoff maps through which it is possible estimate runoff generation in a certain region.

There are three main issues to be considered when choosing methods for the construction of runoff maps (Gottschalk & Krasovskaia, 1998): the method to be used for interpolation, the scale of fundamental units on the map, and the available observations that can be used to resolve the variability at different spatial scales. The method for interpolation can be either manual contouring (subjective methods in the meteorological terminology), or automatic interpolation (objective methods). The automatic interpolation can, in its turn, be divided into deterministic and stochastic approaches. In both cases a formula representing a weighted average is applied. Weighted averages include a wide class of methods from a simple averaging of point observations to stochastic interpolation with local support considering the extension of drainage basins. In this latter case it is assumed that the total area to be mapped is divided into fundamental units by means of subdividing a larger drainage basin into sub-basins or into a regular grid network. The second topic that needs attention is the scale. On a macro scale, drainage basins used for interpolation can become small in comparison with the total area to be mapped. They can therefore be approximated by points in space. This usually also implies that the simplified "vertical perspective" on runoff (or rather rainfall excess) is accepted and runoff is mapped in the same way as, for instance, precipitation. The simplest method is to use an average of the runoff from all the small basins which fall within a grid cell. A fundamental condition is,

of course, that all cells contain observation points. Arnell (1995), for example, has applied this method across the FRIEND region. In order to overcome the problem of having empty cells and to allow a more sophisticated consideration of the difference in geographical location, a deterministic interpolation method can be utilized. Bishop & Church (1992) and later Arnell (1995) have applied the TIN method with this purpose (i.e. linear interpolation within the facets of the Triangulated Irregular Network defined by the gauging station considered as nodes). Conventional stochastic interpolation is also appropriate at the macro scale. Such methods are standard methods for interpolation of stochastic fields in meteorology and climatology (objective methods, Gandin interpolation) (Gandin, 1963; Daley, 1961). They are parallels to kriging, widely applied to interpolation problems in hydrogeology (Matheron, 1965; Delhomme, 1978). Kriging and Gandin interpolation are also of wide use for interpolation and integration of precipitation fields (Lenton & Rodriguez- Iturbe, 1977; Creutin & Obled, 1982; Tabios & Salas, 1985; Dingman et al, 1988; Barancourt et al, 1992). There are also examples of the application of such methods to simplified assumptions for interpolation of runoff as a point process (Villeneuve et al, 1979, Hisdal & Tveito, 1993). If this approach is used properly, only data from small drainage basins can be applied so that a "point" covariance model can be constructed. On meso and micro scales (say, grid cell sizes in the range 10 km x 10 km to 1 km x 1 km and less than 1 km x 1 km, respectively), the area of drainage basins needs to be taken into account in the interpolation procedure, which has several advantages compared to the "point" interpolation. When basins are considered as "points" in a continuous space, the lateral aspects of the runoff process are neglected. Therefore, one cannot expect that runoff in this case, when integrated over a river basin, coincides with measured streamflow in the main rivers. These observations in the main rivers are, as a rule, avoided and not included in the data set. The information that they can add to the variation pattern of runoff is thus lost. A further practical aspect is that small basins are often situated in the headwaters to a river system leading to an overestimation of the total runoff (Arnell, 1995). In a collaborative paper on grid estimation of runoff (IIASA, 1990), the catchmentbased area-weighted average method (the "nested approach") is proposed for the calculation of runoff for fundamental units (grid cells) as weighted averages with a consideration of the drainage basin areas. Two cases are distinguished. In the first case, at least one basin with observations is within the grid cell. The runoff estimate is calculated from all measured runoff values in the grid cell as a weighted average (aggregation). In the second case, one drainage basin with observed runoff covers more than one grid cell. A dis-

aggregation of the observed runoff for the basin has to be made. The runoff for each grid cell fully within the basin is found by interpolation of runoff (disaggregation) utilizing the relationship between the area of the whole basin and that of a grid cell and eventually other factors. Predeek & Isele (1992) calculated runoff for the Aller River (tributary of the Weser River) by applying this method. The drainage basin area was subdivided into grid cells of 0.5° x 0.5°. The method has also been applied to the FRIEND database (Arnell, 1995). It is a general opinion that, in most cases, the catchment-based area-weighted average method gives the closest estimate to observed runoff in comparison with the methods referred to above. Gottschalk (1993a,b) has developed an alternative stochastic approach for the interpolation of runoff. It takes full account of the fact that runoff is to be integrated to streamflow, thus considering the hierarchical structure of the basin drainage system. To achieve this, distance is measured along the river network and the covariogram for points must be replaced by a covariogram for the drainage basins, i.e. a covariogram model for the whole river system needs to be developed. The third topic to be considered is the type of observations at hand to resolve the variability across space at different scales. The estimated spatial variability from a regional set of observations can be expected to depend on the size of the basins involved—the higher the variability, the smaller the basins. This fact has two implications: the first is that the size of fundamental units of a map and the size of basins used for interpolation must be of a comparable scale and the second is related to the number of fundamental units on the map with respect to the number of basins available as a background for the interpolation procedure. If exactly the same set of runoff observations is used for interpolation to different fundamental units, the basic difference is in the estimation error. It will be larger, the smaller the fundamental units are. The eventual larger detail that a map, based on small fundamental units, reveals is counterbalanced by a larger estimation error.

A hierarchical approach for interpolation is elaborated in Sauquet et al. (2000), with a consideration of the specific topics discussed above. The point of departure is the stochastic interpolation procedure developed by Gottschalk (1993a,b). The territory (major drainage basin) to be mapped is divided into sub-basins in a hierarchy of scales. The number of levels in this hierarchy is determined mainly by the amount of available observations, which also indicates the level of detail that can be achieved (size and number of fundamental units of the map). The first level in a larger drainage basin is usually already well defined by existing observation stations in the main rivers constituting the first level of sub-basins. These basins are, in their turn, divided into a second level of sub-basins (or grid cells), and observation stations with appropriate

basin scales are chosen as the background for the interpolation. The interpolation procedure guarantees that the water balance equation is satisfied so that the sum of runoff from this second level of basins is equal to that of the first order basin accommodating them. The procedure can be repeated to a third level and so on. At each step new information must be added. Auxiliary runoff values to supplement or replace observed runoff values can be calculated for points in space in a regular or irregular pattern by means of empirical relationships and water balance models. The results were compared to an established method (the nested approach, (IISA, 1990)), and a cross-validation was performed for each mapping technique. The disaggregation approach appears to give the most consistent results. Finally, two gridded maps were derived by applying the disaggregation twice to assess water depth on an increasingly finer grid mesh (32 x 32 km resolution and 16 x 16 km resolution, with respect to water balance on 32 x 32 km grid cell estimation based on disaggregation procedure). The global constraint of water balance was applied to each element of the coarser mesh to give estimates for the finer one.

Another work that uses a geostatistical approach is Skøien and Bloschl (2005). They propose a method of geostatistical estimation, *Top-kriging*, or topological kriging, as a method for estimating streamflow-related variables in ungauged catchments. The main appeal of the method is that it is a *best linear unbiased estimator*(BLUE) adapted for the case of stream networks without any additional assumptions. This method of geostatistical estimation on stream networks is built on the work of Sauquet et al. (2000). But it extends the original work in a number of ways. First, they suggest that the interpolation method can be used, in an approximate way, for a range of streamflow-related variables including variables that are not fully mass conserving. Sauquet et al. (2000) interpolated mean annual runoff which is a mass conserving variable. Second, they use variograms while Sauquet et al. (2000) used covariances. This allows them to deal with variables that are non-stationary. Third, they account for local uncertainties of the measurements that may differ between locations. This allows to exploit short records. Last, they illustrate the potential of the approach for estimating the uncertainty of the variable of interest in ungauged catchments. This approach is applied to the case of estimating the 100 year specific flood in ungauged catchments in Austria. This includes a comparison of the estimates with Ordinary Kriging as well as an analysis of the estimation uncertainties in ungauged catchments. Moreover, the authors suggest that Top-kriging can be used for spatially interpolating a range of streamflow-related variables including mean annual discharge, flood characteristics, low flow characteristics, concentrations, turbidity and stream temperature.

In this thesis, only the rainfall, temperature and runoff data spatial structural dependence is used to recostruct missing rainfall data, neglecting then the spatial-temporal dependence. In particular, the considered variables has been studied taking into account their characteristics. So for the variables as precipitation and temperature that are point processes, different algorithms used for the spatial interpolation are applied (inverse distance weighting, radial basis function with thin plate spline, simple linear regression, multiple regression, geographically weighted regression, artificial neural network, ordinary kriging, residual ordinary kriging). Whereas for the variable as runoff, that can be assimilated to an areal process, it considered appropriate to take into account a stochastic method to construction of runoff maps with a geostatistical approach (stocastic interpolation system that can be assimilated to kriging system with consideration about the runoff variable as an areal process).

Chapter 2

Methods of spatial interpolation for point climatic variable

The methos used for the estimation of climatic variables for obtaining a robust and reliable dataset when measurements are either unavailable or affected by gaps, are numerous. It has been explained in the state of the art reported in chapter 1. Due to the spatial-temporal structure of these data, methods using time-based, space-based or time and space-based approaches, respectively, exist.From the study of comprehensive set of previously published papers, in this work a spatial approach has been chosen, using spatial interpolation methods and other methods that can be assimilated to them, due to the characteristics of the input variables (geographical coordinates, altitude, topographic characteristics, etc.).Particular attention was paid to the different characteristics of the considered climatic variables and to the choice of the spatial methods application. Climatic variables, such as precipitation and temperature, can be assimilated to point process while climatic variables, such as runoff, depends strongly on the area when they are measured, and, in particular, on the hierarchical structure of the basin river network. Since this latter variables can not be compared to a point process, thus the spatial interpolation methods have to be applied with appropriate variation.

The spatial interpolation methods used in this study can be divided into two classes: *deterministic methods*and *stocastic methods*.

The deterministic methods are models where arbitrary or empirical model parameters are used. From the observed value of a generic variable, once applied a given

transformation, it can be obtained an estimated value with the same value of the observed one. They are mathematical - based methods that can be adapted to the studied phenomenon by means of hypotheses. The term "model" is used to indicate the theoretical framework, which reduces a complex behavior upon a few variables, related to each other according to their functional dependency. No estimate of the model error is available and usually no strict assumptions about the variability of a feature exist. The most known techniques that belong to this group are: thiessen polygons, inverse distance interpolation, splines, etc.

The stocastic models are statistical - based models. In the case of these models, the parameters are commonly estimated in an objective way, following the probability theory. From the observed value of a generic variable , once applied a given transformation, it can be obtained an estimated value that will have a different value from that observed. The predictions are accompanied with the estimate of the prediction error. A drawback is that the input dataset usually need to satisfy strict statistical assumptions. There are at least four groups of statistical models: Bayesian-based models (e.g. Bayesian Maximum Entropy), environmental correlation (e.g. regression-based), kriging (plain geostatistics), mixed models (regression-kriging).

Spatial prediction models can also be grouped based on the:

1. *smoothing effect*, whether the model smooths predictions at sampling locations or not:
   - exact (measured and estimated values coincide);
   - inexact (measured and estimated values do not have to coincide);
2. *proximity effect*, whether the model uses all sampling locations or only locations in local proximity:
   - local (a local sub-sample; local models applicable);
   - global (all samples; the same model for the whole area);
3. *convexity effect*— whether the model makes predictions outside range of the data:
   - convex (all predictions are within the range);
   - non-convex (some predictions might be outside the range);
4. *support size*— whether the model predicts at points or for blocks of land:

- point-based or punctual prediction models;
- area-based or block prediction models;

Here, some of this method are described with the following subdivision:

1. Deterministic Model:
   - *Radial Basis Function* with *thin plate spline*
   - *Inverse Distance Weighting*
2. Stochastic Model:
   - **data-driven models**:
     - *spatial regressive models*
     - *geografically Weighted Regression*
     - *artificial Neural Networks*
   - **geostatistic models**:
     - *kriging*
     - *residual Kriging.*

In the following, the detailed description of the methods previously listed is given. It should be noted that the general formulation of such methods is suitable to interpolate climatic variables that can be assimilated to point process. On the other hand, a modified formulation of one of this method, i.e. the geostatistic method, appropriate for interpolation of a climatic variable, such as runoff, that depends strongly on the area when it is measured, will be presented in the next chapter.

2.1 Deterministic methods - Mechanical spatial prediction models.

Deterministic spatial prediction models can be very flexible and easy to use. They can be considered to be subjective or empirical techniques because the parameters of the model are arbitrarily selected, often without any deeper statistical analysis. Most commonly, a user typically accepts the default parameters suggested by some software, and for this reason they are also called mechanical models. The most used mechanical spatial prediction models are Thiessen polygons, inverse distance interpolation and

splines, although the list could be extended (Lam, 1983; Myers, 1994). In general, mechanical prediction models are more primitive than the statistical models and often sub-optimal, however, there are situations where they can perform as good as the statistical models.

The deterministic spatial interpolation approach typically involves generating a fine rectangular grid covering the study region, and then estimation of the surface value or height for every grid intersection or cell. The estimation process involves the use of a simple linear expression in order to compute grid values:

$$\hat{z}\left(\boldsymbol{x}_0\right) = \sum_{i=1}^{N} \lambda_i z\left(\boldsymbol{x_i}\right) \tag{2.1}$$

where $\hat{z}\left(\boldsymbol{x}_0\right)$ is the z-value to be estimated for location $\boldsymbol{x}_0$, the λ_i are a set of estimated weights and the z_i are the known (measured) values at points (x_i, y_i). As $\hat{z}\left(\boldsymbol{x}_0\right)$ is a simple weighted average an additional constraint is required ensuring that the sum of the weights adds up to 1:

$$\sum_{i=1}^{N} \lambda_i = 1 \tag{2.2}$$

The interpolation problem is to determine the optimum weights to be used. If $\lambda_i = 0$ for all i except for the measured point closest to the grid intersection then this would represent a form of nearest-neighbor interpolation. If all n points in the dataset were used and weighted equally every point would have weights $1/n$ and would be given the same z-value. In many cases, as per Tobler's First Law, measured points closer to $\hat{z}\left(\boldsymbol{x}_0\right)$ are more likely to be similar to $\hat{z}\left(\boldsymbol{x}_0\right)$ than those further afield and hence warrant weighting more strongly than observations that are a long way away. Indeed, there may be little to gain from including points that lie beyond a given radius, or more than $m < n$ points away, or points that lie in certain directions.

2.1.1 Inverse distance weighting

Probably one of the oldest spatial prediction technique is the inverse distance interpolation (Shepard, 1968).

Inverse distance weighting models work on the premise that observations further away should have their contributions diminished according to how far away they are. The simplest model involves dividing each of the observations by the distance it is

from the target point raised to a power r:

$$\hat{z}\left(\boldsymbol{x}_0\right) = \sum_{i=1}^{N} \lambda_i(\boldsymbol{x_o}) z\left(\boldsymbol{x_i}\right) \tag{2.3}$$

where the weights λ_i are expressed as function of distance as follows:

$$\lambda_i = \frac{d_{i0}^{-r}}{\sum_{i=1}^{N} d_{i0}^{-r}} \tag{2.4}$$

The basic idea for IDW method is that observations that are close to each other on the ground tend to be more similar than those further apart (Tobler, 1970), hence observations closer to $\boldsymbol{x_0}$ receive a larger weight.

A faster rate of distance decay may be provided, by including a power function of distance, $r > 1$, rather than simple linear distance. While any r value convenient for a given application may be used, common practice is to use distance ($r = 1$) or distance squared ($r = 2$). Moreover, this exact interpolation method requires, in addition to the choice of the exponent r, the choice ofa search radius R (specifying the search direction and the search shape) or alternatively the minimum number $N$of points required (limiting the number of points included) for the interpolation.

Figure 6.34 and Figure 6.35 provide illustrations of the method applied to the test data for Pentland Hills OS NT04 (Smith et al., 2008). The surface plot shows how simple IDW with no smoothing and power 2 distance decay results in dips and peaks around the data points but is otherwise relatively smooth in appearance. Figure 6.35A shows the source data (spot heights and contours) with Figure 635B and Figure 635C illustrating the surface obtained using parameters of $r = 1$ and $r = 2$. Both exhibit the familiar bull's eye effect of standard IDW. Figure 635D is markedly different and seems much closer to the source contours. In this case we have selected $r = 3$, a smoothing factor of $t = 2$, and an anisotropy (directional bias) of 45° degrees using an elliptical search region with a ratio of 2:1. The selection of these values was made after limited experimentation using simple cross-validation and comparison with additional information.

2.1.2 Natural neighbor

Natural neighbor interpolation creates weights for each of the input points based on their assumed "area of influence". These areas are determined by the generation of Voronoi polygons around each input point. In principle every grid intersection created would be in one of these polygons and could be assigned the value of the point around which the polygon has been created. This would result in a step-like surface of patches. This is the kind of result that is obtained from Nearest Neighbor interpolation. A far more effective approach involves a development of this idea, as described below. The end result is a smooth surface with discontinuities at the input points.

The first step in the process is to create a Delaunay triangulation of the *j=1, 2...62*input data points (Figure 636A) as a preliminary stage in the creation of Voronoi polygons.

The second stage is to generate a set of Voronoi polygons for the study region (Figure 636B). Each of the points, j, in the source dataset has its own Voronoi polygon, which has an area A_j. In order to determine the estimated value at a sample point P the point P is temporarily added to the set (so there are now 63 points) and the Voronoi polygons are re-computed. Adding point Presults in a new Voronoi polygon and redefinition of those immediately surrounding it (Figure 636C). This new polygon has an area A_p.

Effectively this new point has "borrowed" some of the area of influence from each of the nearby points. This can be seen in Figure 636D, where the new region has been overlaid on the original set. There are $k = 5$ original polygons that the new polygon has borrowed area from. Let us call these borrowed areas $A_{ip,i} = 1, \ldots k$ then the total area of P's Voronoi polygon is:

$$A_P = \sum_{i=1}^{N} A_{iP} \tag{2.5}$$

and thus the proportion borrowed from each of the original points is:

$$\lambda_i = \frac{A_{iP}}{\sum_{i=1}^{k} A_{iP}} \tag{2.6}$$

These proportions are the weights used to compute the estimated value at P,

based on the standard linear weighting equation:

If the point Pcoincides with one of the existing points its area of overlap with that point would be 100%, hence its weight would be 1, as required. If the Voronoi polygon for Pdoes not overlap a region the weight associated with that region is 0. One of the main advantages of this method of interpolation is that it requires no decision-making regarding the number of points to use, the radius or direction of search, or any other parameters.

2.1.3 Radial Basis Function with Thin Plate Spline

Radial basis interpolation is the name given to a large family of exact interpolators. In many ways the methods applied are similar to those used in geostatistical interpolation, but without the benefit of prior analysis of variograms. On the other hand they do not make any assumptions regarding the input data points (other than they are not co-linear) and provide excellent interpolators for a wide range of data.

A great deal of research has been conducted into the quality of these interpolators, across many disciplines. For terrain modelling and earth sciences generally the so-called multiquadric function has been found particularly effective, as have thin plate splines. The simplest variant of this method, without smoothing (i.e. as an exact interpolator) can be viewed as a weighted linear function of distance (or inverse distance) from grid point to data point, plus a "bias" factor, μ. The model is of the form:

$$\hat{z}(\boldsymbol{x}_0) = \sum_{i=1}^{N} \lambda_i \phi(||\boldsymbol{x_i} - \boldsymbol{x_0}||) + \mu \tag{2.7}$$

or the equivalen model, using the untransformed data values and data weights λ_i

$$\hat{z}(\boldsymbol{x}_0) = \sum_{i=1}^{N} \lambda_i z(\boldsymbol{x}_i) \tag{2.8}$$

where $\hat{z}(\boldsymbol{x}_0)$ is the estimated value for the surface at $\boldsymbol{x}_0$, $\phi(\rho)$ is the radial basis function selected, being ρ the radial distance from point $\boldsymbol{x_0}$ to the i-th data point $\boldsymbol{x_i}$ ($\rho = ||\boldsymbol{x_i} - \boldsymbol{x_0}||$), λ_i are the weights to be estimated together to the bias value μ (or Lagrangian multiplier). This requires solving a system of n linear equations. Using the second of the two models above the procedure is then essentially the same as for Ordinary Kriging (explained further, subsection....). For clarity we outline this latter

procedure below as a series of steps using matrix notation:

- compute the $n \times n$ matrix, $\boldsymbol{D}$, of inter-point distances between all $(\boldsymbol{x_i}, \boldsymbol{y_i})$ pairs in the source dataset (or a selected subset of these);
- Apply the chosen radial basis function, $\varphi()$, to each distance in $\boldsymbol{D}$, to produce a new array $\boldsymbol{\Phi}$;
- augment $\boldsymbol{\Phi}$ with a unit column vector and a unit row vector, plus a single entry 0 in position $(n+1),(n+1)$. Call this augmented matrix A (see below for an illustration);
- compute the column vector $\boldsymbol{r}$ of distances from the grid point, $\boldsymbol{x}_0$, to each of the source data points used to create $\boldsymbol{D}$;
- apply the chosen radial basis function to each distance in $\boldsymbol{r}$, to produce a column vector $\boldsymbol{\varphi}$ and then create the (n+1) column vector c as $\boldsymbol{\varphi}$ plus a single 1 as the last entry;
- compute the matrix product $\boldsymbol{b} = \boldsymbol{A}^{-1}\boldsymbol{c}$. This provides the set of n weights to be used in the calculation of the estimated value at $\boldsymbol{x}_0$, plus the Lagrangian value μ using the linear equation 2.7.

In matrix form the system of linear equations being solved is of the form:

$$\begin{bmatrix} \boldsymbol{\Phi} & \mathbf{1} \\ \mathbf{1}' & 0 \end{bmatrix} \times \begin{bmatrix} \boldsymbol{\Lambda} \\ \mu \end{bmatrix} = \begin{bmatrix} \boldsymbol{Z} \\ 0 \end{bmatrix} \tag{2.9}$$

(i.e., $\boldsymbol{Ab} = \boldsymbol{c}$ hence $\boldsymbol{b} = \boldsymbol{A}^{-1}\boldsymbol{c}$)

where $\boldsymbol{\Phi}$ is a matrix whose i,j - element is equal to $\phi\left(||\boldsymbol{x_i} - \boldsymbol{x_j}||\right)$, 1 is a column vector containing all ones and $\boldsymbol{Z}$ is a column vector containing the z value at gauged sites.

A variety of different radial basis functions may be used. A selection of those commonly available follows. The parameter c in these expressions (not to be confused with the vector $\boldsymbol{c}$ above), which may be 0, determines the amount of smoothing.

- Multiquadric:

$$\varphi_1(\rho) = \sqrt{\rho^2 + c^2} \tag{2.10}$$

If the application of this function is restricted to some range, a, and we write $h = r/a$, then

$$\varphi_1(h) = \sqrt{h^2 + c^2} \tag{2.11}$$

is approximately linear over the range $[c, 1]$ and approximately c over the range $[0, c]$.

- Inverse multiquadric:

$$\varphi_2(\rho) = \frac{1}{\sqrt{\rho^2 + c^2}} \tag{2.12}$$

- Thin plate spline:

$$\varphi_3(\rho) = c^2\rho^2 ln(c\rho) \tag{2.13}$$

- Multilog:

$$\varphi_4(\rho) = ln\left(c^2 + \rho^2\right) \tag{2.14}$$

- Natural cubic spline:

$$\varphi_5(\rho) = \left(c^2 + \rho^2\right)^{\frac{3}{2}} \tag{2.15}$$

The radial basis function here used is the *thin plate spline*, (equation 2.13)

$$\phi(\rho) = c^2\rho^2 \ln(c\rho) \tag{2.16}$$

where c is the smoothing parameter that can be found by minimizing the prediction root mean square error (RMSE). This interpolation method requires the choice of a minimum number of neighbor points.

2.2 Stochastic methods

In the case of statistical models, coefficients/rules used to derive outputs are derived in an objective way following the theory of probability. Unlike deterministic models, in the case of statistical models, we need to follow several statistical data analysis steps before we can generate maps. This makes the whole mapping process more complicated but it eventually helps us: (a) produce more reliable/objective maps, (b) understand the sources of errors in the data and (c) depict problematic areas/points that need to be revisited. As said before, here this kind of methods are subdivided in *data-driven* and *geostatistic*. Before the description, done according to this classification, the concept of auxiliary or aided variables is introduced. They are of paramount importance in the application of the methods described hereinafter.

2.4 Environmental correlation - Generalized Linear Models (GLMs) and General Additive Models (GAMs).

If some exhaustively-sampled auxiliary variables or covariates are available in the area of interest and if they are significantly correlated with our target variable (spatial cross-correlation), and assuming that the point-values are not spatially autocorrelated, predictions can be obtained by focusing only on the deterministic part of variation:

$$Z(x) = f\left\{q_k(x)\right\} + \varepsilon \tag{2.17}$$

where q_k are the auxiliary predictors that can be used to explain the deterministic part of spatial variation. This approach to spatial prediction has a strong physical interpretation. Consider Rowe and Barnes (1994) observation that earth surface energy-moisture regimes at all scales/sizes are the dynamic driving variables of functional ecosystems at all scales/sizes. The concept of vegetation/soil-environment relationships has frequently been presented in terms of an equation with six key environmental factors as:

$$V \times S[x,\, y,\, \widetilde{t}] = f \begin{cases} s\left[x,\, y,\, \widetilde{t}\right]\ c\left[x,\, y,\, \widetilde{t}\right]\ o\left[x,\, y,\, \widetilde{t}\right] \\ r\left[x,\, y,\, \widetilde{t}\right]\ p\left[x,\, y,\, \widetilde{t}\right]\ a\left[x,\, y,\, \widetilde{t}\right] \end{cases} \tag{2.18}$$

where V stands for vegetation, S for soil, c stands for climate, o for organisms (including humans), r is relief, p is parent material or geology, a is age of the system, x, y are the coordinates and t is time dimension. This means that the predictors which are

available over entire areas of interest can be used to predict the value of an environmental variable at unvisited locations, first by modelling the relationship between the target and auxiliary environmental predictors at sample locations, and then by applying it to unvisited locations using the known value of the auxiliary variables at those locations. Common auxiliary environmental predictors used to map environmental variables are land surface parameters, remote sensing images, and geological, soil and land-use maps (McKenzie and Ryan, 1999). Because many auxiliary predictors are now also available at low or no cost, it makes this approach to spatial prediction ever more important (Hengl et al., 2007b). Functional relations between environmental variables and factors are in general unknown and the correlation coefficients can differ for different study areas, different seasons and different scales. However, in many cases, relations with the environmental predictors often reflect causal linkage: deeper and more developed soils occur at places of higher potential accumulation and lower slope; different type of forests can be found at different expositions and elevations; soils with more organic matter can be found where the climate is cooler and wetter etc. This makes this technique especially suitable for natural resource inventory teams because it allows them to validate their empirical knowledge about the variation of the target features in the area of interest. There are (at least) four groups of statistical models that have been used to make spatial predictions with the help of environmental factors (Chambers and Hastie, 1992; McBratney et al., 2003; Bishop and Minasny, 2005):

- **Classification-based model**. Classification models are primarily developed and used when we are dealing with discrete target variables (e.g. land cover or soil 1.3 Statistical spatial prediction models 21 types). There is also a difference whether Boolean (crisp) or Fuzzy (continuous) classification rules are used to create outputs. Outputs from the model fitting process are class boundaries (class centres and standard deviations) or classification rules.

- **Tree-based models**. Tree-based models are often easier to interpret when a mix of continuous and discrete variables are used as predictors (Chambers and Hastie, 1992). They are fitted by successively splitting a dataset into increasingly homogeneous groupings. Output from the model fitting process is a decision tree, which can then be applied to make predictions of either individual property values or class types for an entire area of interest.

- **Regression models**. Regression analysis employs a family of functions called **Generalized Linear Models** (GLMs), which all assume a linear relationship

between the inputs and outputs (Neter et al., 1996). Output from the model fitting process is a set of regression coefficients. Regression models can be also used to represent non-linear relationships with the use of **General Additive Models** (GAMs). The relationship between the predictors and targets can be solved using one-step data-fitting or by using iterative data fitting techniques (neural networks and similar).

The hereafter the methods that are part of the Regression model are described. These are: Linear, Multiple and Geographically Weighted Regression for **Generalized Linear Models** and Artificial Neural Network for **General Additive Models.**

2.5 Data - Driven methods

2.5.1 Linear regression and multiple regression

Regression describes the functional relationship between a dependent variable and one or more independent variable.Linear regression analysis is often used by earth scientists. For example, the equation for the regression of one variable on another may suggest hypotheses about why the two variables are related. More practically, regression can be used in situations where the dependent variable is difficult, expensive or impossible to measure, but its values can be predicted from another easily measured variable to which it is functionally related.

The simplest formulation of regression models is the simple linear regression. A linear regression analysis gives an equation for a line that describes the functional relationship between two variables and tests whether the statistics that describe this line are significantly different from zero. The simplest functional relationship between a dependent and independent variable is a straight line. The position of any point on a straight line can be described by the equation:

$$\hat{z}_i = \beta_0 + \beta_1 \, x_i \tag{2.19}$$

The simple linear regression model can also be considered in terms of the mean function and the variance function:

$$E(Z|X = x) = \beta_0 + \beta_1 x \tag{2.20}$$

$$Var(Z|X = x) = \sigma^2 \tag{2.21}$$

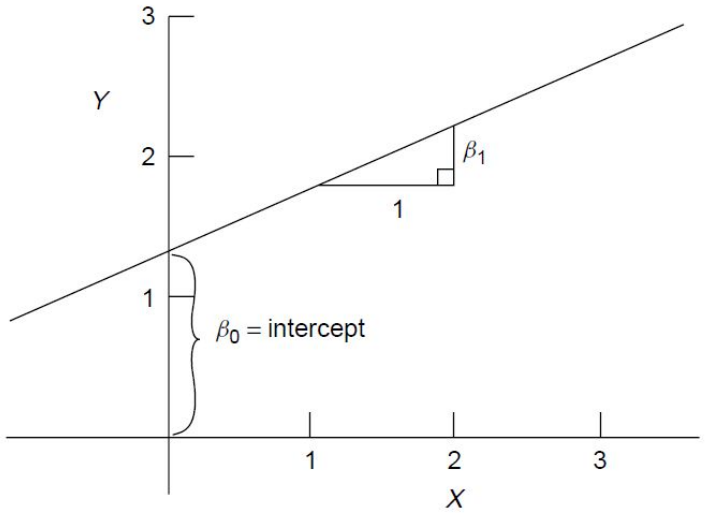

Figure 2.1: Equation of a straight line $E(Y|X = x) = \beta_0 + \beta_1 x$.

The parameters in the mean function are the intercept β_0, which is the value of $E(Z|X = x)$ when x equals zero, and the slope β_1, which is the rate of change in $E(Z|X = x)$ for a unit change in X (Figure2.1). The parameters are unknown and must be estimated using data. The variance function in 2.21 is assumed to be constant, with a positive value σ^2 that is usually unknown. Because the variance $\sigma^2 > 0$, the observed value of the ith response z_i will typically not equal its expected value $E(Z|X = x_i)$. To account for this difference between the observed data and the expected value, statisticians have invented a quantity called a statistical error, or e_i , for case i defined implicitly by the equation $x_i = E(Z|X = x_i) + e_i$ or explicitly by $e_i = z_i - E(Z|X = x_i)$. The errors e_i depend on unknown parameters in the mean function and so are not observable quantities. They are random variables and correspond to the vertical distance between the point z_i and the mean function $E(Z|X = x_i)$.

We make three important assumptions concerning the errors. First, we assume that $E(e_i|x_i) = 0$, so if we could draw a scatterplot of the e_i versus the y_i, we would have a null scatterplot, with no patterns. The second assumption is that the errors are all independent, meaning that the value of the error for one case gives no information about the value of the error for another case. The third assumption is that errors are normally distributed.

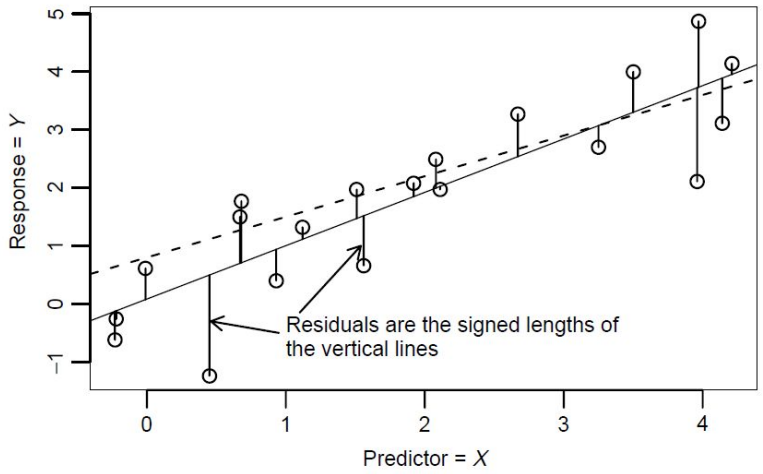

Figure 2.2: A schematic plot for ols fitting. Each data point is indicated by a small circle, and the solid line is a candidate ols line given by a particular choice of slope and intercept. The solid vertical lines between the points and the solid line are the residuals. Points below the line have negative residuals, while points above the line have positive residuals.

Many methods have been suggested for obtaining estimates of parameters in a model. Among this methods the most used is *ordinary least squares*, or *OLS*, in which parameter estimates are chosen to minimize a quantity called the residual sum of squares.

The criterion function for obtaining estimators is based on the residuals, which geometrically are the vertical distances between the fitted line and the actual y values, as illustrated in Figure 2.2. The residuals reflect the inherent asymmetry in the roles of the response and the predictor in regression problems.

The OLS estimators are those values β_0 and β_1 that minimize the function Residual Sum of Squares (RSS):

$$RSS(\beta_0, \beta_1) = \sum_{i=1}^{n} [z_i - (\beta_0 + \beta_1 x_i)]^2 \tag{2.22}$$

When evaluated at $(\hat{\beta}_0, \hat{\beta}_1)$, we call the quantity $RSS(\hat{\beta}_0, \hat{\beta}_1)$ the *residual sum of squares*, or just RSS.

The least squares estimates are given by the expressions

$$\hat{\beta}_1 = \frac{SXZ}{SXX} = r_{xz}\frac{SD_z}{SD_x} = r_{xz}\left(\frac{SZZ}{SXX}\right)^2 \tag{2.23}$$

$$\hat{\beta}_0 = \overline{z} - \hat{\beta}_1\overline{x} \tag{2.24}$$

where $\overline{y} = \sum \frac{y_i}{n}$ is the simple average of y , $\overline{z} = \sum \frac{z_i}{n}$ is the simple average of z, $SYY = \sum (y_i - \overline{y})^2$ is the sum of squares for y's, $SZZ = \sum (z_i - \overline{z})^2$is the sum of squares for z's, $SYZ = \sum (y_i - \overline{y})(z_i - \overline{z})$ is sum of cross-product, $SD_y = \sqrt{\frac{SYY}{n-1}}$ is sample standard deviation of the y's, $SD_z = \sqrt{\frac{SZZ}{n-1}}$ is sample standard deviation of the z's.

The several forms for $\hat{\beta}_1$ are all equivalent. We emphasize again that OLS produces estimates of parameters but not the actual values of the parameters.

Simple regression techniques have been widely applied in spatial analysis for very long time. Another common regression-based approach to spatial prediction is the multiple linear regression (Draper and Smith, 1998).

The general multiple linear regression model with response Z and terms $x_1, ..., x_p$ will have the form

$$E(Z|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{2.25}$$

The symbol X in $E(Z|X)$ means that we are conditioning on all the terms on the right side of the equation. Similarly, when we are conditioning on specific values for the predictors $x_1, ..., x_p$ that we will collectively call x, we write

$$E(Z|X = x) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p \tag{2.26}$$

As in the case of simple linear regression, the β_i (with $i = 1, ...p$) are unknown parameters we need to estimate. Equation 2.25 is a linear function of the parameters, which is why this is called linear regression. When $p = 1$, X has only one element, and we get the simple regression problem. When $p = 2$, the mean function 2.25 corresponds to a plane in three dimensions, as shown in Figure 2.3. When $p > 2$, the fitted mean function is a hyperplane, the generalization of a p-dimensional plane in a $(p + 1)$-dimensional space.

From the initial collection of potential predictors, we have computed a set of $p+1$ terms, including an intercept, $X = (X_0, X_1, ..., X_p)$. The mean function and variance function for multiple linear regression are

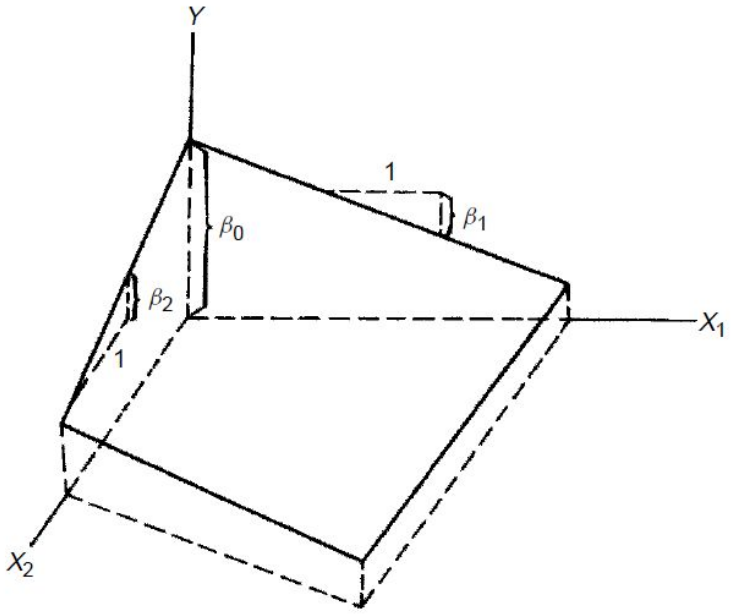

Figure 2.3: A linear regression surface with $p = 2$ predictors.

$$E(Z|X) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p \tag{2.27}$$

$$Var(Z|Y) = \sigma^2 \tag{2.28}$$

Both the β_i and σ^2 are unknown parameters that need to be estimated.

Suppose we have observed data for n cases or units, meaning we have a value of Y and all of the terms for each of the n cases. We have symbols for the response and the terms using matrices and vectors. We define:

$$\mathbf{Z} = \begin{pmatrix} z_1 \\ z_2 \\ \vdots \\ z_n \end{pmatrix} \qquad \mathbf{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \tag{2.29}$$

so $\boldsymbol{Z}$ is an $n \times 1$ vector and $\boldsymbol{X}$ is an $n \times (p+1)$ matrix. We also define $\boldsymbol{\beta}$ to be a $(p+1) \times 1$ vector of regression coefficients and $\boldsymbol{e}$ to be the $n \times 1$ vector of statistical

errors,

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \qquad and\ \boldsymbol{e} = \begin{pmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{pmatrix} \tag{2.30}$$

The matrix $\boldsymbol{Y}$ gives all of the observed values of the terms. The ith row of $\boldsymbol{X}$ will be defined by the symbol $\boldsymbol{x^T}_i$, which is a $(p+1) \times 1$ vector for mean functions that include an intercept. Even though $\boldsymbol{x}_i$ is a row of $\boldsymbol{X}$, we use the convention that all vectors are column vectors and therefore need to write $\boldsymbol{x^T}_i$ to represent a row. An equation for the mean function evaluated at $\boldsymbol{x}_i$ is

$$E(Z|X = \boldsymbol{x}_i) = \boldsymbol{x_i^T}\beta = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} \tag{2.31}$$

In matrix notation, we will write the multiple linear regression model as

$$\boldsymbol{Z} = \boldsymbol{X}\beta + \boldsymbol{e} \tag{2.32}$$

The ith row of 2.32 is $z_i = \boldsymbol{y_i^T}\beta + e_i$.

The least squares estimate $\hat{\beta}$ of β is chosen to minimize the residual sum of squares function

$$RSS(\beta) = (z_i - \boldsymbol{x_i^T}\beta)^2 = (\boldsymbol{Z} - \boldsymbol{X}\beta)^T\,(\boldsymbol{Z} - \boldsymbol{X}\beta) \tag{2.33}$$

The OLS estimates can be found from 2.33 by differentiation in a matrix. The ols estimate is given by the formula

$$\hat{\boldsymbol{\beta}} = (\boldsymbol{X}\,\boldsymbol{X})^{-1}\,\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Z} \tag{2.34}$$

provided that the inverse $\left(\boldsymbol{X^T}\,\boldsymbol{X}\right)$ exists. The estimator $\hat{\boldsymbol{\beta}}$ depends only on the sufficient statistics $\left(\boldsymbol{X^T}\,\boldsymbol{X}\right)$ and $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{Z}$, which are matrices of uncorrected sums of squares and cross-products.

Linear least-squares estimates can behave badly when the error distribution is not normal, particularly when the errors are heavy-tailed. One remedy is to remove influential observations, as said before, from the least-squares fit. Another approach, termed *robust regression*, is to employ a fitting criterion that is not as vulnerable as least squares to unusual data. The most common general method of robust regression is *M-estimation*, introduced by Huber (1964). Consider the linear model

$$z_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon = \boldsymbol{x_i^T \beta} + e_i \tag{2.35}$$

for the ith of n observations.

The fitted model is

$$z_i = \hat{\beta}_0 + \hat{\beta}_1 y_{i1} + \hat{\beta}_2 y_{i2} + \cdots + \hat{\beta}_p y_{ip} + e_i = \boldsymbol{x_i^T \hat{\beta}} + \hat{e}_i \tag{2.36}$$

The general M-estimator minimizes the objective function

$$\sum_{i=1}^{n} \rho\left(e_i\right) = \sum_{i=1}^{n} \rho\left(z_i - \boldsymbol{x}^T \hat{\beta}\right) \tag{2.37}$$

where the function ρ gives the contribution of each residual to the objective function. A reasonable ρ should have the following properties:

- $\rho(e) \geq 0$
- $\rho(0) = 0$
- $\rho(e) = \rho(-e)$
- $\rho(e_i) \geq \rho(e_i')\ for\ |e_i| > |e_i'|$

For example, for least-squares estimation, $\rho(e_i) = e_i^2$. Let $\psi = \rho$' be the derivative of ρ. Differentiating the objective function with respect to the coefficients, $\hat{\beta}$, and setting the partial derivatives to 0, produces a system of $k+1$ estimating equations for the coefficients:

$$\sum_{i=1}^{n} \psi\left(z_i - \boldsymbol{x_i^T \hat{\beta}}\right) \boldsymbol{x_i^T} = \mathbf{0} \tag{2.38}$$

Define the *weight function* $w(\hat{e}) = \psi(\hat{e})/\hat{e}$, and let $w_i = w(\hat{e}_i)$. Then the estimating equations may be written as

$$\sum_{i=1}^{n} w_i \left(z_i - \boldsymbol{x_i^T \hat{\beta}}\right) \boldsymbol{x_i^T} = 0 \tag{2.39}$$

Solving the estimating equations is a weighted least-squares problem, minimizing $\sum w_i^2 e_i^2$. The weights, however, depend upon the residuals, the residuals depend upon the estimated coefficients, and the estimated coefficients depend upon the weights.

Method	*Objective Function*	*Weight Function*
Least-Squares	$\rho_{\mathrm{LS}}(e) = e^2$	$w_{\mathrm{LS}}(e) = 1$
Huber	$\rho_{\mathrm{H}}(e) = \begin{cases} \frac{1}{2}e^2 & \text{for } \lvert e\rvert \le k \\ k\lvert e\rvert - \frac{1}{2}k^2 & \text{for } \lvert e\rvert > k \end{cases}$	$w_{\mathrm{H}}(e) = \begin{cases} 1 & \text{for } \lvert e\rvert \le k \\ k/\lvert e\rvert & \text{for } \lvert e\rvert > k \end{cases}$
Bisquare	$\rho_{\mathrm{B}}(e) = \begin{cases} \frac{k^2}{6}\left\{1-\left[1-\left(\frac{e}{k}\right)^2\right]^3\right\} & \text{for } \lvert e\rvert \le k \\ k^2/6 & \text{for } \lvert e\rvert > k \end{cases}$	$w_{\mathrm{B}}(e) = \begin{cases} \left[1-\left(\frac{e}{k}\right)^2\right]^2 & \text{for } \lvert e\rvert \le k \\ 0 & \text{for } \lvert e\rvert > k \end{cases}$

Table 2.1: Objective function and weight function for least-squares, Huber, and bisquare estimators.

An iterative solution (called *iteratively reweighted least-squares*, IRLS) is therefore required:

1. Select initial estimates $\hat{\boldsymbol{\beta}}^{(0)}$, such as the least-squares estimates.

2. At each iteration t, calculate residuals $\hat{\beta}_i^{(t-1)}$ and associated weights $w_i^{(t-1)} = w\left[\hat{\beta_i^{(t-1)}}\right]$ from the previous iteration.

3. Solve for new weighted-least-squares estimates

$$\hat{\boldsymbol{\beta}}^{(t)} = \left[\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}^{(\boldsymbol{t-1})}\boldsymbol{X}\right]^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{W}^{(\boldsymbol{t-1})}\boldsymbol{z} \tag{2.40}$$

where $\boldsymbol{X}$ is the model matrix, with $\boldsymbol{x}_i^T$ i as its ith row, and $\boldsymbol{W}^{(t-1)} = diag\left\{w_i^{(t-1)}\right\}$is the current weight matrix.

Steps 2. and 3. are repeated until the estimated coefficients converge.

The asymptotic covariance matrix of $\hat{\boldsymbol{\beta}}$ is

$$\nu\left(\hat{\boldsymbol{\beta}}\right) = \frac{E\left(\psi^2\right)}{\left[E\left(\psi'\right)\right]^2}\left(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}\right)^{-1} \tag{2.41}$$

Using $\sum\left[\psi\left(e_i\right)\right]^2$to estimate $E\left(\psi^2\right)$, and $\sum\left[\psi'\left(e_i\right)/n\right]^2$to estimate $\left[E\left(\psi'\right)\right]^2$ produces the *estimated* asymptotic covariance matrix, $\hat{\nu}\left(\hat{\boldsymbol{\beta}}\right)$ (which is not reliable in small samples).

Concerning objective Functions, Figure2.4 compares the objective functions, and the corresponding ψ and weight functions for three M-estimators: the familiar least-squares estimator; the Huber estimator; and the Tukey *bisquare* (or *biweight*) estimator. The objective and weight functions for the three estimators are also given in Table2.1.

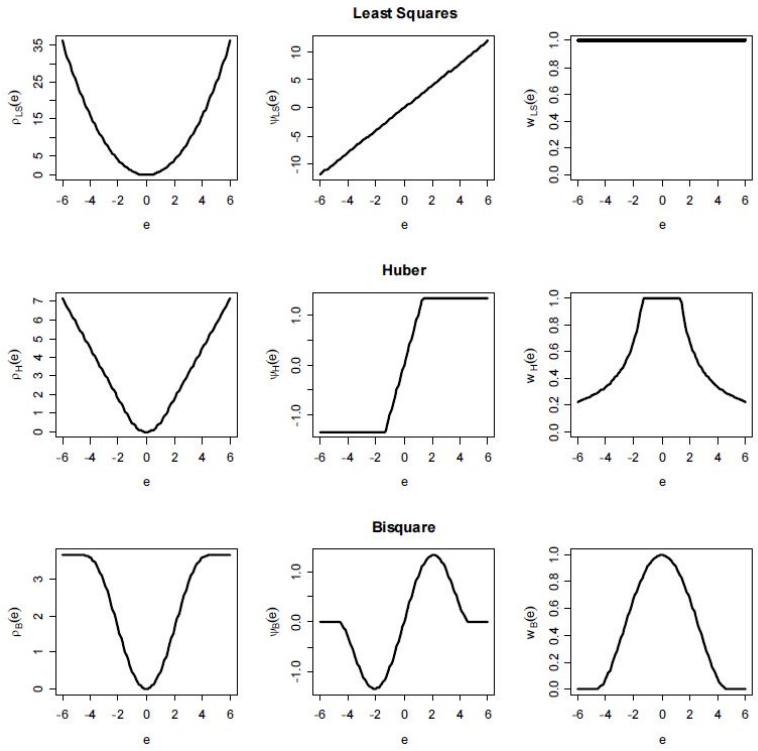

Figure 2.4: Objective, ψ, and weight functions for the least-squares (top), Huber (middle), and bisquare (bottom) estimators. The tuning constants for these graphs are p = 1.345 for the Huber estimator and p = 4.685 for the bisquare. (One way to think about this scaling is that the standard deviation of the errors, σ, is taken as 1.)

2.5.2 Geografically Weighted Regression

The biggest criticism of pure regression approach to spatial prediction is that the position of points in the geographical space is completely ignored, both during the model fitting and prediction. Imagine if we are dealing with two point datasets where one data set is heavily clustered, while the other is well-spread over the area of interest — these has to be a way to account for the clustering of the points so we take the model derived using the clustered points with much bigger caution. One way to account for this problem is to take the distance between the points into account during the estimation of the regression coefficients. This can be achieved by using the Geographically Weighted Regression (GWR) (Fotheringham et al., 2002).

Consider a global regression model written as:

$$z_i = \beta_0 + \sum_p \beta_p x_{ip} + e_i \tag{2.42}$$

GWR extends this traditional regression framework by allowing local rather than global parameters to be estimated so that the model is rewritten as:

$$z_i = \beta_0(x_i, y_i) + \sum_p \beta_p(x_i, y_i) x_{ip} + e_i \tag{2.43}$$

where (x_i, y_i) denotes the coordinates of the ith point in space and $\beta_p(x_i, y_i)$ is a realization of the continuous function $\beta_p(x_i, y_i)$ at point i. That is, we allow there to be a continuous surface of parameter values, and measurements of this surface are taken at certain points to denote the spatial variability of the surface. Note that equation 2.42 is a special case of equation 2.43 in which the parameters are assumed to be spatially invariant. Thus the GWR equation in 2.43 recognizes that spatial variations in relationships might exist and provides a way in which they can be measured.

The calibration process in GWR can be thought of as a trade-off between bias and standard error. Assuming the parameters exhibit some degree of spatial consistency, then values near to the one being estimated should have relatively similar magnitudes and signs. Thus, when estimating a parameter at a given location little can approximate 2.43 in the region of i by 2.42, and perform a regression using a subset of the points in the data set that are close to i. Thus, the $\beta_p(x_i, y_i)$ are estimated for i in the usual way and for the next i, a new subset of 'nearby' points is used, and so on. These estimates will have some degree of bias, since the coefficients of 2.43 will exhibit some drift across the local calibration subset. However, if the local sample is large enough, this will allow a calibration to take place, albeit a biased one. The greater the size of

the local calibration subset, the lower the standard errors of the coefficient estimates; but this must be offset against the fact that enlarging this subset increases the chance that the coefficient *drift* introduces bias. To reduce this effect, one final adjustment to this approach may also be made. Assuming that points in the calibration subset farther from i are more likely to have differing coefficients, a weighted calibration is used, so that more influence in the calibration is attributable to the points closer to i.

As noted above, the calibration of equation 2.43 assumes implicitly that observed data near to location i have more of an influence in the estimation of the $\beta_p(x_i, y_i)$ than to data located farther from i (see Figures 2.6and 2.7). In essence, the equation measures the relationships inherent in the model around each location i. Hence weighted least squares provides a basis for understanding how GWR operates. In GWR an observation is weighted in accordance with its proximity to location i so that the weighting of an observation is no longer constant in the calibration but varies with i. Data from observations close to i are weighted more than data from observations farther away. That is,

$$\hat{\boldsymbol{\beta}}(x_i, y_i) = \left(\boldsymbol{Z^T W}(x_i, y_i)\boldsymbol{Z}\right)^{-1} \boldsymbol{Z^T W}(x_i, y_i)\boldsymbol{z} \tag{2.44}$$

where the bold type denotes matrix, $\hat{\boldsymbol{\beta}}$ represents an estimate of $\boldsymbol{\beta}$ and $\boldsymbol{W}(x_i, y_i)$ is an n by n matrix whose off-diagonal elements are zero and whose diagonal elements denote the geographical weighting of each of the n observed data for regression point i.

To see this more clearly, consider the classical regression, previously seen in section 2.5.1, equation in matrix form:

$$\boldsymbol{Z} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{2.45}$$

where the vector of parameters to be estimated, β, is constant over space and is estimated by

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X^T X}\right)^{-1} \boldsymbol{X^T Z} \tag{2.46}$$

The GWR equivalent is

$$\boldsymbol{Z} = (\boldsymbol{\beta} \times \boldsymbol{X})\mathbf{1} + \boldsymbol{\varepsilon} \tag{2.47}$$

where $\times$ is a logical multiplication operator in which each element of $\boldsymbol{\beta}$ is multiplied

by the corresponding element of $\boldsymbol{Y}$. If there are n data points and p explanatory variables, both $\boldsymbol{\beta}$ and $\boldsymbol{Y}$ will have dimensions of $n \times (p+1)$ and $\mathbf{1}$ is a $(p+1) \times 1$ a vector of 1s. The matrix $\boldsymbol{\beta}$ now consists of n sets of local parameters and has the following structure:

$$\boldsymbol{\beta} = \begin{pmatrix} \beta_0(x_1, y_1) & \beta_1(x_1, y_1) & \cdots & \beta_k(x_1, y_1) \\ \beta_0(x_2, y_2) & \beta_1(x_2, y_2) & \cdots & \beta_k(x_2, y_2) \\ \vdots & \vdots & \ddots & \vdots \\ \beta_0(x_n, y_n) & \beta_0(x_n, y_n) & \cdots & \beta_0(x_n, y_n) \end{pmatrix} \tag{2.48}$$

The parameters in each row of the: both matrix are estimated by

$$\hat{\boldsymbol{\beta}} = \left(\boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{W}(i) \boldsymbol{X}\right)^{-1} \boldsymbol{X}^{\boldsymbol{T}} \boldsymbol{W}(i) \boldsymbol{Z} \tag{2.49}$$

where i represents a row of the matrix in 2.48 and $\boldsymbol{W}(i)$ is an n by n spatial weighting matrix of the form

$$\boldsymbol{W}(i) = \begin{pmatrix} w_{i1} & 0 & \cdots\cdots & 0 \\ 0 & w_{i2} & \cdots\cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots\cdots & w_{in} \end{pmatrix} \tag{2.50}$$

where w_{in} is the weight given to data point n in the calibration of the model for location i. The estimator in equation 2.49 is a weighted least squares estimator but rather than having a constant weight matrix, the weights in GWR vary according to the location of point i. Hence the weighting matrix has to be computed for each point i and the weights depict the proximity of each data point to the location of i with points in closer proximity carrying more weight in the estitmation of the parameters for location i. Notice, however. that in equations 2.49 and 2.50 there is no reason that i has to be the location of a data point. Local estitmates of the parameters can in fact be derived for any point in space, regardless of whether or not that point is one at which data have been observed.

One aspect of GWR is that the estimated parameters are, in part dependent on the weighting function or kernel selected. There are two different approachs to select an appropriate kernel to estimate the parameter: *Fixed Spatial Kernels* and *Adaptive Spatial Kernels*. As before explicated, with GWR a region is described around a regression point and all the data points within this region is then used to calibrate a model. This process was repeated for all regression points. Each data point is

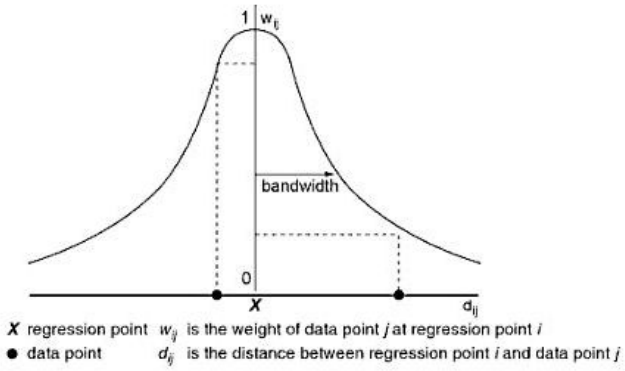

Figure 2.5: A spatial kernel

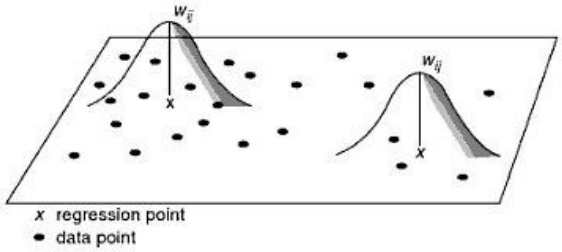

Figure 2.6: GWR with fixed spatial kernels

weighted by its distance from the regression point; hence, data points closer to the regression point are weighted more heavily in the local regression than are data points farther away. Graphically, the method is that of fitting a spatial kernel to the data as described in Figures 2.6 and 2.7. For a given regression point, the weight of a data point is at a maximum when it shares the same location as the regression point. This weight decreases continuously as the distance between the two points increases. In this way, a regression model is calibrated locally simply by moving the regression point across the region. For each location, the data will be weighted differently so that the results of any one calibration are unique to a particular location. By plotting the results of these local calibrations on a map, surfaces of parameter estimates, or any other display which is appropriate, can be generated. In practice, the results of GWR are relatively insensitive to the choice of weighting function but they are sensitive to the bandwidth of the particular weighting function chosen to that the determination of the optimal value of the bandwidth is necessary as part of GWR.

A potential problem that might arise in the application of GWR with fixed spatial

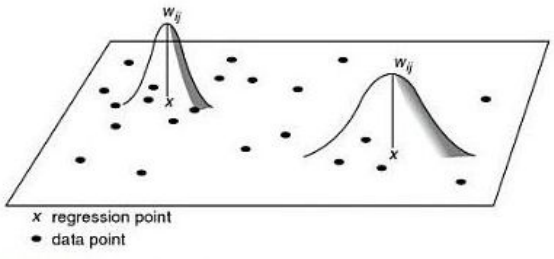

Figure 2.7: GWR with Adaptive Spatial Kernels

kernels is that of some regression points, where data are sparse, the local models might be calibrated on very few data points, giving rise to parameter estimated with large standard errors and resulting surfaces which are *undersmoothed*. In extreme cases, the estimation of some parameters might be impossible due to insufficient variation in small samples. Accordingly, to reduce these problems, the spatial kernels in GWR can be made to adapt themselves in size to variations in the density of the data so that the kernels have larger bandwidths where the data are sparse and have smaller bandwidths where the data are plentiful. There are various ways in which such adaptive bandwidths can be achieved and some of these are described in the following section. Graphically, the application of GWR with adaptive spatial kernels is described in Figure 2.7.

As until this point said, in GWR ,$\boldsymbol{W}(x_i, y_i)$ or, in more convenient terms, $\boldsymbol{W}(i)$, is a weighting scheme based on the proximity of the regression point i to the data points around i without an explicit relationship being stated. The choice of such a relationship is an important issue to application of this method, so it will be considered here. First, consider the implicit weighting scheme of the OLS framework in equation 2.42. Here

$$w_{ij} = 1 \quad \forall i, j \tag{2.51}$$

where j represents a specific point in space at which data are observed and i represents any point in space for which parameters are estimated. That is, in the global model each observation has a weight of unity. An initial step towards weighting based on locality might be to exclude from the model calibration observations that are further than some distance d if from the regression point. This would be equivalent to setting their weights to zero, giving a weighting function of

$$\begin{aligned} w_{ij} &= 1 \quad if\ d_{ij} < d \\ w_{ij} &= 0 \quad otherwise \end{aligned} \tag{2.52}$$

The use of this weighting scheme would simplify the calibration procedure because at every regression point only a subset of the data points would be used to calibrate the model. However, this spatial weighting function has the problem of discontinuity. As the regression point changes, the estimated coefficients could change drastically as a data point moves into or out of the window around i. Although sudden changes in the parameters over space might genuinely occur, in this case changes in their estimates would occur as artefacts of the arrangement of data points, rather than necessarily depicting any underlying process in the relationship under investigation. One way to combat the problem of discontinuities of weights is to specify w_{ij} as a continuous function of d_{ij}, the distance between i and j. One obvious choice is:

$$w_{ij} = exp\left[-\frac{1}{2}\left(\frac{d_{ij}}{b}\right)^2\right] \tag{2.53}$$

where b is referred to as the bandwidth. If i and j coincide (that is, i also happens to be a point in space at which data are observed), the weighting of data at that point will be unity and the weighting of other data will decrease according to a Gaussian cunt as the distance between i and j increases. The empirical GWR results described above are based on Gaussian or near-Gaussian weighting func-tions. An alternative kernel utilises the bi-square function,

$$\begin{aligned} w_{ij} &= \left[1 - \left(\tfrac{d_{ij}}{b}\right)^2\right]^2 \quad if\ d_{ij} < d \\ w_{ij} &= 0 \qquad\qquad\qquad otherwise \end{aligned} \tag{2.54}$$

which is particularly useful because it provides a continuous, near-Gaussian weighting function up to distance b from the regression point and then zero weights any data point beyond b.

The above types of spatial kernel are fixed in terms of their shape and magnitude over space and belong to a class of kernels exemplified in Figure 2.11. It could be argued that the kernels should be allowed to vary spatially, with the kernels being smaller in regions where the density of data points is high and larger where the density of data points is low, as shown in Figure 2.13. The rationale for this is twofold: (i) where data points are dense there is more scope for examining changes in relationships over relatively small distances and such changes might be missed with

larger kernels; and (ii) in regions where data are scarce, the standard errors of the coefficients estimated in GWR when fixed kernels are used will be high because the number of data points used will be small. In essence the problem of fixed kernels in regions where data are dense is that the kernels are larger than they need be and hence the estimates obtained from them are more likely to suffer from bias. Conversely, the problem with fixed kernels in regions where data are scarce is one of inefficiency: the kernels are smaller than they need be to estimate the parameters reliably. Both problems can be reduced by performing GWR with spatially varying kernels.

At least three methods of producing spatially varying kernels exist. One is to rank the data points in terms of their distance front each point i so that R_{ij} is the rank of the jth point from i in terms of the distance j is from i. The closest data point to i has a weight of 1 and the weights decrease as the rank increases according to some continuous function such as:

$$w_{ij} = exp\left(\frac{-R_{ij}}{b}\right) \tag{2.55}$$

This is will automatically reduce the bandwidth of kernels in regions with large amounts of data because the distance to say the 10th nearest data point will be much less than when the regression point is in a region with relatively few data points.

A second, and more complex, method of producing spatially varying kernels is to ensure that the sum of the weights for any point i is a constant, C. In areas where the density of data points is high, the kernel will have to contract to ensure the sum of the weights is equal to C whereas in areas where the density of data points is low, the kernel will have to expand. Formally,

$$\sum_j w_{ij} = C \quad for \quad all \quad i \tag{2.56}$$

which is a constraint that has to be attached to one of the stationary weighting functions described in equations 2.53 and 2.54. One could simply choose an appropriate value of C and then calibrate the weighting function given this value of C.

A third spatially varying weighting method involves a function that is related to the Nth nearest neighbours of point i. That is,

$$\begin{gathered} w_{ij} = 1 \; if \; j \; is \; one \; of \; the \; Nth \; nearest \; neighbours \; of \; i \\ w_{ij} = 0 \; otherwise \end{gathered} \tag{2.57}$$

or, given that (2.28) re-introduces discontinuities

$$w_{ij} = \left[1 - \left(\frac{d_{ij}}{b}\right)^2\right]^2 \; if \; j \; is \; one \; of \; the \; Nth \; nearest \; neighbours \; of \; i \; and \; b \; is \; the \; distance \; to \; the \; N$$
$$w_{ij} = 0 \; otherwise \tag{2.58}$$

If either of the weighting functions in equations 2.57 or 2.58 is used in the GWR, the calibration of the model involves the estimation of N. The value of N is the number of data points to he included within the calibration of the local model and the weighting function determines the weight of each data point up to the Nth one. Weights for all data points beyond the Nth one are set to zero. Hence, the bi-square weighting function, which reaches zero at the Nth data point is a logical one to employ in GWR. The empirical examples of GWR with an adaptive bandwidth described above use nearest neighbour weighting with a bi-square decay function. A comparison of discrete and continuous weighting functions is described by Fotheringham et al. (1997).

Whatever the specific weighting function employed, the essential idea of GWR is that for each regression point i there is a 'bump of influence' around i described by the weighting function such that sampled observations near to i have more influence in the estimation of the parameters than do sampled observations farther away.

2.5.3 Artificial Neural Network

Artificial Neural Networks are in use today in a multitude of applications including, but by no means limited to: data analysis (e.g. data mining), robotics (e.g. control of adaptable robots), time series prediction and estimation (e.g. weather forecasting), bioinformatics (e.g. DNA sequencing), pattern recognition (e.g. speech recognition), financial modelling (e.g. stockmarket prediction). In this work, artificial neural networks (ANN) are applied for analysing spatial hydroclimatic data .

A neural network is an interconnected assembly of simple processing elements, *units* or *nodes*, whose functionality is loosely based on the complex structure of biologic nervous system. The processing ability of the network is stored in the interunit connection strengths, or *weights*, obtained by a process of adaptation to, or *learning* from, a set of training patterns. The main characteristics of the neural networks are: adaptation and learning capability, generalization capability, pattern recognition and data classification capability. The neural networks can be defined as a parallel distributed processing (PDP), i.e. the units that constitute the neural network can

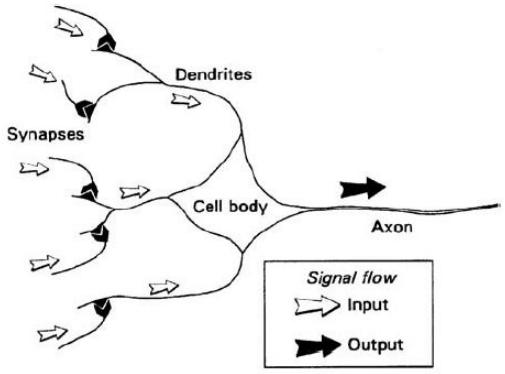

Figure 2.8: Essential components of a neuron shown in stylized form.

carry out their computations simultaneously.

In order to flesh out this definition it is necessary first take a quick look at some basic neurobiology. The human brain consists of an estimated 10^{11} (100 billion) nerve cells or neurons (Figure2.8).

Neurons communicate through electrical signals that are short-lived impulses or "spikes" in the voltage of the cell wall or membrane. The interneuron connections are mediated by electrochemical junctions called *synapses*, which are located on branches of the cell referred to as *dendrites*. Each neuron typically receives many thousands of connections from other neurons and is therefore constantly receiving a multitude of incoming signals, which eventually reach the *cell body*. Here, they are integrated or summed together in some way and, roughly speaking, if the resulting signal exceeds some threshold then the neuron will "fire" or generate a voltage impulse in response. This is then transmitted to other neurons via a branching fibre known as the *axon*. In determining whether an impulse should be produced or not, some incoming signals produce an inhibitory effect and tend to prevent firing, while others are excitatory and promote impulse generation. The distinctive processing ability of each neuron is then supposed to reside in the type, excitatory or inhibitory, and strength of its synaptic connections with other neurons. It is this architecture and style of processing that are to incorporate in neural networks and, because of the emphasis on the importance of the interneuron connections, this type of system is sometimes referred to as being *connectionist* and the study of this general approach as *connectionism*. This terminology is often the one encountered for neural networks in the context of psychologically

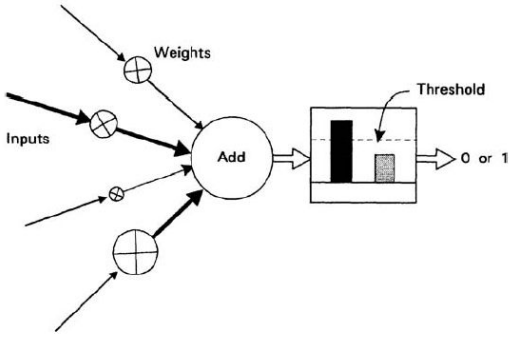

Figure 2.9: Simple artificial neuron.

inspired models of human cognitive function. The artificial equivalents of biological neurons are the *nodes* or *units* in a preliminary definition and a prototypical example is shown in Figure 2.9. Synapses are modelled by a single number or weight so that each input is multiplied by a weight before being sent to the equivalent of the cell body. Here, the weighted signals are summed together by simple arithmetic addition to supply a node activation. In the type of node shown in Figure 2.9, the so-called threshold logic unit (TLU), the activation is then compared with a threshold; if the activation exceeds the threshold, the unit produces a highvalued output (conventionally "1"), otherwise it outputs zero. In the figure, the size of signals is represented by the width of their corresponding arrows, weights are shown by multiplication symbols in circles, and their values are supposed to be proportional to the symbol's size; only positive weights have been used. The TLU is the simplest (and historically the earliest (McCulloch and Pitts, 1943)) model of an artificial neuron.

The term "network" will be used to refer to any system of artificial neurons. This may range from something as simple as a single node to a large collection of nodes in which each one is connected to every other node in the net. One type of network is shown in Figure2.10. Each node is now shown by only a circle but weights are implicit on all connections. The nodes are arranged in a layered structure in which each signal emanates from an input and passes via two nodes before reaching an output beyond which it is no longer transformed. This feedforward structure is only one of several available and is typically used to place an input pattern into one of several classes according to the resulting pattern of outputs. Notice, moreover, the emphasis on learning from experience. In real neurons the synaptic strengths may, under certain circumstances, be modified so that the behaviour of each neuron can

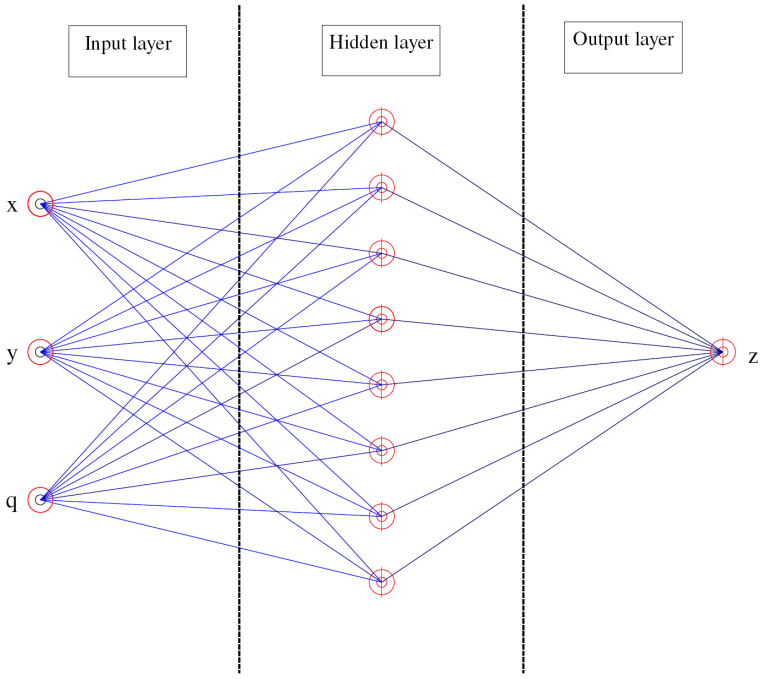

Figure 2.10: Simple example of neural network.

change or adapt to its particular stimulus input. In artificial neurons the equivalent of this is the modification of the weight values. In terms of processing information, there are no computer programs here — it is supposed that the "knowledge" network to be stored in its weights, which evolve by a process of adaptation to stimulus from a set of pattern examples. In one training paradigm called *supervised learning*, used in conjunction with nets of the type shown in Figure 2.10, an input pattern is presented to the net and its response then compared with a target output.

2.5.3.1 Neural Network Structure

As said before we can defined neural network as a distributed parallel processing (PDP). There are eight major aspects of a parallel distributed processing model:

- a set of processing units (neurons);
- a state of activation;
- an output function for each unit;

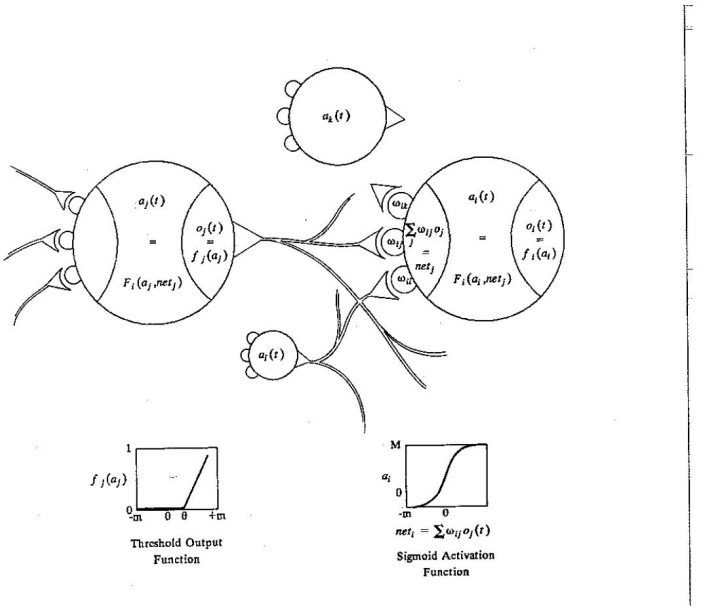

Figure 2.11: Basic components of a distributed parallel processing

- a pattern of connectivity among units;
- a propagation rule for propagating patterns of activities through the network of connectivities;
- an activation rule for combining the inputs impinging on a unit with the current state of that unit to produce a new level of activation for the unit.
- a learning rule whereby patterns of connectivity are modified by experience;
- an environment within which the system must operate.

Figure 2.11 illustrates the basic aspects of these systems. There is a set of processing units generally indicated by circles in the diagrams; at each point in time, each unit u_i, has an activation value, denoted in the diagram as $a_i(t)$; this activation value is passed through a function f_i to produce an output value $o_i(t)$. This output value can be seen as passing through a set of unidirectional connections (indicated by lines or arrows in our diagrams) to other units in the system. There is associated with each connection a real number, usually called the *weight* or *strength* of the connection designated w_{ij} which determines the amount of effect that the first unit has on the second. All of the inputs must then be combined by some operator (usually addition)—and the combined inputs to a unit, along with its current activation value, determine, via a function F, its new activation value. The figure shows illustrative

examples of the function f and F. Finally, these systems are viewed as being plastic in the sense that the pattern of interconnections is not fixed for all time; rather, the weights can undergo modification as a function of experience. In this way the system can evolve. What a unit represents can change with experience, and the system can come to perform in substantially different ways. Now the major aspects of a parallel distributed processing model are described:

1. *A set of processing units.* All of the processing of a PDP model is carried out by these units. They are only relatively simple units, each doing it own relatively simple job. A unit's job is simply to receive input from its neighbors and, as a function of the inputs it receives, to compute an output value which it sends to its neighbors. The system is inherently parallel in that many units can carry out their computations at the same time. Within any system here modelled, it is useful to characterize three types of units: *input*, *output*, and *hidden*. Input units receive inputs from sources external to the system under study. These inputs may be either sensory input or inputs from other parts of the processing system in which the model is embedded. The output units send signals out of the system. They may either directly affect motoric systems or simply influence other systems external to the ones modelled. The hidden units are those whose only inputs and outputs are within the system we are modeling. They are not "visible" to outside systems.

2. *The state of activation.* In addition, to the set of units, the state of the system at time t has to be represented. This is primarily specified by a vector of N real numbers, $\boldsymbol{a}(t)$, representing the pattern of activation over the set of processing units. Each element of the vector stands for the activation of one of the units at time t. The activation of unit u_i at time t is designated $a_i(t)$. It is the pattern of activation over the set of units that captures what the system is representing at any time. It is useful to see processing in the system as the evolution, through time, of a pattern of activity over the set of units. Activation values may be continuous or discrete. If they are continuous, they may be unbounded or bounded. If they are discrete, they may take binary values or any of a small set of values.

3. *Output of the units.* Units interact by transmitting signals to their neighbors. The strength of their signals, and therefore the degree to which they affect their neighbors, is determined by their degree of activation. Associated with each unit, u_i, there is an output function, $f_i(a_i(t))$, which maps the current state

of activation $a_i(t)$ to an output signal $o_i(t)$ (i.e., $o_i(t) = f(a_i(t))$). In vector notation, we represent the current set of output values by a vector, $\boldsymbol{o}(t)$. In some of our models the output level is exactly equal to the activation level of the unit. In this case f is the identity function $f(x) = x$. More often, however, f is some sort of threshold function so that a unit has no affect on another unit unless its activation exceeds a certain value. Sometimes the function f is assumed to be a stochastic function in which the output of the unit depends in a probabilistic fashion on its activation values.

4. *The pattern of connectivity.* Units are connected to one another. In a pattern of connectivity that constitutes what the system knows and determines how it will respond to any arbitrary input. Specifying the processing system and the knowledge encoded therein is, in a parallel distributed processing model, a matter of specifying this pattern of connectivity among the processing units. In many cases, it can be assumed that each unit provides an additive contribution to the input of the units to which it is connected. In such cases, the total input to the unit is simply the weighted sum of the separate inputs from each of the individual units. That is, the inputs from all of the incoming units are simply multiplied by a weight and summed to get the overall input to that unit. In this case, the total pattern of connectivity can be represented by merely specifying the weights for each of the connections in the system. A positive weight represents an *excitatory* input and a negative weight represents an *inhibitory* input. It is often convenient to represent such a pattern of connectivity by a weight matrix $\boldsymbol{W}$ in which the entry w_{ij} represents the strength and sense of the connection from unit u_j to unit u_i. The weight w_{ij} is a positive number if unit u_j excites unit u_i; it is a negative number if unit u_j inhibits unit u_i; and it is 0 if unit u_j has no direct connection to unit u_i. The absolute value of w_{ij} specifies the *strength of the connection.* A given unit may receive inputs of different kinds whose effects are separately summated. Sometimes, more complex inhibition/excitation combination rules are required. In such cases it is convenient to have separate connectivity matrices for each kind of connection. Thus, the pattern of connectivity can be represented by a set of connectivity matrices, $\boldsymbol{W}$, one for each type of connection. It is cornmon, for example, to have two types of connections in a model: an inhibitory connection and an excitatory connection. When the models assume simple addition of inhibition and excitation they do not constitute different types of connections in our present

sense. They only constitute distinct types when they combine through some more complex rules. The pattern of connectivity is very important. Since it determines what each unit represents. Many of the issues concerning whether *top-down* or *bottom-up* processing systems are correct descriptions or whether a system is hierarchical and if so how many levels it has, etc., are all issues of the nature of the connectivity matrix. One important issue that may determine both how much information can be stored and how much serial processing the network must perform is the *fan-in* and *fan-out* of a unit. The *fan-in* is the number of elements that either excite or inhibit a given unit. The *fan- out* of a unit is the number of units affected directly by a unit. Note, in some cases we need more general patterns of connectivity. Specifying such a pattern in the general case is complex and will be addressed in a later section of this chapter.

5. *The rule of propagation.* A rule which takes the output vector, $\boldsymbol{o}(t)$, representing the output values of the units and combines it with the connectivity matrices to produce a net input for each type of input into the unit has been presented. net_{ij} is the net input of type i to unit u_j. Whenever only one type of connectivity is involved, it is suppressed the first subscript and use net_j to mean the net input into unit u_j. In vector notation $\boldsymbol{net_i}(t)$ represents the net input vector for inputs of type i. The propagation rule is generally straightforward. For example, there are two types of connections, inhibitory and excitatory, the net excitatory input is usually the weighted sum of the excitatory inputs to the unit. This is given by the vector product $\boldsymbol{net_e} = \boldsymbol{W_e o}(t)$. Similarly, the net inhibitory effect can be written as $\boldsymbol{net_i} = \boldsymbol{W_i o}(t)$. When more complex patterns of connectivity are involved, more complex rules of propagation are required.

6. *Activation rule.* Also a rule whereby the net inputs of each type impinging on a particular unit are combined with one another and with the current state of the unit to produce a new state of activation is presented. A function, $\boldsymbol{F}$, which takes $\boldsymbol{a}(t)$ and the vectors $\boldsymbol{net_j}$ for each different type of connection and produces a new state of activation is given. In the simplest cases, when $\boldsymbol{F}$ is the identity function and when all connections are of the same type, it can write $\boldsymbol{a}(t+1) = \boldsymbol{W o}(t)\boldsymbol{net}(t)$. Sometimes $\boldsymbol{F}$ is a threshold function so that the net input must exceed some value before contributing to the new state of activation. Often, the new state of activation depends on the old one as well as the current input. In general, however, there is $\boldsymbol{a}(t+1) = \boldsymbol{F}(\boldsymbol{a}(t), \boldsymbol{net}(t)_i, \boldsymbol{net}(t)_2, ...)$; the function $\boldsymbol{F}$ itself is what can be called the *activation rule.* Usually, the

function is assumed to be deterministic. Thus, for example, if a threshold is involved it may be that $a_i(t) = 1$ if the total input exceeds some threshold value and equals 0 otherwise. Other times it is assumed that $\boldsymbol{F}$ is stochastic. Sometimes activations are assumed to decay slowly with time so that even with no external input the activation of a unit will simply decay and not go directly to zero. Whenever $a_i(t)$ is assumed to take on continuous values it is common to assume that $\boldsymbol{F}$ is a kind of sigmoid function. In this case, an individual unit can *saturate* and reach a minimum or maximum value of activation. Perhaps the most common class of activations functions is the *quasi- linear* activation function. In this case the activation function, $\boldsymbol{F}$, is a nondecreasing function of a single *type* of input. In short, $a_i(t+1) = \boldsymbol{F}(net_i(t)) = F\left(\sum_j w_{ij} o_j\right)$. It is sometimes useful to add the constraint that $\boldsymbol{F}$ be a differentiable function. A differentiable quasi-linear activation functions can be considered as semilinear functions .

7. *Modifying patterns of connectivity as a function of experience.* Changing the processing or knowledge structure in a parallel distributed processing model involves modifying the patterns of interconnectivity. In principle this can involve three kinds of modifications: 1) the development of new connections. 2) the loss of existing connections; 3) the modification of the strengths of connections that already exist. Very little work has been done on (1) and (2) above. To a first order of approximation, however, (1) and (2) can be considered a special case of (3). Whenever the strength of connection away from zero to some positive or negative value is changed, it has the same effect as growing a new connection. Whenever the strength of a connection to zero, that has the same effect as losing an existing connectionis changed. Thus, in this section the rules whereby strengths of connections are modified through experience have been taken into account. Virtually all learning rules for models of this type can be considered a variant of the Hebbian learning rule suggested by Hebb (1949). Hebb's basic idea is this: If a unit, u_i, receives a input from another unit, u_j ; then, if both are highly active, the weight, w_{ij} , from u_j to u_i should be *strengthened*. This idea has been extended and modified so that it can be more generally stated as $\Delta w_{ij} = [g(a_i(t),\, t_i(t))\, h\,(o_j(t), w_{ij})]$, where $t_i(t)$ is a kind of teaching input to u_i. Simply stated, this equation says that the change in the connection from u_j, to u_i, is given by the product of a function, $g()$, of the activation of u_i, and its teaching input t_i, and another function, $h()$, of the output value of u_j and

the connection strength w_{ij}. In the simplest versions of Hebbian learning there is no teacher and the functions g and h are simply proportional to their first arguments. Thus there is $\Delta w_{ij} = \eta \alpha_i o_j$, where η is the constant of proportionality representing the learning rate. Another common variation is a rule in which $h(o_j(t), w_{ij}) = o_j(t)$ and $g(a_i(t), t_i(t))1) = \eta(t_i(t)—a_i(t))$. This is often called the Widrow-Hoff rule (Sutton & Barto, 1981). However, the *delta rule* is called in this manner because the amount of learning is proportional to the *difference* (or delta) between the actual activation achieved and the target activation provided by a teacher. In this case we have $\Delta w_{ij} = \eta \left(t_i(t) - a_i(t)\right) o_j(t)$. This is a generalization of the *perceptron* learning rule for which the famous *perception convergence theorem* has been proved. Still another variation has $\Delta w_{ij} = \eta a_i(t) \left(o_j(t) - w_{ij}\right)$.This is a rule employed by Grossberg (1976). There are many variations on this generalized rule, and we will describe some of them in more detail when we discuss various specific models below.

8. *Representation of the environment.* It is crucial in the development of any model to have a clear model of the environment in which this model exists. In PDP models, we represent the environment as a time-varying stochastic function over the space of input patterns. That is, in any point in time, there is some probability that any of the possible set of input patterns is impinging on the input units. This probability function may in general depend on the history of inputs to the system as well as outputs of the system. In practice, most PDP models involve a much simpler characterization of the environment. Typically, the environment is characterized by a stable probability distribution over the set of possible input patterns independent of past inputs and past responses of the system. In this case, the set of possible inputs to the system and numbering them from 1 to M can be listed. The environment is then characterized by a set of probabilities, p_i for $i = 1, ..., M$. Since each input pattern can be considered a vector, it is sometimes useful to characterize those patterns with nonzero probabilities as constituting *orthogonal* or *linearly independent* sets of vectors. Certain PDP models are restricted in the kinds of patterns they are able to learn: some being able to learn to respond correctly only if the input vectors form an orthogonal set; others if they form a linearly independent set of vectors; and still others are able to learn to respond to essentially arbitrary patterns of inputs.

2.5.3.2 The perceptron

Single-layer networks, with threshold activation function, were studied by Rosenblatt (1962) who called them *perceptrons*. Rosenblatt also built hardware implementation of these networks, which incorporated learning using an algorithm to be discussed below. Let's start with Rosenblatt's perceptron learning algorithm as the foundation of the area of neural networks. Consider a set of data shown in Figure 2.12. This is a 2-dimensional data set in which 2 variables, x_1 and x_2 have been measured for each observation. Each observation is a member of one of two mutually exclusive classes labelled "$\times$" and "$+$" in the figure. To apply the perceptron algorithm it is necessary to have a numeric code for each of the classes. It is possible use:

$$y = \begin{cases} 1 & for\, class\ \times\ and \\ -1 & for\, class + \end{cases} \tag{2.59}$$

The model then consists of a function f, called an "activation function," such that:

$$f = \begin{cases} 1 & w_0 + w^\tau x \geq 0 \\ -1 & otherwise \end{cases} \tag{2.60}$$

The perceptron learning algorithm tries to minimize the distance of a misclassified point to the straight line $w_0 + w_1 x_1 + w_2 x_2$. This line forms the "decision boundary" in that points are classified as belonging to class "$\times$" or "$+$" depending on which side of the line they are on. The misclassification rate is the proportion of observations that are misclassified (Figure 2.12 it is 1/9). Two classes that can be separated with 0 misclassification error are termed "linearly separable." The algorithm uses a cyclic procedure to adjust the estimates of the w parameters. Each point x is visited in turn and the w are updated by

$$w_i \leftarrow w_i + \eta \left[y - f\left(w_0 + w^\tau x \right) \right] x \tag{2.61}$$

This means that only incorrectly classified points move the decision boundary. The term η has to be set in advance and determines the step size. This is generally set to a small value in order to try to prevent overshooting the mark.

Where the classes are linearly separable it can be shown that the algorithm converges to a separating hyperplane in a finite number of steps. Where the data are not linearly separable, the algorithm will not converge and will eventually cycle through the same values. If the period of the cycle is large this may be hard to detect.

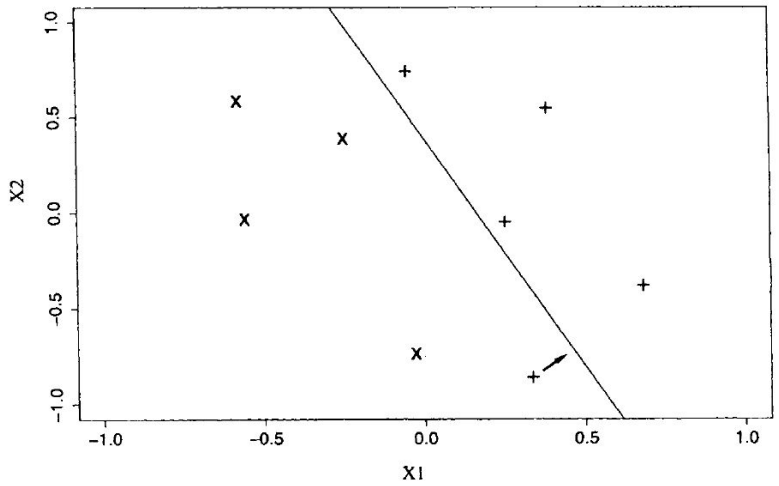

Figure 2.12: A 2-dimensional data consisting points in two classes, labelled "×" and "+". A perceptron decision boundary $w_0 + w_1 x_1 + w_2 x_2$ is also shown. One point is misclassified, that is, it is on the wrong side of the decision boundary. The margin of its misclassification is indicated by an arrow. On the next iteration of the perceptron fitting algorithm (1.1) the decision boundary will move to correct the classification that point. Whether it changes the classification in one iteration depends on the value of η, the step size parameter.

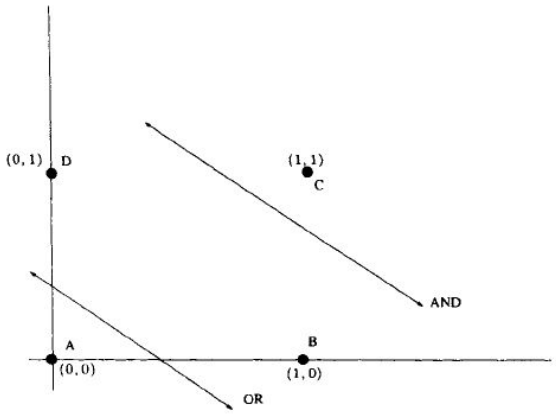

Figure 2.13: The problem of learning a logical function can be recast as a geometric problem by encoding $\{TRUE, FALSE\}$as $\{1, 0\}$. The figure shows decision boundaries that implement the function OR and AND. The XOR function would have to have points A and C on one side of the line and B and D on the other. It is clear that no single line can achieve that, although a set of lines defining a region or a non-linear boundary can achieve it.

Where then is the connection with brains and neurons? It lies in the fact that the algorithm can be represented in the form shown in Figure 2.9 where a processing node (the "neuron") receives a number of weighted inputs, forms their sum, and gives an output that is a function of this sum.

Interest in the perceptron as a computational model flagged when Minsky and Papert (1969) showed that it was not capable of learning some simple functions. Consider two logical variables A and B that take values in the set {TRUE, FALSE}. Now consider the truth values of the logical functions AND, OR, and XOR (exclusive OR, which is true if and only if one of its arguments is true) shown in Table 1.2. The problem of learning a logical function a geometric problem by encoding {TRUE, FALSE} as {l , 0} can be recasted. Now for the XOR function, in order to get a 0 classification error, the perceptron would have to put the points {1 , 1} and {0, 0} on one side of a line and {l, 0} and {0, 1} on the other. Clearly this is not possible (see Figure 2.13).

From these considerations, it is evident that the neural approach and the traditional statistics approach are different each others.

Linear discriminant analysis is the classical statistical technique for classification. It can not achieve a zero error on the geometric XOR problem any more than the perceptron can. Using a layered structure of perceptron shown in Figure2.14 over-

came this problem. These are the "multi-layer perceptrons" (MLPs). They required a different learning algorithm to the single perceptron and require that f be a differentiable function. It was the development of such algorithms that was the first step in their use. This has appeared several times in the literature, common early references being Werbos (1974) and Rumelhart et al. (1986).

A MLP is a feedforward neural network consisting of a number of units (*perceptrons* or *neurons*) which are connected by weighted links. We introduce the notation to be used in the following and describe the MLP model.

Given a set of training data, $D = \{x_n, t_n\}_{n=1}^{N}$ where each $x_n \in R^P$is a feature vector of length P, and $g_n \in \{1, ..., Q\}$ is the corresponding class label. A data matrix $\boldsymbol{X}$ containing the training data is:

$$X = \begin{pmatrix} x_1^T \\ \vdots \\ x_N^T \end{pmatrix} \tag{2.62}$$

where $x_n^T = (1, x_{n1}, ..., x_{nP})$. Hence the data matrix consists of N observations, each of length $P + 1$, and each containing a feature vector augmented by the addition of a 1 in the first coordinate position. The 1 is multiplied by a parameter known as "bias," which is equivalent to the "intercept" or β_0 term in a linear regression. Comparing Figure 2.14 and Figure 2.9, we can see that the essential structure of the MLP is that of perceptrons interconnected in layers, with the outputs from one layer forming the inputs to the next layer.

The first stage in constructing the model is to form the product which relates the input layer to the hidden layer,

$$\underset{H\times 1}{y} = \underset{H\times(P+1)(P+1)\times 1}{\Omega} \; x \tag{2.63}$$

where Ω is the first layer of adjustable parameters or weights in the MLP model. Then the function f_H (termed the activation function of the network) is applied to each element of y. In the diagrammatic representation of the MLP (Figure 2.14), this is applied separately at each unit in the hidden and output layers (the circles represent the computational units where this is done).

However, in the algebraic representation, we consider f_H a mapping from R^H to $(0, 1)^H$ that applies the same one-variable function, f_1 to each of its coordinates.

The resulting vector $f_H(y)$ s then augmented by the addition of a 1 so that

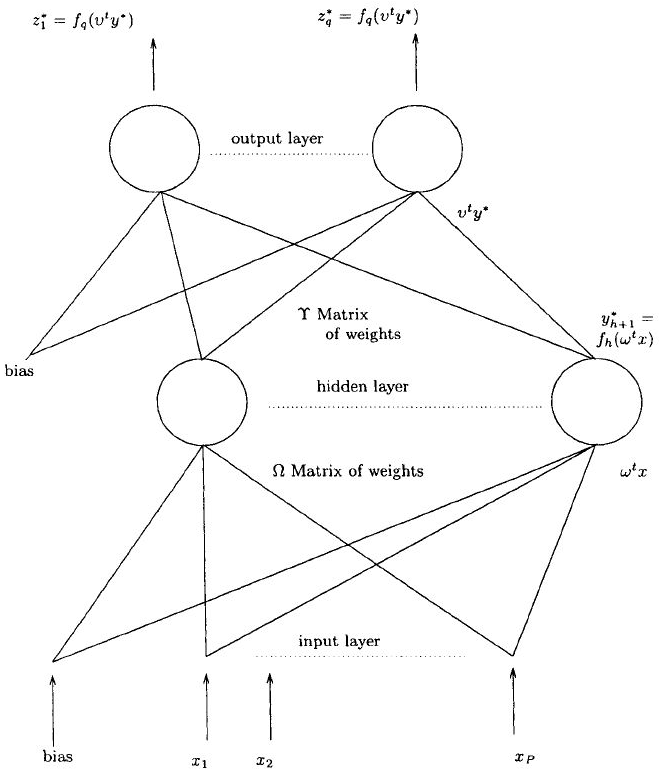

Figure 2.14: In the multi-layer pereeptron model, each operational unit, represented in the figure by a circle with a number of input lines, is a perceptron. The outputs of the perceptrons on one layer form the inputs to the perceptrons on the next layer. Each component of the input pattern is presented separately at the nodes of the input layer; the components are weighted, summed, and passed through the perceptron activation function. The outputs of the first, or hidden, layer form the inputs to the nodes of the output layer. While the figure has only two layers of perceptrons, it is of course possible to have an arbitrary number of such layers.

$$y^* = \left[1, \{f_H(y_1, ..., y_H)\}^T\right]^T \tag{2.64}$$

(just as the original data vector x was augmented). This now forms the data input to the next and, in this case final, layer of the network. So we have

$$\underset{Q\times 1}{z} = \underset{Q\times(H+1)}{Y}\ \underset{(H+1)\times 1}{y^*} \tag{2.65}$$

and

$$mlp(x) = z^* = f_Q\left(Y\left[1, \{f_H(\Omega x)\}^T\right]^T\right) \tag{2.66}$$

The MLP described here has two layers of adjustable weights and no skip or intra-layer connections and can be described by the parameters P, H, Q that is, P input units, H units in the hidden layer and Q output units. This is a two-layer MLP, i.e., MLP with two layers of adjustable weights; however, the terminology is not standardized and some authors would call it a threelayer MLP (input, hidden and output) and some a one-layer MLP (one hidden layer). The weights are considered as going from the input to a hidden-layer unit, for example, as the weights "fanning into" the unit and the weights connecting the unit to the output units as the weights "fanning out" of the unit.

While other choices are possible, f_H here is chosen to be the logistic activation function,

$$f_H(x) = \frac{1}{1 + exp(-x)} \tag{2.67}$$

For general regression problems f_Q may be chosen as the identity function so that $z* = z$ and the output of the MLP is not restricted to the interval (0,l). As we restrict our attention to classification problems, $f_Q$will be either the logistic activation function like f_H, or the softmax activation function,

$$f_Q\left(z_q^*\right) = \frac{exp\left(z_q\right)}{\sum_{q_1=1}^{Q} exp(z_{q_1})} \tag{2.68}$$

A target matrix T of dimension $N \times Q$, consisting of a high value in position (n, q) if observation n is in class q, and a low value otherwiseis also formed. When the high and low values are 1 and 0, respectively, this is referred to as a one of Q encoding. The target vector, t_n, is the n^{th} row of T and is an encoding of the class label, g_n, a

target vector of length Q and t_{nq} is the $(n,q)^{th}$ element of the target matrix.

The units on each layer (input, hidden and output) are indicated with the indices p, h and q respectively, so that w_{ph} is the weight between the pth input and the p^{th} hidden-layer unit. When additional indices are required it can be used p_1, h_1, q_1 etc.

Having set up the MLP model with output $z*$, it is then choosen the weight matrices Y and Q to minimize the value of some penalty function. Two penalty functions here are considered: one is the sum of squares penalty function

$$\rho_l = \sum_{n=1}^{N}\sum_{q=1}^{Q}\frac{1}{2}\left(t_{nq} - z^*_{nq}\right)^2 \tag{2.69}$$

and the other is the cross-entropy penalty function

$$\rho_l = \sum_{n=1}^{N}\sum_{q=1}^{Q} t_{nq}\, log\left(\frac{t_{nq}}{z^*_{nq}}\right) \tag{2.70}$$

In the standard implementation, the weights are initially assigned random values selected uniformly in some small interval, often (- 1, l). While it is generally possible write $\omega \in \Omega$ and $\nu \in \Upsilon$ in some contexts, it is natural to consider ρ as a function of a vector of parameters. In this case it is possible write $\omega = \left\{vec\left(\Upsilon^T\right)^T, vec\left(\Omega^T\right)^T\right\}$ and index ω by $r = 1, ..., R$ where $R = (P+1)H + (H+1)Q$ is the total number of weights.

2.5.3.3 Training an MLP – Backpropagation.

The perhaps most straightforward way to design a training algorithm for the MLP is to use the gradient descent algorithm. The model output $\hat{y}$ have to be differentiable with respect to all the parameters ω_{jk} and ν_j. A training data set $X = x(n)$, $y(n)_{n=1,...,N}$ with N observations is defined, and all the weights in the network by $\boldsymbol{W} = \omega_j, \nu$ are denoted.

The batch form of gradient descent then goes as follows:

1. Initialize $\boldsymbol{W}$ with e.g. small random values.
2. Repeat until convergence (either when the error E is below some preset value or until the gradient ∇wE is smaller than a preset value), t is the iteration number
   (a) compute the update $\Delta\boldsymbol{W} = -\eta\nabla w\, E\,(t))\eta = \eta\sum_{n=1}^{N} e\,(n,t)\,\nabla w\,\hat{y}\,(n,t)$ where $e\,(n,t) = (y\,(n) - \hat{y}\,(n,t))$

(b) update the weights $W(t+1) = W(t) + W(t)$

(c) compute the error$E(t+1)$

As an example, the weight updates for the special case of a multilayer perceptron with one hidden layer, using the transfer function $\phi(z)$ (e.g. $tanh(z)$), and one output unit with the transfer function $\theta(z)$ (e.g. logistic or linear) have been computed. It can be used half the mean square error

$$E = \frac{1}{2N}\sum_{n=1}^{N}\left[y\left(n\right) - \hat{y}\left(n\right)\right]^2 = \frac{1}{2N}\sum_{n=1}^{N} e^2\left(n\right) \tag{2.71}$$

and the following notation

$$\hat{y}(x) = \theta[a(\boldsymbol{x})] \tag{2.72}$$

$$a(x) = \nu_0 + \sum_{j=1}^{M} \nu_j h_j(\boldsymbol{x}) \tag{2.73}$$

$$h_j(x) = \phi[b_j(\boldsymbol{x})] \tag{2.74}$$

$$bj(x) = \sum_{k=1}^{D} w_{jk} x_{jk} \tag{2.75}$$

Here, ν_j are the weights between the hidden layer and the output layer, and w_{jk} are the weights between the input and the hidden layer.

For weight ν_i is given

$$\frac{\partial E}{\partial \nu_i} = -\frac{1}{N}\sum_{n=1}^{N} e(n)\frac{\partial \hat{y}}{\partial \nu_i} = \frac{1}{N}\, e(n)\, \theta\left[a\left(n\right)\right] \frac{\partial a_n}{\partial \nu_i} = -\frac{1}{N}\sum_{n=1}^{N} e(n)\, \theta'\left[a\left(n\right)\right]\, h_i\left(n\right) \tag{2.76}$$

$$\Rightarrow \Delta\nu_i = -\eta\frac{\partial E}{\partial \nu_i} = \eta\frac{1}{N}\sum_{n=1}^{N} e(n)\, \theta'\left[a\left(n\right)\right]\, h_i\left(n\right) \tag{2.77}$$

with the definition $h_0\left(n\right) \equiv 1$. If the output transfer function is linear, i.e. $\theta\left(z\right) = z$, then $\theta'\left(z\right) = 1$. If the output transfer function is logistic, i.e. $\theta\left(z\right) = \left[1 + exp(-z)\right]^{-1}$, then $\theta'\left(z\right) = \theta\left(z\right)\left[1 - \theta\left(z\right)\right]$.

For weight w_{il} is given

$$\begin{gathered}\frac{\partial E}{\partial w_{il}} = -\frac{1}{N}\sum_{n=1}^{N} e(n)\frac{\partial \hat{y}(n)}{\partial w_{il}} = -\frac{1}{N}\sum_{n=1}^{N} e(n)\,\theta'\left[a\left(n\right)\right]\frac{\partial a(n)}{\partial w_{il}} = \\ = -\frac{1}{N}\sum_{n=1}^{N} e(n)\,\theta'\left[a\left(n\right)\right]\,\nu_i\frac{\partial h_i(n)}{\partial w_{il}} = -\frac{1}{N}\sum_{n=1}^{N} e(n)\,\theta'\left[a\left(n\right)\right]\nu_i\phi'\left[b_i\left(n\right)\right]\frac{\partial b_i(n)}{\partial w_{il}} = \\ = -\frac{1}{N}\sum_{n=1}^{N} e(n)\,\theta'\left[a\left(n\right)\right]\nu_i\phi'\left[b_i\left(n\right)\right]x_l\end{gathered} \tag{2.78}$$

$$\Rightarrow \Delta w_{il} = -\eta\frac{\partial E}{\partial w_{il}} = \eta\frac{1}{N}\sum_{n=1}^{N} e(n)\,\theta'\left[a\left(n\right)\right]\nu_i\phi'\left[b_i\left(n\right)\right]\,x_l\left(n\right) \tag{2.79}$$

with definition $x_0\left(n\right) \equiv 1$. If hidden unit transfer function is hyperbolic tangent function, i.e. $\phi\left(z\right) = tanh\left(z\right)$,then $\phi'\left(z\right) = 1 - \phi^2\left(z\right)$.

This gradient descent method for updating the weights has become known as the "backpropagation" training algorithm. The motivation for the name becomes clear if this notation has introduced

$$\delta\left(n\right) = e\left(n\right)\theta'\left[a\left(n\right)\right] \tag{2.80}$$

$$\delta_i\left(n\right) = \delta\left(n\right)\nu_i\phi'\left[b\left(n\right)\right] \tag{2.81}$$

and it can be write

$$\Delta\nu_i = \eta\,\frac{1}{N}\sum_{n=1}^{N}\delta\left(n\right)h_i\left(n\right) \tag{2.82}$$

$$\Delta w_{il} = \eta\,\frac{1}{N}\sum_{n=1}^{N}\delta_i\left(n\right)x_l\left(n\right) \tag{2.83}$$

which is very similar to the old LMS algorithm. Expression 2.81 corresponds to a propagation of $\delta\left(n\right)$ backwards through the network.

The gradient descent learning algorithm corresponds to backprop in its *batch* form, where the update is computed using all the available training data. There is also an "*on-line*" version where the updates are done after each pattern $\boldsymbol{x}(n)$ without averaging over all patterns.

Bishop (1995) and Haykin (1999) discuss variants of backprop, e.g. using a "momentum" term and adaptive learning rate . Backprop with "momentum" amounts to using an update equal to

$$\Delta \boldsymbol{W}(t) = -\eta \nabla w E(t) + \alpha \Delta \boldsymbol{W}(t-1) \tag{2.84}$$

where $0 < \alpha < 1$. This enables the learning to gain momentum (hence the name). If there are several updates in the same direction, then the effective learning rate is increased. If there are several updates in opposite directions then the effective learning rate is decreased.

Backpropagation is, in general, a very slow learning algorithm (even with momentum) and there are many better algorithms which will be discussed below. However, backpropagation was very important in the beginning of the 1980, because it was used to demonstrate that multilayer perceptrons can learn things.

2.5.3.4 Resilient backropagation (RPROP)

A very useful gradient based learning algorithm, that is not discussed in Bishop (Bishop 1995), is the "resilient backropagation" (RPROP) algorithm (Riedmiller & Braun 1993). It uses individual adaptive learning rates combined with the so-called "Manhattan" update step. The standard backpropagation updates the weights according to

$$\Delta w_{il} = -\eta \frac{\partial E}{\partial w_{il}} \tag{2.85}$$

The "Manhattan" update step, on the other hand, uses only the sign of the derivative (the reason for the name should be obvious to anyone who has seen a map of Manhattan), i.e.

$$\Delta w_{il} = -\eta \, sign \left[\frac{\partial E}{\partial w_{il}} \right] \tag{2.86}$$

The RPROP algorithm combines this Manhattan step with individual learning rates for each weight, and the algorithm goes as follows

$$\Delta w_{il}(t) = -\eta_{il}(t) \; sign \left[\frac{\partial E}{\partial w_{il}} \right] \tag{2.87}$$

where w_{il} denotes any weight in the network (e.g. also hidden to output weights).

The learning rate $\eta_{il}(t)$ is adjusted according to

$$\eta_{il}(t) = \begin{cases} \gamma^{+}\eta_{il}(t-1) & if\ \partial_{il}E(t)\cdot\partial_{il}E(t-1) > 0 \\ \gamma^{-}\eta_{il}(t-1) & if\ \partial_{il}E(t)\cdot\partial_{il}E(t-1) < 0 \end{cases} \tag{2.88}$$

whereγ^{+} and $\gamma^{-}$are different growth/shrinking factors ($0 < \gamma^{-} < 1 < \gamma^{+}$). The RPROP algorithm is a *batch*algorithm, since the learning rate update becomes noisy and uncertain if the error $E$is evaluated over only a single pattern.

2.5.3.5Second order learning algorithms

Backpropagation, i.e. gradient descent, is a *first order* learning algorithm. This means that it only uses information about the first order derivative when it minimizes the error. The idea behind a first order algorithm can be illustrated by expanding the error E in a Taylor series around the current weight position $\boldsymbol{W}$

$$E(\boldsymbol{W}+\Delta\boldsymbol{W}) = E(\boldsymbol{W}) + \nabla w\, E(\boldsymbol{W})^{T}\Delta\boldsymbol{W} + \mathcal{O}\left(||\Delta\boldsymbol{W}||^{2}\right) \tag{2.89}$$

The vector $\boldsymbol{W}$ contains all the weights w_{jk} and ν_j (and others if other network architectures is consided) and $\Delta\boldsymbol{W}$ have to be small.

The notation $\mathcal{O}\left(||\Delta\boldsymbol{W}||^{2}\right)$ denotes all the terms that contains the small weight step $\Delta\boldsymbol{W}$ multiplied by itself at least once, and by "small" it means that $\Delta\boldsymbol{W}$ is so small that the gradient term $\nabla w\, E(\boldsymbol{W})\Delta\boldsymbol{W}$ is larger than the sum of the higher order terms. In that case the higher order terms can be ignored and it is possible write:

$$E(\boldsymbol{W}+\Delta\boldsymbol{W}) \approx E(\boldsymbol{W}) + \nabla w\, E(\boldsymbol{W})^{T}\Delta\boldsymbol{W} \tag{2.90}$$

Now, the weights have been changed so that the new error $E(\boldsymbol{W}+\Delta\boldsymbol{W})$ is smaller than the current error $E(\boldsymbol{W})$. One way to guarantee this is to set the weight update $\Delta\boldsymbol{W}$ proportional to the negative gradient, i.e. $\Delta\boldsymbol{W} = -\eta\nabla w\, E(\boldsymbol{W})$, in which case

$$E(\boldsymbol{W}+\Delta\boldsymbol{W}) \approx E(\boldsymbol{W}) - \eta\,||\nabla w\, E(\boldsymbol{W})||^{2} \leq E(\boldsymbol{W}) \tag{2.91}$$

However, this of course requires that $\Delta\boldsymbol{W}$ is so small that we can motivate (2.90).

This and also consider the second order term can be extended in the Taylor expansion. That is

$$E(\boldsymbol{W}+\Delta\boldsymbol{W}) = E(\boldsymbol{W}) + \nabla w E(\boldsymbol{W})^T \Delta \boldsymbol{W} + \frac{1}{2}\Delta\boldsymbol{W}^T H(\boldsymbol{W})\Delta\boldsymbol{W} + \mathcal{O}\left(||\Delta\boldsymbol{W}||^3\right) \tag{2.92}$$

where

$$H(\boldsymbol{W}) = \nabla w \nabla w^T E(\boldsymbol{W}) \tag{2.93}$$

is the Hessian matrix with elements $H_{ij}(\boldsymbol{W}) = \frac{\partial^2 E(\boldsymbol{W})}{\partial w_i \partial w_j}$. The Hessian is symmetric (all eigenvalues are consequently real and H with an orthogonal transformation can be diagonalized).

If the higher order terms in equation2.92 can be ignored, then there will be

$$E(\boldsymbol{W}+\Delta\boldsymbol{W}) \approx E(\boldsymbol{W}) + \nabla w E(\boldsymbol{W})^T \Delta \boldsymbol{W} + \frac{1}{2}\Delta\boldsymbol{W}^T H(\boldsymbol{W})\Delta\boldsymbol{W} \tag{2.94}$$

It is possible change the weights so that the new error $E(\boldsymbol{W}+\Delta\boldsymbol{W})$ is smaller than the current error $E(\boldsymbol{W})$. Furthermore, it has to be as small as possible. That is, $E(\boldsymbol{W}+\Delta\boldsymbol{W})$ can be minimized by choosing $\Delta\boldsymbol{W}$ appropriately. The requirement that extremum point is reached

$$\begin{aligned} &\nabla w\, E(\boldsymbol{W}+\Delta\boldsymbol{W}) = 0 \\ \Rightarrow\ &\nabla w\, E(\boldsymbol{W}) + H(\boldsymbol{W})\Delta\boldsymbol{W} \end{aligned} \tag{2.95}$$

which yields the optimum weight update as

$$\Delta\boldsymbol{W} = H^{-1}(\boldsymbol{W})\ \nabla w\, E(\boldsymbol{W}) \tag{2.96}$$

To guarantee that this is a mininum point it is necessary also require that the Hessian matrix is positive definite. This means that all the eigenvalues of the Hessian matrix must pe positive. If any of the eigenvalues are zero then there will be saddle point and $H(\boldsymbol{W})$ is not invertible. If any of the eigenvalues of $H(\boldsymbol{W})$ are negative then there will be a maximum point for at least one of the weights w_j and (2.96) will actually move away from the minimum.

The update step (7.32) is usually referred to as a "Newton-step", and the minimization method that uses this update step is the Newton algorithm. Some problems with "vanilla" Newton learning (7.32) are:

- The Hessian matrix may not be invertible, i.e. some of the eigenvalues are zero.

- The Hessian matrix may have negative eigenvalues.
- The Hessian matrix is expensive to compute and also expensive to invert. The learning may therefore be slower than a first order method.

The first two problems are handled by regularizing the Hessian, i.e. by replacing $H(\boldsymbol{W})$ by $H(\boldsymbol{W}) + \lambda \boldsymbol{I}$. This effectively filters out all eigenvalues that are smaller than λ. The third problem is handled by "Quasi-Newton" methods that iteratively try to estimate the inverse Hessian using expressions of the form $H^{-1}(\boldsymbol{W} + \Delta \boldsymbol{W}) \approx H^{-1}(\boldsymbol{W})+$ correction.

2.5.3.6 The Levenberg-Marquardt algorithm

The Levenberg-Marquardt, see e.g. (Bishop 1995), is a very efficient second order learning algorithm that builds on the assumption that the error E is a quadratic error (which it usually is), like half the mean square error. In this case there will be

$$H_{ij} = \frac{1}{2}\frac{\partial^2 MSE}{\partial w_i \partial w_j} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial \hat{y}(n)}{\partial w_i}\frac{\partial \hat{y}(n)}{\partial w_j} + \frac{1}{N}\sum_{n=1}^{N} e(n)\frac{\partial^2 \hat{y}(n)}{\partial w_i \partial w_j} \tag{2.97}$$

If the residual $e(n)$ is symmetrically distributed around zero and small then it is possible assume that the second term in equation 2.97 is very small compared to the first term. If so, then it is possible approximate

$$H_{ij} \approx \frac{1}{N}\sum_{n=1}^{N}\frac{\partial \hat{y}(n)}{\partial w_i}\frac{\partial \hat{y}(n)}{\partial w_j} \tag{2.98}$$

$$H(\boldsymbol{W}) \approx \frac{1}{N}\sum_{n=1}^{N}\boldsymbol{J}(n)\boldsymbol{J}^T(n) \tag{2.99}$$

where this notation can be used

$$\boldsymbol{J}(n) = \nabla w \hat{y}(n) \tag{2.100}$$

and $\boldsymbol{J}$ is the "Jacobian". This approximation is not as costly to compute as the exact Hessian, since no second order derivatives are needed. The fact that the Hessian is approximated by a sum of outer products JJ^T means that the rank of H is at most N. That is, there must be at least as many observations as there are weights in the network.

This approximation of the Hessian is used in a Newton step together with a regularization term, so that the Levenberg-Marquardt update is

$$\Delta \boldsymbol{W} = \left[\frac{1}{N} \sum_{n=1}^{N} \boldsymbol{J}(n) \boldsymbol{J}^T(n) + \lambda \boldsymbol{I} \right]^{-1} \nabla w E(\boldsymbol{W}) \tag{2.101}$$

The Levenberg-Marquardt update is a very useful learning algorithm, and it represents a combination of gradient descent and a Newton step search.There will be

$$\Delta \boldsymbol{W} \rightarrow \begin{cases} \frac{1}{\lambda} \nabla E(\boldsymbol{W}) & when\, \lambda \rightarrow \infty \\ \left[\frac{1}{N} \sum_{n=1}^{N} \boldsymbol{J}(n) \boldsymbol{J}^T(n) \right]^{-1} & when\, \lambda \rightarrow 0 \end{cases} \tag{2.102}$$

which corresponds to gradient descent, with $\eta = \frac{1}{\lambda}$, when is large, and to Newton learning when λ is small.

2.6 Geostatistic methods

2.6.1 kriging

Kriging has for many decades been used as a synonym for geostatistical interpolation. It originated in the mining industry in the early 1950's as a means of improving ore reserve estimation. The original idea came from the mining engineers D. G. Krige and the statistician H. S. Sichel. The technique was first published in Krige (1951), but it took almost a decade until a French mathematician G. Matheron derived the formulas and basically established the whole field of linear geostatistics (Cressie, 1990; Webster and Oliver, 2001; Zhou et al., 2007).

A standard version of kriging is called *ordinary kriging* (OK). Here the predictions are based on the model:

$$Z(\boldsymbol{x}) = \mu + \varepsilon'(\boldsymbol{x}) \tag{2.103}$$

where μ is the constant stationary function (global mean) and $\varepsilon'(\boldsymbol{x})$ is the spatially correlated stochastic part of variation. The predictions are made with following formula:

$$\hat{z}_{OK}(\boldsymbol{x}_0) = \sum_{i=1}^{N} \lambda_i(\boldsymbol{x_o}) \cdot z(\boldsymbol{x_i}) = \boldsymbol{\lambda_0^T} \cdot \boldsymbol{z} \tag{2.104}$$

where $\boldsymbol{\lambda}_0$ is the vector of kriging weights (w_i), $\boldsymbol{z}$ is the vector of n observations at primary locations. In a way, kriging can be seen as a sophistication of the inverse distance interpolation. For inverse distance weighted interpolation the key problem is to determine how much importance should be given to each neighbour. Intuitively thinking, there should be a way to estimate the weights in an objective way, so the weights reflect the true spatial autocorrelation structure. The novelty that Matheron (1962) and Gandin (1963) introduced to the analysis of point data is the derivation and plotting of the so-called semivariances, differences between the neighbouring values:

$$\gamma(\boldsymbol{h}) = \frac{1}{2} E\left[(z(\boldsymbol{x_i}) - z(\boldsymbol{x_i} + h))^2\right] \tag{2.105}$$

where $z(\boldsymbol{x_i})$ is the value of target variable at some sampled location and $z(\boldsymbol{x}_i + \boldsymbol{h})$ is the value of the neighbour at distance $\boldsymbol{x_i} + \boldsymbol{h}$.

Suppose that there are n point observations, this yields $n \cdot (n-1)/2$ pairs for which a semivariance can be calculated. It is possible plot all semivariances versus their distances, which will produce a variogram cloud as shown in Fig. 2.15b. Such clouds are not easy to describe visually, so the values are commonly averaged for standard distance called the lag. If such averaged data are displayed, then it is possible get a standard experimental variogram as shown in Fig.2.15c. It is usually see that semivariances are smaller at shorter distance and then they stabilize at some distance. This can be interpreted as follows: the values of a target variable are more similar at shorter distance, up to a certain distance where the differences between the pairs are more less equal the global variance. This is known as the spatial auto-correlation effect. Once an experimental variogram has been calculated, it is possible fit it using some of the authorized variogram models, such as *linear*, *spherical*, *exponential*, *circular*, *Gaussian*, *Bessel*, *power* and similar (Isaaks and Srivastava, 1989; Goovaerts, 1997). The variograms are commonly fitted by iterative reweighted least squares estimation, where the weights are determined based on the number of point pairs or based on the distance. Most commonly, the weights are determined using N_j/h_j^2 , where N_j is the number of pairs at certain lag, and $\boldsymbol{h}_j$ is the distance (Fig.2.15d). This means that the algorithm will give much more importance to semivariances with large number of point pairs and to the shorter distances. Fig.2.15d shows the result of automated variogram fitting given an experimental variogram (Fig.2.15c) and using the N_j/h_j^2 - weights: in this case, an exponential model with the nugget parameter = 26, sill parameter = 440, and the range parameter = 478 m has been obtained. Note that this is only a *sample variogram*, if it would go and collect several point samples,

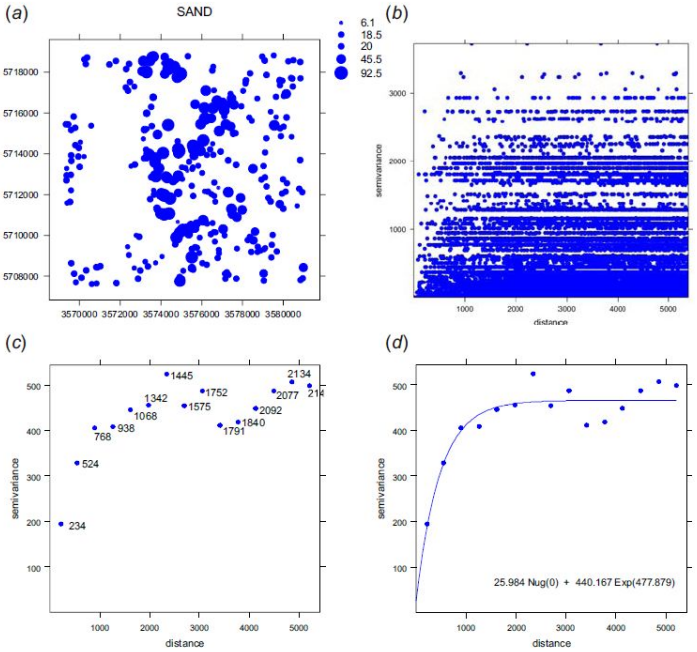

Figure 2.15: Steps of variogram modelling: (a) location of points (300), (b) variogram cloud showing semivariances for 44850 pairs, (c) semivariances aggregated to lags of about 300 m, and (d) the final variogram model fitted using the default settings in gstat.

each would lead to somewhat different variogram plot. The target variable is said to be stationary if several sample variograms are very similar (constant), which is referred to as the covariance stationarity. Otherwise, if the variograms differ much locally and/or globally, then one can speak about a non-stationary inherent properties. In principle, assumptions of kriging are that the target variable is stationary and that it has a normal distribution, which is probably the biggest limitation of kriging. It is also important to note that there is a difference between the range factor and the range of spatial dependence, also known as the practical range. A practical range is the Lag $\boldsymbol{h}$ for which $\gamma(\boldsymbol{h}) = 0.95\,\gamma(\infty)$, i.e. that distance at which the semivariance is close to 95% of the sill (Fig. 2.16b).

Once the variogram model has been estimated, it is possible use it to derive semivariances at all locations and solve the kriging weights . The kriging OK weights are solved by multiplying the covariances:

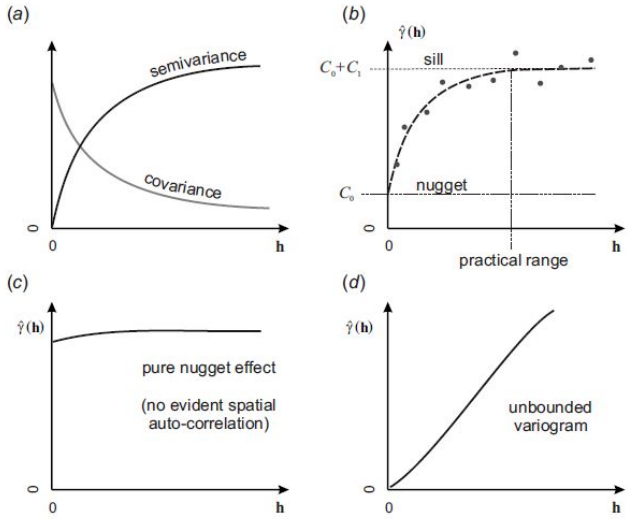

Figure 2.16: Some basic concepts of variograms: (a) the difference between semivariance and covariance; (b) it often important in geostatistics to distinguish between the sill variation (C_0+C_1) and the sill parameter (C_1) and between the range parameter (R) and the practical range; (c) a variogram that shows no spatial correlation can be defined by a single parameter (C_0); (d) an unbounded variogram typically leads to predictions similar to inverse distance interpolation.

$$\lambda_0 = \boldsymbol{C}^{-1} \cdot \boldsymbol{C_0} \qquad C\left(|\boldsymbol{h}| = 0\right) = C_0 + C_1 \tag{2.106}$$

where $\boldsymbol{C}$ is the covariance matrix derived for $n \times n$ observations and $\boldsymbol{C_0}$ is the vector of covariances at new location. Note that the C is in fact (n + 1) × (n + 1) matrix if it is used to derive kriging weights. One extra row and column are used to ensure that the sum of weights is equal to one:

$$\begin{bmatrix} C(\boldsymbol{x}_1, \boldsymbol{x}_1) & \cdots & C(\boldsymbol{x}_1, \boldsymbol{x}_n) & 1 \\ \vdots & \vdots & \vdots & \vdots \\ C(\boldsymbol{x}_n, \boldsymbol{x}_1) & \cdots & C(\boldsymbol{x}_n, \boldsymbol{x}_n) & 1 \\ 1 & \cdots & 1 & 0 \end{bmatrix}^{-1} \cdot \begin{bmatrix} C(\boldsymbol{x}_0, \boldsymbol{x}_1) \\ \vdots \\ C(\boldsymbol{x}_0, \boldsymbol{x}_n) \\ 1 \end{bmatrix} = \begin{bmatrix} \lambda_1(\boldsymbol{x}_0) \\ \vdots \\ \lambda_n(\boldsymbol{x}_0) \\ \mu \end{bmatrix} \tag{2.107}$$

where μ is the so-called Langrange multiplier.

In addition to estimation of values at new locations, a statistical spatial prediction technique offers a measure of associated uncertainty of making these estimations by using a given model. In geostatistics, this is often referred to as the prediction variance, i.e. the estimated variance of the prediction error. OK variance is defined as the weighted average of covariances from the new point ($\boldsymbol{x}_0$) to all calibration points ($x_1, ...x_n$), plus the Lagrange multiplier (Webster and Oliver, 2001):

$$\hat{\sigma}^2_{OK} = (C_0 + C_1) - \boldsymbol{C}_0^T \cdot \lambda_0 = C_0 + C_1 - \sum_{i=1}^{N} \lambda_i(\boldsymbol{x_o}) \cdot C(\boldsymbol{x_0}, \boldsymbol{x_i}) + \mu \tag{2.108}$$

where $C(\boldsymbol{x_0}, \boldsymbol{x_i})$ is the covariance between the new location and the sampled point pair, and μ is the Lagrange multiplier, as shown in Eq.(1.3.5).

As you can notice, outputs from any statistical prediction model are always two maps: (1) predictions and (2) prediction variance. The mean of the prediction variance at all location can be termed the overall prediction variance, and can be used as a measure of how precise is our final map: if the overall prediction variance gets close to the global variance, then the map is 100% imprecise; if the overall prediction variance tends to zero, then the map is 100% precise. Note that a common practice in geostatistics is to model the variogram using a semivariance function and then, for the reasons of computational efficiency, use the covariances. In the case of solving the kriging weights, both the matrix of semivariances and covariances give the same

results, so you should not really make a difference between the two. The relation between the covariances and semivariances is (Isaaks and Srivastava, 1989):

$$C(\boldsymbol{h}) = C_0 + C_1 - \gamma(\boldsymbol{h}) \tag{2.109}$$

where $C(\boldsymbol{h})$ is the covariance, and $\gamma(\boldsymbol{h})$ is the semivariance function (Fig. 2.16a).

So for example, exponential model can be written in two ways:

$$\begin{aligned} \gamma(\boldsymbol{h}) &= \begin{cases} 0 & if \quad |h| = 0 \\ C_0 + C_1 \cdot \left[1 - e^{-(\frac{h}{R})}\right] & if \quad |h| > 0 \end{cases} \\ C(\boldsymbol{h}) &= \begin{cases} C_0 + C_1 & if \quad |h| = 0 \\ C_1 \cdot \left[e^{-(\frac{h}{R})}\right] & if \quad |h| > 0 \end{cases} \end{aligned} \tag{2.110}$$

The covariance at zero distance ($C(0)$) is by definition equal to the mean residual error (Cressie, 1993); $C(\boldsymbol{h}_{11})$ also written as $C(\boldsymbol{x}_1, \boldsymbol{x}_1)$, and which is equal to

$$C(0) = C_0 + C_1 = Var\{z(x)\} \tag{2.111}$$

The variogram models can be extended to even larger number of parameters if either (a) anisotropy or (b) smoothness are considered in addition to modelling of nugget and sill variation. The experimental semivariogram is also a function of direction. In fact, spatial variation of considered variables might not be the same in all directions and it might be observed the presence of anisotropic patterns. This different behavior of the experimental variogram in different directions is called *anisotropic* behavior. As variogram models are defined for the isotropic case, it is need to examine transformations of the coordinates which allow to obtain anisotropic random functions from the isotropic models. In practice anisotropies are detected by inspecting experimental variograms in different directions and are included into the model by tuning predefined anisotropy parameters. In 2D-space a representation of the behavior of the experimental variogram can be made by drawing a map of iso-variogram lines as a function of a vector $\boldsymbol{h}$. Ideally if the iso-variogram lines are circular around the origin, the variogram obviously only depends on the length of the vector $\boldsymbol{h}$ and the phenomenon is isotropic. If not, the iso-variogram lines can in many applications be approximated by concentric ellipses defined along a set of perpendicular main axes of anisotropy.

Now it will try to better clarify this concept. Given a coordinate system for $\boldsymbol{h} = (h_1, h_2, ..., h_n)$ with n coordinates. In this coordinate system the surfaces of

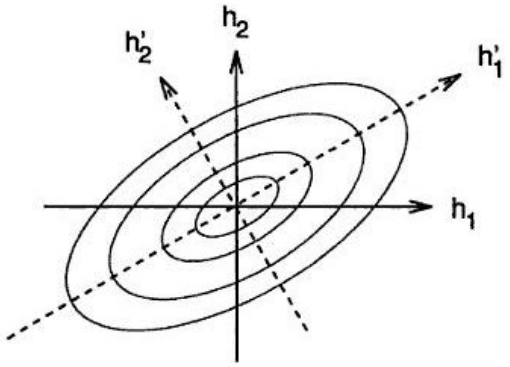

Figure 2.17: The coordinate system for $\boldsymbol{h} = (h_1, h_2)$ is rotated into system $\boldsymbol{h'}$parallel to the mai axes of the concentric ellipses.

constant variogram describe an ellipsoid and we search a new coordinate system for $\widetilde{\boldsymbol{h}}$ in which the iso-variogram lines are spherical. As a first step a rotation matrix $\boldsymbol{Q}$ is sought which rotates the coordinate system $\boldsymbol{h}$ into a coordinate system $\boldsymbol{h'} = \boldsymbol{Q} \cdot \boldsymbol{h}$ that is parallel to the principal axes of the ellipsoid, as shown on Figure in the 2D case. The directions of the principal axes should be known from experimental variogram calculations.

In 2D the rotation is given by the matrix

$$\boldsymbol{Q} = \begin{bmatrix} cos\theta & sin\theta \\ -sin\theta & cos\theta \end{bmatrix} \tag{2.112}$$

where $\theta$is the rotation angle. The second step in the transformation is to operate or dilatation of the principal axes of the ellipsoid using a diagonal matrix

$$\sqrt{\Lambda} = \begin{bmatrix} \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \sqrt{\lambda_n} \end{bmatrix} \tag{2.113}$$

which transforms the system $\boldsymbol{h'}$ into a new system $\widetilde{\boldsymbol{h}}$ in which the ellipsoids become spheres

$$\widetilde{\boldsymbol{h}} = \sqrt{\boldsymbol{\Lambda}}\boldsymbol{h'} \tag{2.114}$$

Conversely, if r is the radius of a sphere around the origin in the coordinate system

of the isotropic variogram, it is obtained by calculating the length of any vector $\widetilde{\boldsymbol{h}}$ pointing on the surface of the sphere

$$r = |\widetilde{\boldsymbol{h}}| = \sqrt{\widetilde{\boldsymbol{h}}^{\boldsymbol{T}}\widetilde{\boldsymbol{h}}} \tag{2.115}$$

This yelds the equation of an ellipsoid in the $\boldsymbol{h'}$coordinate system

$$(\boldsymbol{h'})^T \boldsymbol{\Lambda h'} = r^2 \tag{2.116}$$

The diameters d_p(principal axes) of the ellipsoid along the principal directions are thus

$$d_p = \frac{2r}{\sqrt{\lambda_p}} \tag{2.117}$$

and the principal directions are the vectors $\boldsymbol{q_p}$ of the rotation matrix.

Finally once the ellipsoid is determinaned the anisotropic variogram is specified on the basis of an isotropic variogram by

$$\gamma(r) = \gamma\left(\sqrt{\boldsymbol{h}^{\boldsymbol{T}}\boldsymbol{B}\boldsymbol{h}}\right) \tag{2.118}$$

where $\boldsymbol{B} = \boldsymbol{Q}^{\boldsymbol{T}}\boldsymbol{\Lambda}\boldsymbol{Q}$.

So investigating the spatial variability of the variables taken into account it will observed the different kind of anisotropy. In particular, if the initial range of the variogram changes with direction and a simple transformation of the coordinates will remove it, then this is known as *geometric anisotropy*(Fig. 2.18a), if the sill variance fluctuates with changes in direction, this might indicate the presence of preferentially orientated zones with different means. This is known as *zonal anisotropy*(Fig.2.18 b).

The geometric anisotropy, can be represented by a linear transformation of the spatial coordinates of a corresponding isotropic model. It allows to relate the class of ellipsoidally anisotroipic random functions to a corresponding isotropic random function. This is essential because variogram models are defined for the isotropic case. The linear transformation extends in a simple way a given isotropic variugram to a whole class of ellipsoidally anisotropic variograms. While if experimental variograms calculated in different directions suggest a different value for the sill, as before said, this is a zonal anisotropy (for example, in 2D the sill along the x_2 coordinate might be much larger than along x_1). In such a situation a common strategy is to fit first to an isotropic model $\gamma_1(\boldsymbol{h})$ to the experimental variogram along the x_1 direction.

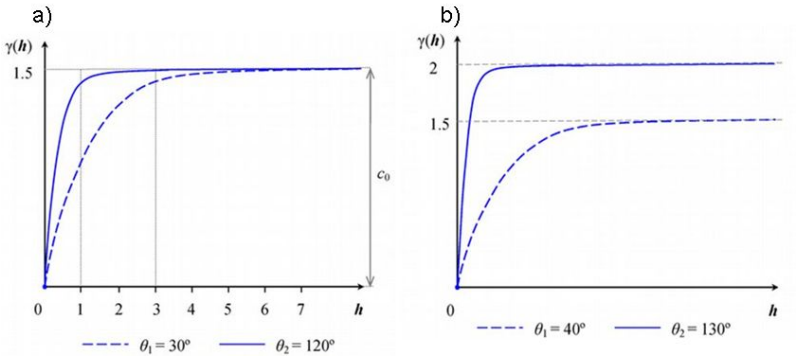

Figure 2.18: a) an example of geometric anisotropy; b) an example of zonal anisotropy.

Second, to add a geometrically anisotropic variogram $\gamma_2(\boldsymbol{h})$, which is designed to be without effect along the x_1 coordinate by providing it with a very large range in that direction through an anisotropy coefficient. The final variogram model is then

$$\gamma(\boldsymbol{h}) = \gamma_1(\boldsymbol{h}) + \gamma_2(\boldsymbol{h}) \tag{2.119}$$

in which the main axis of the anisotropy ellipse for $\gamma_2(\boldsymbol{h})$ is very large in the direction x_1.

The underlying random function model overlays two uncorrelated processes $Z_1(\boldsymbol{x})$ and $Z_2(\boldsymbol{x})$

$$Z(\boldsymbol{x}) = Z_1(\boldsymbol{x}) + Z_2(\boldsymbol{x}) \tag{2.120}$$

From the point of view of the regionalized variable, the anisotropy of $\gamma_2(\boldsymbol{h})$ can be due to morphological objects which are extremely elongated in the direction of x_1, crossing the borders of the domain. These units slice up the domain along thus creating a zonation along x_1, which explains the additional variability to be read on the variogram in that direction.

Another important aspect of using kriging is the issue of the support size. In geostatistics, one can control the support size of the outputs by averaging multiple (randomized) point predictions over regular blocks of land. This is known as block prediction (Heuvelink and Pebesma, 1999). A problem is that it is possible sample elevation at point locations, and then interpolate them for blocks of e.g. 10×10 m, but it could also take composite samples and interpolate them at point locations. This often confuses GIS users because as well as using point measurements to interpolate

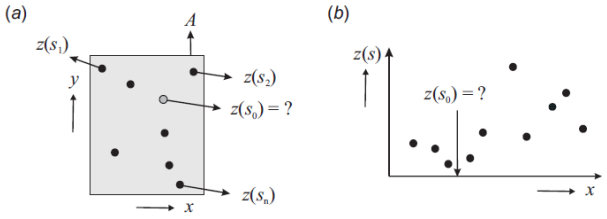

Figure 2.19: Spatial prediction implies application of a prediction algorithm to an array of grid nodes (point to point spatial prediction). The results are then displayed using a raster map.

values at regular point locations (e.g. by point kriging), and then display them using a raster map (see Fig.2.19), one can also make spatial predictions for blocks of land (block kriging) and display them using the same raster model (Bishop and McBratney, 2001). For simplicity, in the case of block-kriging, one should always use the cell size that corresponds to the support size.

2.3 RK

The residual kriging method is presented here taking into account the particular case of the regression-kriging. The same consideration can be made about other inexact deterministic interpolation methods.

Matheron (1969) proposed that a value of a target variable at some location can be modelled as a sum of the deterministic and stochastic components:

$$Z(x) = m(s) + \varepsilon'(x) + \varepsilon'' \tag{2.121}$$

which he termed universal model of spatial variation. Both deterministic and stochastic components of spatial variation can be modelled separately. By combining the two approaches, it is possible to obtain:

$$\begin{aligned} \hat{z}(\boldsymbol{x}_0) &= \hat{m}(\boldsymbol{x}_0) + \hat{e}(\boldsymbol{x}_0) \\ &= \sum_{k=0}^{p} \hat{\beta}_k \cdot y_k(\boldsymbol{x}_0) + \sum_{i=1}^{n} \lambda_i \cdot e(\boldsymbol{x}_i) \end{aligned} \tag{2.122}$$

where $\hat{m}(\boldsymbol{x}_0)$ is the fitted deterministic part, $\hat{e}(\boldsymbol{x}_0)$ is the interpolated residual, $\hat{\beta}_k$ are estimated deterministic model coefficients ($\hat{\beta}_0$ is the estimated intercept), λ_i are

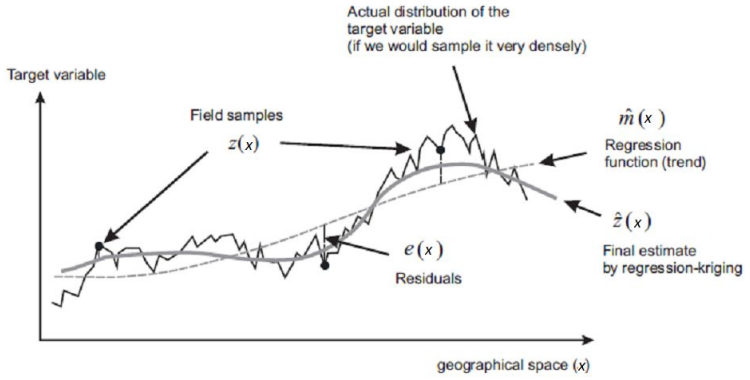

Figure 2.20: A schematic example of regression-kriging: fitting a vertical cross-section with assumed distribution of an environmental variable in horizontal space.

kriging weights determined by the spatial dependence structure of the residual and where $e(\boldsymbol{x}_i)$ is the residual at location $\boldsymbol{x}_i$. The regression coefficients $\hat{\beta}_k$ can be estimated from the sample by some fitting method, e.g. ordinary least squares (OLS) or, optimally, using Generalized Least Squares (Cressie, 1993). Once the deterministic part of variation has been estimated, the residual can be interpolated with kriging and added to the estimated trend (Fig. 2.20).

In matrix notation, regression-kriging is commonly written as (Christensen, 2001, p.277):

$$\hat{z}_{RK}\left(\boldsymbol{x}_0\right) = \boldsymbol{y}_0^T \cdot \hat{\beta}_{GLS} + \lambda_0^T \cdot \left(\boldsymbol{z} - \boldsymbol{y} \cdot \hat{\beta}_{GLS}\right) \tag{2.123}$$

where $\hat{\beta}_{GLS}$ is the vector of estimated regression coefficients, $\hat{z}_{RK}\left(\boldsymbol{x}_0\right)$ is the predicted value at location $\boldsymbol{x}_0$, $\boldsymbol{y}_0$ is the vector of $p+1$ estimators and λ_0 is the vector of n kriging weights used to interpolate the residuals. The model in Equation 2.123 is considered to be the Best Linear Estimate of spatial data. It has a prediction variance that reflects the position of new locations (extrapolation) in both geographical and feature space:

$$\begin{gathered}\hat{\sigma}_{RK}^2 = \left(C_0 + C_1\right) - \boldsymbol{c}_0^T \cdot \boldsymbol{C}^{-1} \cdot \boldsymbol{c}_0 \\ + \left(\boldsymbol{y}_0 - \boldsymbol{y}^T \cdot \boldsymbol{C}^{-1} \cdot \boldsymbol{c}_0\right)^T \cdot \left(\boldsymbol{y}^T \cdot \boldsymbol{C}^{-1} \cdot \boldsymbol{y}\right)^{-1} \cdot \left(\boldsymbol{y}_0 - \boldsymbol{y}^T \cdot \boldsymbol{C}^{-1} \cdot \boldsymbol{c}_0\right)\end{gathered} \tag{2.124}$$

where $C_0 + C_1$ is the sill variation and c_0 is the vector of covariances of residuals at the unvisited location (ungauged site). Obviously, if the residuals show no spatial auto-correlation (pure nugget effect), the regression-kriging (Eq. 2.123) converges to pure multiple linear regression because the covariance matrix ($\boldsymbol{C}$) becomes identity matrix:

$$\boldsymbol{C} = \begin{bmatrix} C_0 + C_1 & \cdots & 0 \\ \vdots & C_0 + C_1 & 0 \\ 0 & 0 & C_0 + C_0 \end{bmatrix} = (C_0 + C_1) \cdot \boldsymbol{I} \tag{2.125}$$

so the kriging weights at any location predict the mean residual, i.e. 0 value.

Chapter 3

Method of spatial interpolation for areal climatic variable

As said in Chapter 1, interpolation of runoff is more complex that interpolation of the variables usually assimilated to a point process. Runoff is a generalized random space-time process with a local support equal to the basin area. Because of this, in the application of a spatial interpolation methods, the areal nature of runoff can not be neglected.

As anticipated in the previous chapter, among the methods presented in literature, that better accounting for the specific characteristich of the runoff has been chosen. Furthermore, among the different approaches described in literature, the geostatistical one has been selected. As a matter of fact, it is a method that analyse with a better accuracy the stocastic variables and their spatial variations. This method will be better explained in the following.

In particular a method for the runoff estimation in ungauged basins is presented here. This method is based on the the solution of a system of equations similar to those used for kriging method. It takes both the area and the nested nature of catchments into account. The main appeal of the method is that it is a best linear unbiased estimator (BLUE) adapted for the case of stream networks without any additional assumptions. The presented method can be seen as an interesting approach to address the problem of Prediction in Ungauged Basins (PUB) (Sivapalan et al., 2003), i.e. to estimate streamflow and streamflow-related variables at locations where no measurements are available.

This chapter starts with a short presentation of basically concepts about the components of water balance for a drainage basins. In the second and third sections the interpolation of runoff as point process and the interpolation of runoff as areal process are treated. The fourth section shows a detailed description of the stocastic interpolation system for runoff as an areal process, applied to one nested basin. The fifth section presents some consideration about the distance measures for hydrological data having support.

3.1 Water balance components. Mapping runoff

Dooge (2005) writes: "Hydrologists are lucky that, in progressing from the continuum scale to the global scale, the equation of continuity can be integrated in order to move from a lower scale to a higher scale. This useful result occurs because the equation of continuity can be written in a linear form which contains no empirical coefficients. None of the other basic equations of hydrology possess these two properties and hence we can identify the equation of continuity as the fundamental equation of hydrology and its validity as the fundamental theorem of hydrology. The basic requirement in hydrologic analysis is to satisfy this equation and then to tackle the problem of the remaining equations which are non-linear."

The equation of continuity is the basic principle for establishing the water balance equation for a drainage basin. It reads (Fig 3.1):

$$p + g_{in} = g_{out} + et + \Delta s \tag{3.1}$$

where p is precipitation, et evapotranspiration, g groundwater flow (in; out), q stream outflow and s storage. The water balance equation is the simplest model we can think of for a drainage basin. However it is important to note that the drainage basin determined from the topographic water divider might not have that expected fundamental role for a balance calculation. In principle it is assumed that the stream receives water, i.e. an effluent stream so that the discharge increase downstream. This is typically the situation in temperate and humid tropic climates. Exceptions are local areas with deep groundwater percolation, and sandy or karstic areas. These features are especially sensitive for small headwater basins. For arid and semi-arid climatic situations the drainage basin may loose its meaning and streams become influent and discharge decreases downstream giving rise to intermittent and ephemeral streams.

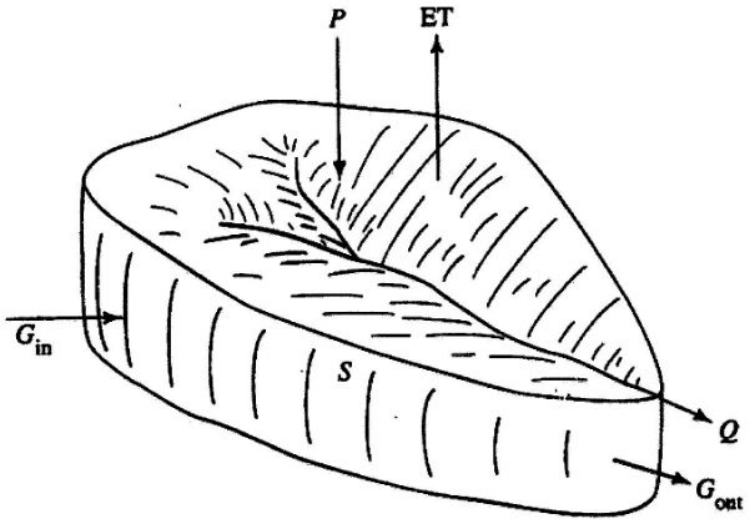

Figure 3.1: The components of the water balance for a drainage basin: p is precipitation, et evapotranspiration, g ground-water flow (in; out), q stream outflow and s storage.

This latter situation is not treated here. It is in the following assumed that discharge increases downstream.

The water balance perspective is extended in this chapter from the small headwater drainage basin downstream to the larger basin or to a region to be able to make maps of the three basic components of the water balance – precipitation, evapotranspiration and runoff. Such maps have three main uses (Gottschalk and Krasovskaia, 1997). First, they indicate the quantity of water resources available in a region, and are valuable water resource planning tools. Second they can be used to determine regional and continental water budgets, and thus contribute to the understanding of global water fluxes. Third, mapped water balance components provide validation data for atmospheric simulation models and macro-scale hydrological models. A fourth use might be added namely as a diagnostic tool for checking the uncertainty of available hydrologic data of the water balance components.

In this chapter we want to focus more on the estimate of the runoff to obtain maps of this variable. There are examples of use of conceptual runoff models for making runoff maps. In Sweden a regionalised HBV-model (Jutman, 1992) was used for calculation of the runoff for the period 1961-90 and a similar approach was used

by Beldring et al.(2002). The hydrological model guaranties that the water balance is satisfied and basic observed data are adjusted accordingly (as a rule to give optimal fit to observed runoff). A standard procedure is to include a precipitation correction as one of model parameters. It is an ill posed problem if the tuning of the model also should account for the correction of basic input data. Before going into modelling exercises there is a need to grasp the basic uncertainty in observed data and treat the possible correction of systematic errors in data as an independent problem and not as a result of calibration procedure.

A starting point for this is ordinary stochastic interpolation with grid or areas support. The target areas for this interpolation can either be elements of a grid network or of sets of sub-basins or successive drainage basins connected to points along rivers from headwaters down to an outlet. In the first case the result will produced in terms of a map while in the other case runoff estimates are given along the main river network only. Interpolation constraints can be introduced in the interpolation procedure so that the water balance is preserved along rivers. Furthermore a correction algorithm can be formulated so that the vertical water balance between precipitation, evapotranspiration and runoff is preserved.

3.2 Interpolation of runoff as point process

Runoff with the dimension of flow per unit area is, as a rule, considered as a point process $q(\boldsymbol{u})$ continuous in space $\boldsymbol{u} = (u_1, u_2)$.. Intuitively the point runoff process is interpreted as a contribution from an arbitrary point in the basin with area A to the observed streamflow $q(A)$ from the basin. In the simplest way it is determined by dividing the observed streamflow by the corresponding drainage area, creating a spatial step function constant over controlled sub-areas defined by the network of discharge stations. Another possible way of defining runoff is to subtract the estimated point actual long-term mean evapotranspiration $et(\boldsymbol{u})$ from the estimated point long-term mean precipitation $p(\boldsymbol{u})$ creating a continuous spatial process of precipitation excess $q(\boldsymbol{u}) = p(\boldsymbol{u}) - et(\boldsymbol{u})$. The integrated value of this spatial function over a basin area A should in principle coincide with the observed streamflow $Q(A)$ for a given time period:

$$Q(A) = \frac{1}{A} \int_A \int q(\boldsymbol{u})\, d\boldsymbol{u} \tag{3.2}$$

A formula (weighted average) for interpolation of $q(\boldsymbol{u}_0)$ as a point process at location u_0 from regional observations of this variable $q(\boldsymbol{u}_i)$ $(i = 1, ..., N)$ is:

$$\hat{q}(\boldsymbol{u}_0) = \sum_{i=1}^{N} \lambda_i x(\boldsymbol{u}_i) = \Lambda^T \boldsymbol{Q} \tag{3.3}$$

where λ_{ij} are the weights, determined so that the interpolation error is minimized .

In the case of stochastic interpolation, optimal weights in equation 3.3 are found by minimizing the estimation variance. Adopting an assumption of local second-order stationarity of the process and under the condition of unbiasedness this leads to the following linear equation system for the calculation of weights:

$$\boldsymbol{C}\Lambda = \boldsymbol{C}_0 \tag{3.4}$$

where

$$\Lambda = \begin{bmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_N \\ \mu \end{bmatrix} \qquad C = \begin{bmatrix} Cov(\boldsymbol{u}_1, \boldsymbol{u}_1) & Cov(\boldsymbol{u}_1, \boldsymbol{u}_2) & \dots & Cov(\boldsymbol{u}_1, \boldsymbol{u}_N) & 1 \\ Cov(\boldsymbol{u}_2, \boldsymbol{u}_1) & Cov(\boldsymbol{u}_2, \boldsymbol{u}_2) & \dots & Cov(\boldsymbol{u}_2, \boldsymbol{u}_N) & 1 \\ \dots & \dots & \dots & \dots & \dots \\ Cov(\boldsymbol{u}_N, \boldsymbol{u}_1) & Cov(\boldsymbol{u}_N, \boldsymbol{u}_2) & \dots & Cov(\boldsymbol{u}_N, \boldsymbol{u}_N) & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \tag{3.5}$$

$$C_0 = \begin{bmatrix} Cov(\boldsymbol{u}_1, \boldsymbol{u}_0) \\ Cov(\boldsymbol{u}_2, \boldsymbol{u}_0) \\ \vdots \\ Cov(\boldsymbol{u}_N, \boldsymbol{u}_0) \\ 1 \end{bmatrix}$$

and μ is a Lagrange multiplier. The unbiasedness is warranted with unit sum weights, i.e.:

$$\sum_{i=1}^{N} \lambda_i^j = 1 \tag{3.6}$$

The optimal weights, i.e. the formal solution to equation 3.4 are calculated from:

$$\Lambda_j = C^{-1}C_0 \tag{3.7}$$

The elements in the matrix $\boldsymbol{C}$ represent the values of the fitted covariance function between each pair of data values located inside the network, while the elements of the column vector $\boldsymbol{C}_0$ are the values of the fitted covariance function between the location of interest and each of the stations. The covariances of the matrix $\boldsymbol{C}$ are between point mean values.

It was stated that the premises of spatial homogeneity are hardly satisfied for hydrologic random processes and that it is common that the correlation structure demonstrates homogeneity while the covariances do not. The reason for this is the fact that the correlation is based on a standardization of the initial data and by this the non-homogeneity in mean and variance is eliminated. The mean values are random variables as they are functions of the random variables p and et, so stochastic interpolation is also in this case the appropriate method. We only have one mean value per point i.e. only one realization of random process. It was recommended to use semivariogram to characterise the spatial variation in case of only one realization of the random process is available. For the interpolation of mean values it is completely equivalent to use covariances or semivariogram for interpolation as both take into consideration the non-homogeneity. In the following covariances will be used.

3.3 Interpolation of runoff as areal process

Interpolation of runoff is more complex that interpolation of the two other hydrological variables. Runoff observations might nested i.e. the drainage basins of one station is contained in a larger basin of another station. It is therefore worthwhile to in a first step make a "denesting" of observed runoff within a larger basin.

A common situation is that a larger drainage basin with a gauging basin at its outlet contain one or several other gauging station upstream. Runoff productionfor the area of the basin between outlet station and an upstream one should in principle be achieved by simply subtracting the discharge of the upstream point from the downstream one. For the simple case illustrated in Fig. 3.2 with only one upstream station the denesting is thus represented by two values only $x(A_1)$ and $x(A_2)$, respectively. A_2 is the upstream area and A_1 is the intermediate area calculated from $A_1 = A - A_2$,

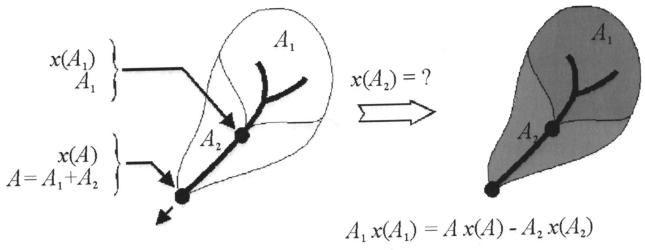

Figure 3.2: The principle of runoff denesting

where A is the area at the outlet. $x(A_1)$ is estimated from:

$$\hat{x}\left(A_1\right) = \frac{\left(Ax\left(A\right) - A_2 x\left(A_2\right)\right)}{A_1} \tag{3.8}$$

The procedure of denesting might be a first good control of the quality of data. Errors are directly detectable. Later when applying more sophisticated procedures for interpolation these errors will be present although less obvious to detect.

An example of denesting is illustrated in Figure 3.3; this procedure has been applied through for the 58 stations of annual mean runoff for the Glomma river. The map confirm the general observation that small basins are found in the basin headwaters, where a relative detailed picture of the variation in runoff is achieved. In the downstream part this variability is not revealed by the existing network of gauging stations. For sure the map indicate stations where the data need to be checked.

We can now turn to the interpolation equation for runoff. The advice is to use denested runoff data. The formulas above for interpolation for point observations to a target area a_i are altered to account for the drainage basin areas (denested areas) A_j for the $j = 1, ..., M_i$ runoff gauging sites used. The interpolation equation is:

$$\hat{q}(a_i) = \sum_{j=1}^{M_i} \lambda_j^i q(A_j) = \mathbf{\Lambda}_j^T \mathbf{Q} \tag{3.9}$$

where $\mathbf{\Lambda}_j^T$ as before is the transposed column vector of lambdas and $\boldsymbol{Q}$ the column vector of long-term runoff at the different . The weights λ_j^i with $j = 1, ..., N_i$, related to each of the grid cells i, are evaluated from $\mathbf{\Lambda}_j = \boldsymbol{C}^{-1}\boldsymbol{C}_{0i}$ where now:

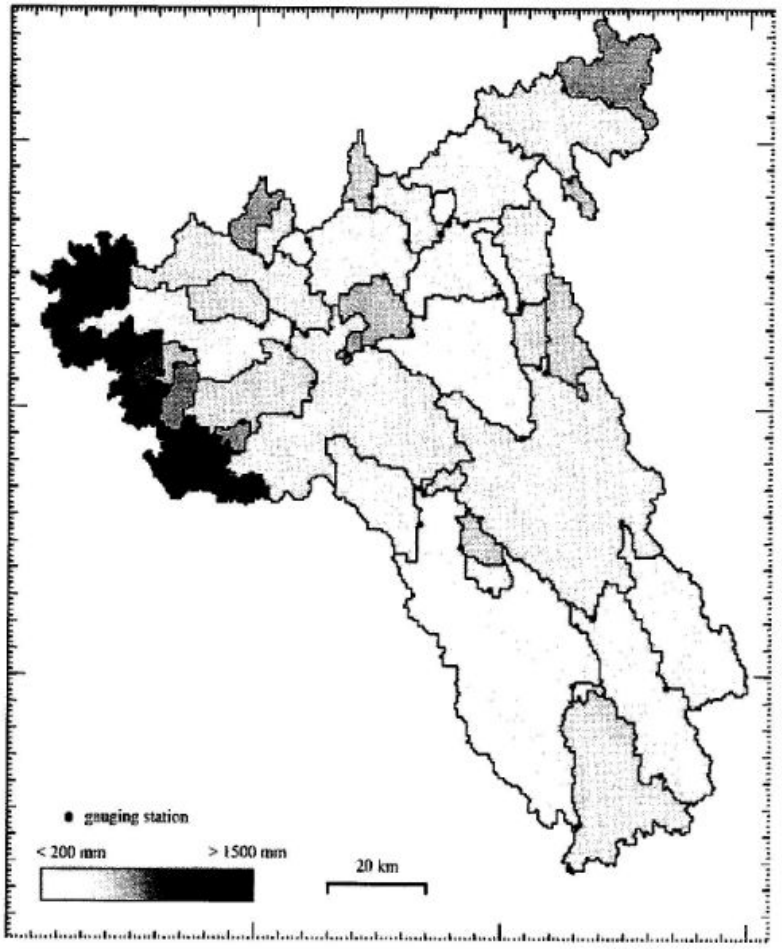

Figure 3.3: Denested runoff derived from observations 58 gauging stations in the Glomma basin, Norway.

$$\boldsymbol{\Lambda}_i = \begin{bmatrix} \lambda_1^i \\ \lambda_2^i \\ \vdots \\ \lambda_N^i \\ \mu^i \end{bmatrix} \quad \boldsymbol{C} = \begin{bmatrix} Cov(A_1, A_1) & Cov(A_1, A_2) & \dots & Cov(A_1, A_{M_i}) & 1 \\ Cov(A_2, A_1) & Cov(A_2, A_2) & \dots & Cov(A_2, A_{M_i}) & 1 \\ \dots & \dots & \dots & \dots & \dots \\ Cov(A_{M_i}, A_1) & Cov(A_{M_i}, A_2) & \dots & Cov(A_{M_i}, A_{M_i}) & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix} \tag{3.10}$$

$$\boldsymbol{C}_{0i} = \begin{bmatrix} Cov(A_1, a_i) \\ Cov(A_1, a_i) \\ \vdots \\ Cov(A_1, a_i) \\ 1 \end{bmatrix}$$

The elements in the matrix $\boldsymbol{C}$ represent covariances between mean annual runoff at the M_i^q observation stations in the search neighbourhood of grid i - $Cov(A_i, A_j) = Cov(\bar{X}(A_i), \bar{X}(A_j))$ - while the elements of the column vector $\boldsymbol{C}_{0i}$ are the covariances between runoff at these observation stations and runoff at grid cell i - $Cov(A_j, a_i) = Cov(\bar{X}(A_j), \bar{X}(a_i))$. These covariances are evaluated from a river covariance model.

An example of an interpolated grid map is shown in Figure 3.4 utilising the denested runoff values of Fig. 4. of mean annual runoff from 57 gauging stations in the drainage basin of the Glomma River has been utilised.

3.4 A detailed description of the stocastic interpolation system for runoff as an areal process. Application to one nested basin

As previously said, the aim of this method is to estimate the values of runoff in areas characterized by a magnitude lower than those of basins with available observed data.

The first step is the application of the denesting procedure at the nested basin, as described in the previous section. A the group of the non-overlapping basins has been obtained. This disaggregation procedure can be called the "*first level of hierarchization*".

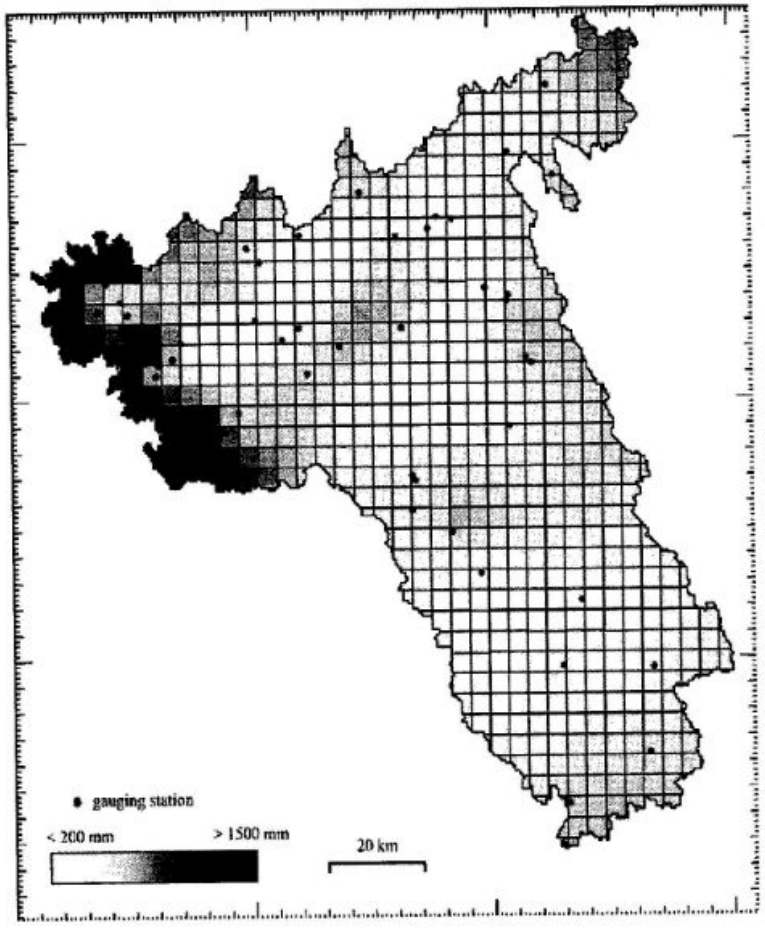

Figure 3.4: Interpolated mean annual runoff [mm/year] to grid cells 8x8 km over the drainage basin of the Glomma River, Norway with data from 57 gauging stations.

The second step is the calculation of distances between each pair of independent drainage basins following the hierarchy of the drainage network. Gottschalk (1993b) has focused on the requirement of a relevant distance definition between drainage basins to identify the spatial structure of runoff. The Euclidean distance measured between the gauging stations is not adapted to the runoff data (Gottschalk, 1993a). The appropriate distance should include the drainage network and the hierarchy of drainage basins in the system. Here, this is made possible by replacing the Euclidean distance by a "geostatistical" distance, called Ghosh distance. This geostatistical distance between drainage basin A and drainage basin B is expressed as the mean of the distances between all possible pairs of points inside A and B, respectively, i.e.:

$$d(A,B) = \frac{1}{AB}\int\limits_{A,\ B}\int ||\boldsymbol{u_A} - \boldsymbol{u_B}||\boldsymbol{du}_A\boldsymbol{du_B} \tag{3.11}$$

This distance allows a better identification of the spatial structure of runoff (Gottschalk, 1993a).

Practically, for the group of the non-overlapping basins obtained by the application of denesting procedure, random point for each basins were produced and The average over all possible distances between pairs points in the respective catchments or the ghosh distance was applied

$$d(A_n,A_m) = \frac{1}{A_nA_m}\int\limits_{A_n,A_m}\int ||\boldsymbol{u_{A_n}} - \boldsymbol{u_{A_m}}||\boldsymbol{du}_{A_n}\boldsymbol{du}_{A_m} \tag{3.12}$$

where$\|\boldsymbol{u_{A_n}} - \boldsymbol{u_{A_m}}\|$is the euclidean distance between all random points in the areas, taken pairwise. $A_n$and $A_m$are the areas of the non-overlapping units contained in the area taken into account and $n = 1, 2, ...N$ and $m = 1, 2, ..., N$ are the number of these non-overlapping basins. In the next section the theoretical basics that led to the concept of Ghosh distance has been explained in detail.

Thus an empirical covariogram can be derived and, under the assumption of second-order stationarity, an empirical corollary covariogram $Cov_e(A_n, A_m)$ is deduced. In particular, the values of semivariance and covariance to represent the empirical variogram and the empirical covariogram are calculated with the following equations:

$$Covariance\,(A_n,A_m) = (Q(A_n) - m_Q)*(Q(A_m) - m_Q) \qquad \forall m,n = 1,...,N \tag{3.13}$$

$$Semivariance\,(A_n, A_m) = 0.5 * (Q(A_n) - Q(A_m))^2 \qquad \forall m, n = 1, ..., N \qquad (3.14)$$

where m and n are the number of areas taken into account and m_Q is the annual mean of the annual runoff of the basins. In this way, the variance and covariance matrices are obtained. These matrices are square with a rank equal to the number of basins considered (for example, considering 32 sub-basins, the variance and covariance matrices have rank 32).

Before making the representation of the semivariogram and the covariogram, the main diagonal of the matrix of distance has been replaced with zeros, since it has been assumed that the distance of an area with itself must be zero whatever the criterion used to calculate it. Furthermore at zero distance the covariance should be equal to the variance and the semivariogram equal to zero, so the main diagonal of the matrix of covariance has been set equal to mean of the main diagonal of the semivariance matrix. On the contrary, the main diagonal of the semivariance matrix has been set equal to zero.

At this point, an experimental covariogram and semivariogram can be drawn. In this study, to the application of this procedure, only the covariogram (Cov_e) is taken into account, adopting an assumption of local second-order stationarity.

The related theoretical covariogram $Cov(A_n, A_m)$ with the local supports A_n and A_m, respectively, is derived in a similar manner by averaging the point process covariance function:

$$Cov\,(A_n, A_m) = \frac{1}{A_n A_m} \int_{A_n, A_m} \int Cov_e\,(||\boldsymbol{u_{A_n}} - \boldsymbol{u_{A_m}}||)\,\boldsymbol{du_{A_n}}\boldsymbol{du}_{A_m} \qquad (3.15)$$

In particular, a selection of possible theoretical models for the covariance function Cov_e it has to be done.

Once this procedure has been carried out will be passed to the "*second level of hierarchization*".

For estimation of runoff (q) as an areal process, the following formula is taken into account:

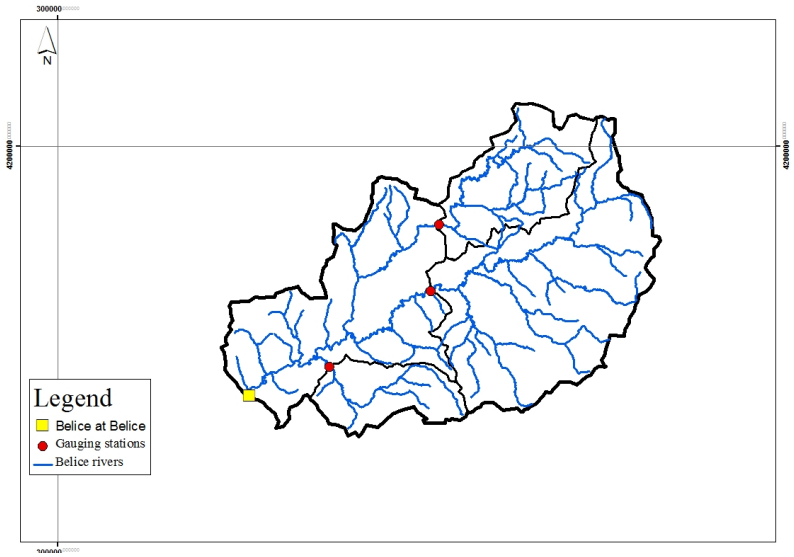

Figure 3.5: An example of nested basin and location of gauging stations

$$\hat{q}(a_0) = \sum_{j=1}^{N} \lambda_j q(A_j) = \tag{3.16}$$

where $a_0$is the fundamental unit of the map (*second level of hierarchization*), A_j and $j = 1, ..., N$ are the areas of drainage basins with observations (first level of hierarchization). $\boldsymbol{Q}$ is the column vector of observations and Λ^T is the transposed column vector of weights λ_j $(j = 1, ..., N)$, associated with the N observations.

The point of departure is a drainage basin A_T where the mean annual discharge Q_T at the outlet point is known from measurements or estimation. In the following this value will be treated as a known constant. In Figure 3.5, A_T is the area of total basin taking into account and Q_T is runoff at the outlet point, (squared yellow marker in figure) that is the sum of discharge of all sub-basins, considering the values obtained to the application of denesting procedure.

The area A_T is approximated by a regular grid of n_T fundamental square cells of area a, so that $A_T = n_T * a$. As only Q_T is available, the way in which the mean flow is distributed within the drainage basin is unknown and some assumptions are needed. It is assumed that the distribution across each fundamental unit (square cells) is uniform. Therefore, discharge data are converted into $mm\ year^{-1}$ using the

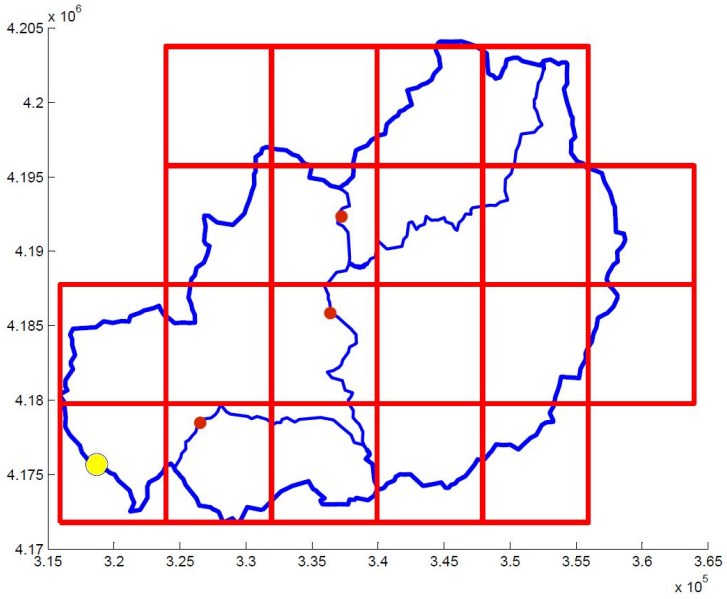

Figure 3.6: An example of an area $A_T$subdivided into M non-overlapping areas ΔA_i applied to the Belice basins (Sicily). The red dots are the gauging stations. The yellow dot is the outlet gauging station of the main drainage basin.

decomposition of each basin in grid cells:

$$q_T = \frac{Q_T}{A_T} = \frac{Q_T}{n_T a} \tag{3.17}$$

After this preliminary stage, a second level of hierarchization is faced.

The total area A_T can be subdivided into M non-overlapping areas ΔA_i, $(i = 1, ..., M)$, as Figure 3.6 shows.

The aim is to estimate the specific discharge $q(\Delta A_i)$ for each of these areas. Such a disaggregation can naturally keep on with a stepwise disaggregation of each of the discharges $q(\Delta A_i)$ into smaller units. In the following, algorithms for the interpolation of runoff based on the unsealed runoff $q(\Delta A_i)$ are developed. Afterwards the interpolated runoff depth can be easily aggregated to the drainage basin ΔA_i, by reversing equation 3.17 to assess discharge values:

$$Q\left(\Delta A_i\right) = n_i a\, q(\Delta A_i) \tag{3.18}$$

where n_i is the area of the unit ΔA_i, measured in terms of number of cells.

Assume that discharge observations are available at N basins A_j $j = 1, ..., N$ as above. Insertion into equation 3.16 yields the following equation for interpolation of the specific runoff:

$$q(\Delta A_i) = \sum_{j=1}^{N} \lambda_j^i q(A_j) = \Lambda^T \mathbf{H} \tag{3.19}$$

The optimal weights in equation 3.19 (or 3.16) are found by minimizing the estimation variance. Adopting an assumption of local second-order stationarity of the process and under the condition of unbiasedness this leads to the following linear equation system for the calculation of weights λ_i^j $(j = 1, ..., N)$

$$\Lambda_j = C^{-1} C_{0j} \tag{3.20}$$

with constraint

$$\sum_{i=1}^{M} \lambda_i^j = 1 \tag{3.21}$$

with

$$\Lambda_i = \begin{bmatrix} \lambda_1^i \\ \lambda_2^i \\ \vdots \\ \lambda_N^i \\ \mu^i \end{bmatrix} \qquad C = \begin{bmatrix} Cov(A_1, A_1) & Cov(A_1, A_2) & \dots & Cov(A_1, A_N) & 1 \\ Cov(A_2, A_1) & Cov(A_2, A_2) & \dots & Cov(A_2, A_N) & 1 \\ \dots & \dots & \dots & \dots & \dots \\ Cov(A_N, A_1) & Cov(A_N, A_2) & \dots & Cov(A_N, A_N) & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix}$$

and

$$C_0 = \begin{bmatrix} Cov(A_1, \Delta A_1) \\ Cov(A_1, \Delta A_2) \\ \vdots \\ Cov(A_1, \Delta A_N) \\ 1 \end{bmatrix}$$

The streamflow at the outlet point of basin A_T is given by:

$$\sum_{i=1}^{M} \Delta A_i q\left(\Delta A_i\right) = \sum_{i=1}^{M} n_i a\, q(\Delta A_i) = \sum_{i=1}^{M} n_i a \left(\sum_{j=1}^{N} \lambda_j^i q(A_j) \right) \tag{3.22}$$

The sum of the interpolated discharge for each of the sub-basins, calculated from equation 3.22, does not necessarily match the discharge Q_T observed downstream. A further step is to include a constraint so that the interpolated lateral inflow is balanced with the observed runoff in the river system. Rearrangement of equation 3.22 yields:

$$\sum_{i=1}^{M} n_i a \left(\sum_{j=1}^{N} \lambda_j^i q(A_j) \right) \Rightarrow \sum_{i=1}^{M} \left(\sum_{j=1}^{N} n_i \lambda_j^i q(A_j) \right) = n_T q_T \tag{3.23}$$

This new constraint supplements the previous ones presented in equation 3.21. The interpolation equation 3.19 remains the same for this case, but the weights λ_j^i ($j = 1, ...N$, $i = 1, ...M$) have to be calculated simultaneously for all M elements. Optimal weights were found through the solution of the system of equations:

$$\Lambda = C^{-1} C_0 \tag{3.24}$$

with

$$\mathbf{\Lambda} = \begin{bmatrix} \mathbf{L_1} \\ \mathbf{L}_2 \\ \vdots \\ \mathbf{L}_M \\ \mu_T \end{bmatrix}, \qquad \mathbf{C} = \begin{bmatrix} \mathbf{K} & 0 & \cdots & 0 & \mathbf{V}_1 \\ 0 & \mathbf{K} & 0 & \dots & \mathbf{V}_2 \\ \vdots & 0 & \ddots & 0 & \vdots \\ 0 & \dots & 0 & \mathbf{K} & \mathbf{V}_M \\ \mathbf{V}_1^T & \mathbf{V}_2^T & \dots & \mathbf{V_M^T} & 0 \end{bmatrix}$$

$$\mathbf{C_0} = \begin{bmatrix} \mathbf{G}_1 \\ \mathbf{G}_2 \\ \vdots \\ \mathbf{G}_M \\ n_T q_T \end{bmatrix} \quad L_i = \begin{bmatrix} \lambda_1^i \\ \lambda_2^i \\ \vdots \\ \lambda_N^i \\ \mu^i \end{bmatrix}$$

$$\mathbf{K} = \begin{bmatrix} Var(A_1) & Cov(A_1, A_2) & \dots & Cov(A_1, A_N) & 1 \\ Cov(A_2, A_1) & Var(A_2) & \dots & Cov(A_2, A_N) & 1 \\ \dots & \dots & \dots & \dots & \dots \\ Cov(A_n, A_1) & Cov(A_n, A_2) & \dots & Var(A_N) & 1 \\ 1 & 1 & 1 & 1 & 0 \end{bmatrix},$$

$$\mathbf{V_i} = \begin{bmatrix} n_i q(A_1) \\ n_i q(A_2) \\ \vdots \\ n_i q(A_N) \\ 0 \end{bmatrix}$$

$$\mathbf{G}_i = \begin{bmatrix} Cov(A_1, \Delta A_1) \\ Cov(A_1, \Delta A_2) \\ \vdots \\ Cov(A_1, \Delta A_N) \\ \mu^i \end{bmatrix}$$

Then, using the 3.19 we obtain an estimate of the runoff areas in the areas ΔA_i. The map that can be obtain in this step is such as show in Figure 3.4. The same procedure will be followed for the third and final levels of hierarchization: the "*third level of hierarchization*".

3.5 Distance measures for hydrological data having a support

As before said, runoff characteristics are, by definition, derived as integrated values over a basin i.e. representing a generalised random space-time process with a local support equal to the basin area. The support, as well as the organisation of the river network, influence the variance-covariance functions and semivariograms estimated from the original data series and also all river flow statistics including low flow and floods and river flow regimes (Gottschalk, 1993, Gottschalk et al., 2006). Covariance and correlation are functions of distance between observation points in space. How can the distance between the irregular supports that basin areas represent be adequately defined? This is the topic that is developed in this section. The Matèrn formulation of the covariance function is exploited. The interest is especially devoted to the determination of the density function $f(\lambda)$ of all possible distances between/within line segments or areas and to what extent the expected value of this distribution m_Λ (here named Ghosh distance) might be a proper distance measure. This was suggested earlier by Gottschalk (1993) and Sauquet et al.(2000) but without giving an in depth motivation.

3.5.1 Distance measures

There are several possible alternatives to define the distance between two basins A_1 and A_2:

1. Euclidian distance between the two outlets of the basins;
2. Distance along rivers between the two outlets;
3. Euclidian distance between the two centres of gravity d_{CG};
4. The expected value m_Λ of all possible distances λ between basins A_1 and A_2, here for short named Ghosh-distance (Ghosh, 1950).

The first two are of interest when representing variance-covariance and correlation along rivers in a graphical form but are not of help when analysing spatial variability of data or interpolation of data. For these latter tasks the distance between centres of gravity and the Ghosh distance are better suited. Distance between basin centres of gravity do not need any further explanations. That is why the interest herein is especially devoted to the determination of the density function $f(\lambda)$ of all possible distances between/within line segments or areas and to what extent the expected value of this distribution m_Λ (called Ghosh distance) might be a proper distance measure for the application developed in this paper.

3.5.2 Distances along a straight line

To start distances along a straight line is analysed that may be a straight rive or a time axis. Two intervals are situated along the line with a lag L between them. This lag L is always measured from the end of the first interval T_1 to the beginning of the second interval T_2 (with sign). It is assumed that $T_1 \leq T_2$, if this is not the case T_1 and T_2 should change place in all expressions developed below. We will now look for the distribution function of all possible distances between these intervals. Let us look first at the special case $T_1 = T_2 = T$; $L = -T\ T$ which has the well known expression for the distribution function (Ghosh, 1951):

$$f(|\lambda|) = \frac{2}{T}\left(1 - \frac{|\lambda|}{T}\right); \quad 0 \leq |\lambda| \leq T \tag{3.25}$$

The moments of this distribution are derived as:

$$E\left[|\lambda|^n\right] = \frac{2T^n}{(n+1)(n+2)} \tag{3.26}$$

The general expression for the n-th order moment allows the determination of the mean m_Λ , the variance σ_Λ^2 and the third order central moment μ_3 as:

$$m_\Lambda = E\left[\lambda\right] = \frac{1}{3}T \tag{3.27}$$

$$\sigma_\Lambda^2 = \mu_2 = E\left[(\lambda - m_\Lambda)^2\right] = \frac{1}{18}T^2 \tag{3.28}$$

$$\mu_3 = E\left[(\lambda - m_\Lambda)^3\right] = \frac{1}{135}T^3 \tag{3.29}$$

$$\mu_4 = E\left[(\lambda - m_\Lambda)^4\right] = \frac{1}{135}T^4 \tag{3.30}$$

For a second case there is no overlap between the intervals T_1 and T_2 ; and $L > 0$ and the distribution becomes:

$$f_1(\lambda) = \frac{1}{T_2}\left(\frac{\lambda - L}{T_1}\right); \qquad L \le \lambda \le T_1 + L \tag{3.31}$$

$$f_2(\lambda) = \frac{1}{T_2}; \qquad T_1 + L \le \lambda \le T_2 + L \tag{3.32}$$

$$f_3(\lambda) = \frac{1}{T_2}\left(1 - \frac{\lambda - L - T_2}{T_1}\right); \qquad T_2 + L \le \lambda \le T_1 + L + T_2 \tag{3.33}$$

The n-th order moment of this distribution is derived as:

$$E\left[\lambda^n\right] = \frac{(T_1 + T_2 + L)^{n+2} - (T_1 + L)^{n+2} - (T_2 + L)^{n+2} + L^{n+2}}{(n+1)(n+2)T_1T_2} \tag{3.34}$$

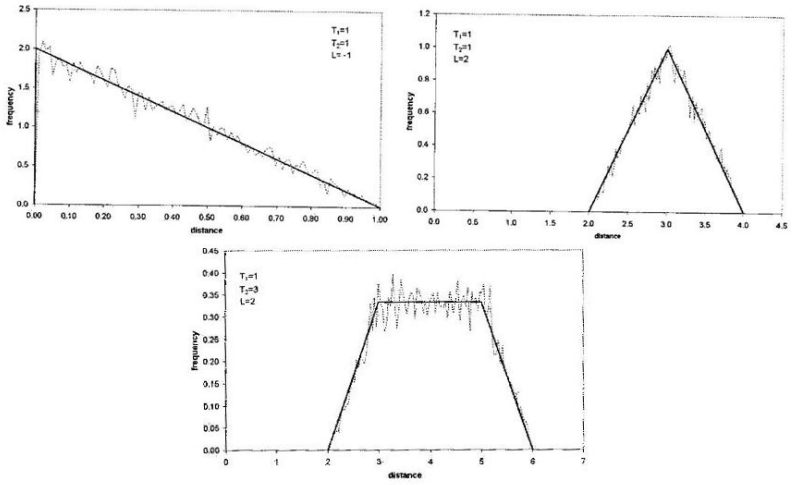

Figure 3.7: Distribution function of distances between to segments T_1 and T_2 along a line L units apart. The full size line shows the theoretical distribution and the dotted the sample distribution of 10000 generated distance.

which allows the determination of the mean m_Λ , the variance σ_Λ^2 and the third order central moment μ_3 as:

$$m_\Lambda = E[\lambda] = L + \frac{1}{2}(T_1 + T_2) \tag{3.35}$$

$$\sigma_\Lambda^2 = \mu_2 = E\left[(\lambda - m_\Lambda)^2\right] = \frac{1}{12}\left(T_1^2 + T_2^2\right) \tag{3.36}$$

$$\mu_3 = E\left[(\lambda - m_\Lambda)^3\right] = 0 \tag{3.37}$$

Examples of density functions for these two cases are illustrated in Fig. 3.7. Their parameters are given in Table 1. For comparison sample distribution functions generated from 10000 random distances drawn along lines are also included in the figure.

Parameter	T_1=1, T_2=1, L=-1	T_1=1, T_2=1, L=2	T_1=1, T_2=3, L=2
mean	0.00	3.00	4.00
std.dev.	0.41	0.41	0.91
skew	0.00	0.00	0.00
median	0.00	3.00	4.00
min.	-1.00	2.00	2.00
max.	1.00	4.00	5.00

Table 3.1: Parameters for density function shown in Fig. 3.7. The moments for the distributions are estimated theoretically

3.5.3 Distance between areas

For area it cannot be expected to find analytical expression exist for $f(\lambda)$ as above for interval along a line. Analytical expressions exist for special case only and Matèrn (1960) and Ghosh (1951) found that for a rectangle the distribution can be expressed as:

$$f(\lambda) = \frac{1}{\sqrt{A}} g\left(\frac{\lambda}{\sqrt{A}}, \sqrt{l_1, l_2}\right) \tag{3.38}$$

with

$$g(\omega, a) = 2\omega\left(g_1(\omega, a)\right) + g_2(\omega a, a) + g_2\left(\frac{\omega}{a}, \frac{1}{a}\right) \tag{3.39}$$

$$g_1(\omega, a) = \begin{cases} \pi + \omega^2 - 2\omega\left(a + \frac{1}{a}\right); & 0 < \omega < \sqrt{(a^2 + a^{-2})} \\ 0 & otherwise \end{cases} \tag{3.40}$$

$$g_2(\omega, a) = \begin{cases} 2\sqrt{\omega^2 - 1} - 2arccos\left(\frac{1}{\omega}\right) - a^{-2}(\omega - 1)^2; & 1 < \omega < \sqrt{1 + a^2} \\ 0 & otherwise \end{cases} \tag{3.41}$$

where l_1 and l_2 are the side lengths in the rectangle.

The moment for this case are derived as (Ghosh, 1950):

$$\alpha_1 = \frac{1}{6}\left\{\frac{l_2^2}{l_1} cosh^{-1}\left(\frac{d}{l_2}\right) + \frac{l_1^2}{l_2} cosh^{-1}\left(\frac{d}{l_1}\right) + \frac{1}{15}\left(\frac{l_1^3}{l_2^2} + \frac{l_2^3}{l_1^2}\right) - \frac{1}{15} d\left(\frac{l_1^2}{l_2^2} + \frac{l_2^2}{l_1^2} - 3\right)\right\} \tag{3.42}$$

$$\alpha_2 = \frac{1}{6} d^2 \tag{3.43}$$

$$\alpha_3 = \frac{1}{20}\left\{\frac{l_2^4}{l_1} cosh^{-1}\left(\frac{d}{l_2}\right) + \frac{l_1^4}{l_2} cosh^{-1}\left(\frac{d}{l_1}\right) + \frac{2}{105}\left(\frac{l_1^5}{l_2^2} + \frac{l_2^5}{l_1^2}\right) - d\left(\frac{2}{105}\left(\frac{l_1^4}{l_2^2} + \frac{l_2^4}{l_1^2}\right) - \frac{5}{84} d^2\right)\right\} \tag{3.44}$$

where $d = \sqrt{l_1^2 + l_2^2}$.

With known values of l_1 and l_2, the values of mean m_Λ, variance σ_Λ^2 and skewness μ_3 can be worked out. For example, the approximate values of the mean i.e. the Ghosh-distance ($m_\Lambda = \alpha_1$) for $l_1 = l_2$, $l_1 = 2l_2$ and $l_1 = 4l_2$ are given by $m_\Lambda = 0.521 l_1$, $m_\Lambda = 0.402 l_1$ and $m_\Lambda = 0.357 l_1$ respectively. For the circle with a diameter l_1, and for ellipses with the eccentricity of 2 and 4 and with the length of principal axis l_1, the corresponding values are $m_\Lambda = 0.453 l_1$, $m_\Lambda = 0.349 l_1$ and $m_\Lambda = 0.309 l_1$, respectively (the values are derived by the Monte Carlo method). These expressions thus allow the estimation of the variance for simple geometries, which eventually can approximately describe a real basin.

As an illustration Fig. 3.8 shows the Ghosh-distances m_Λ within drainage basins as a function of the areas of these basins for the Glomma River basin in Norway (~40000 km2), Moselle River basins in France (~40000 km2) and for all drainage basins in territory of Costa Rica (~50000 km2). Theoretical curves for a square and for rectangles are introduced in the diagrams for comparison. The dotted graphs for the natural basins plots over the whole range of the theoretical curves for small and

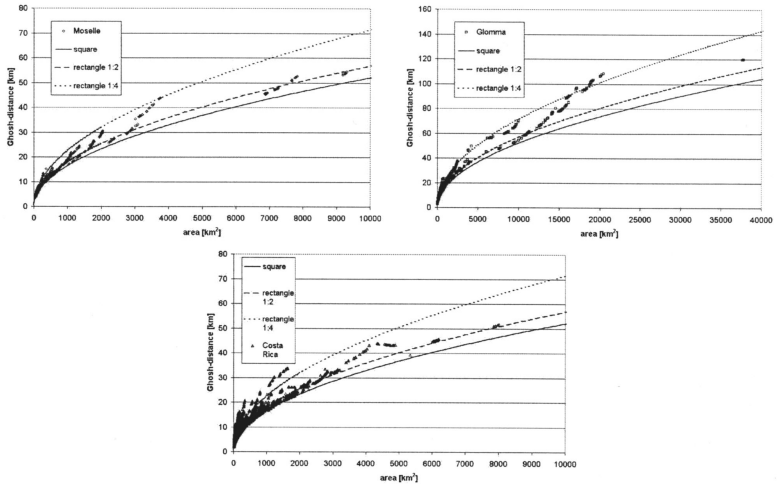

Figure 3.8: Ghosh distances m_Λ as a function of basin area estimated from digital map data for the Moselle River basin (France,) Glomma River basin (Norway) and for all drainage basins in Costa Rica as well as the corresponding theoretical functions for known regular areas.

moderate sized basins but a preference for the stretched rectangle curves is noted for the larger basins. This specially concerns the data from Glomma basins which indeed is dominated by two very long and relatively narrows valleys Osterdalen and Gullbrandsdalen.

The traditional basin distance measure in hydrology has been distance between centres of gravity d_{cG}. What is the difference between this distance measure and the Ghosh distance. This is illustrated in Fig. 3.8 where these two distances are plotted against each other for the Glomma River basin, the Moselle River basin and drainage basins in Costa Rica (∼50000 km2). For the Glomma and Moselle data there may be rather large differences between the two when near in space. For the Costa Rican data there is almost a perfect agreement between the two except for very small basins. The explanation to these different performances is first of all to be found in the existence of nested basins, which are relatively many in case of the Glomma and Moselle and few for Costa Rica. Furthermore both Glomma and Moselle are closed drainage basins and this give rise to the bigger scatter of these cases. The Euclidian distance between centres of gravity can obtain a value of zero even when

the two basins do not coincide, which cannot be the case for the Ghosh - distance. Furthermore the following inequality between these two distance measures is valid:

$$m_{\Lambda}\left(A_1, A_2\right) = E\left[|\lambda|\right] = E\left[|\boldsymbol{u}_1 - \boldsymbol{u}_2|\right] \geq \left|E\left[\boldsymbol{u}_1\right] - E\left[\boldsymbol{u}_2\right]\right| = d_{CG} \tag{3.45}$$

For a landscape with many small non-nested basins, like in e.g. Costa Rica, the difference between Ghosh-distance and distance between centre of gravity is relatively small. On the other hand, the difference is significanr for large drainage basins containing many nested sub-basins, like e.g. Glomma, Norway and Moselle, France.

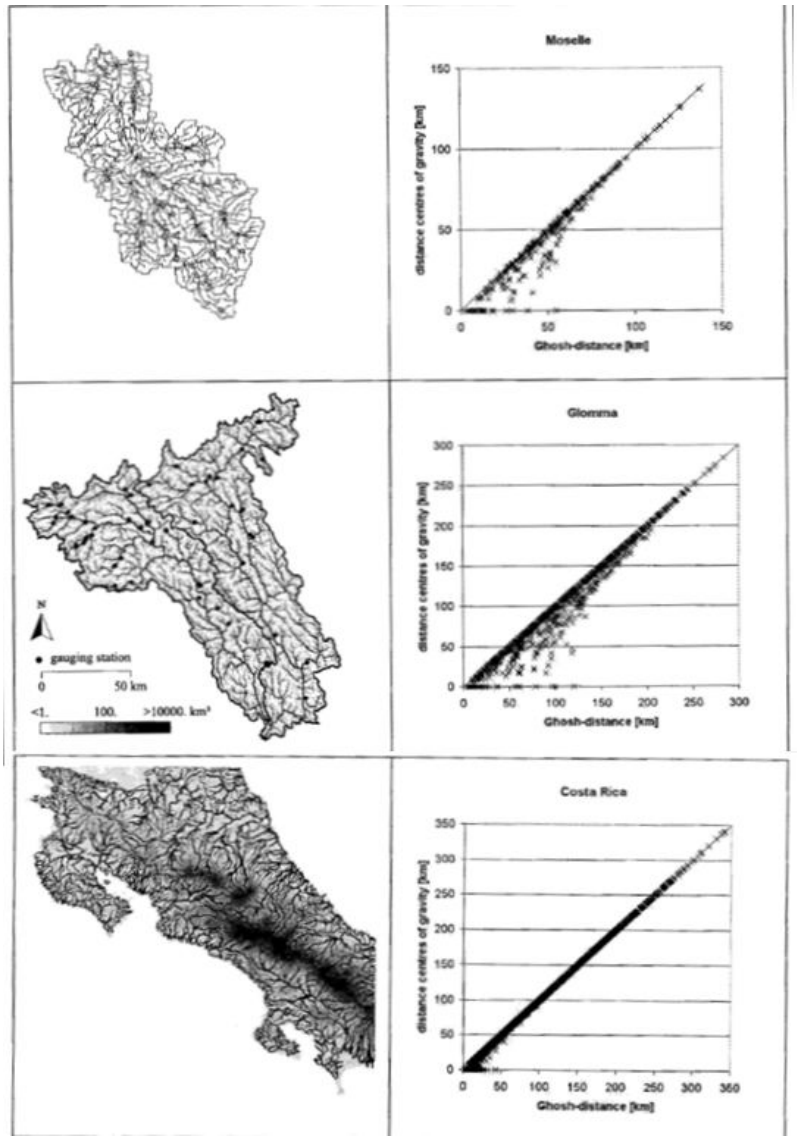

Figure 3.9: Relatioship between Gosh distances $m_\Lambda$and Euclidian distance between centres of gravity for the Glomma and Moselle basins and basins in Costa Rica

Chapter 4

Description of precipitation, temperature and runoff dataset

This chapter describes the characteristics of the dataset used for the application of spatial interpolation methods exposed in the two previous chapters. Starting with a short description of the geographical and climatic characteristics of Sicily island. In the second section, the description of the precipitation and temperature dataset has been made. In particular, a validation procedure used to detect the "uncertain data" within the recorded dataset of precipitation and temperature has been described and applied. In the third section the runoff dataset and the GIS techniques used to obtain chatchment areas, i.e. the spatial supports of runoff data, have been presented.

4.1 Sicily region

Sicily is the greatest island in the Mediterranean sea, with a total surface of about 25.000 km^2. Its geographical coordinates are 36° and 38° north latitude and 12° e 15° east longitude. Despite the complex orographic profile, it is possible to classify the region into three different zones: the northern zone, extending from *Capo Peloro* to *Capo Lilibeo*; southern zone, extending from *Capo Lilibeo* to *Capo Passero* and finally the western zone, extending from *Capo Passero* to *Capo Peloro*. The Sicily orography shows non homogeneity among the different zones. In particular, the northern zone is prevalently mountainous, the middle-southern zone and south-western zone are prevalently hill, the south-eastern zone is dominated by plateus and the eastern zone

has vulcanic characteristics. The zone with the prevalence of mountains is that located in the Tirrenian area, where the mountain chain is considered as the extention of the *Appennino Calabro*, while the eastern extremity of chain includes the Peloritani montains, with steep slopes originating narrow and deep valleys. Going toward west, there is the Nebrodi montain chain, with gentle peak and less steep slopes if compared with the previous ones. In the middle and western regions the following montain groups are present: *Madonie*, *Trabia*, *Palermo* and *Trapani* montains. In the inland the *Sicani* montains are present. In the south of the northern montain chain the landscape is different with prevalence of hills presenting the marks due to rivers that in some cases show the aspect of geological instability. The eastern region of Sicily is characterized by the *Etna Volcano*, that is located in the Catania plain, while the south-eastern region is characterized by the *Ibleo plateu*. The plain areas of the island are as a whole the 7% of the entire region; in particular they are the *alluvional plain of Catania*, the *coast plain* of *Licata* and *Gela*, and the *coast plain* near to *Trapani* and that comprised between *Siracusa* and *Scicli* (*Iblei mountain*). The overall island average elevation is around 400 m, but the range of variation of the elevation is between 0 and 3,263 m of *Etna Volcano*, that is the highest peak of Sicily and the highest volcano in Europe.

Considering the average conditions of the entire region, Sicily can be defined as a region with a wet-mild climate or in other words, mesotermic-wet subtropical climate, with a dry summer, i.e. the typical weather of the Mediterranean area, with an average temperature in the hottest month greater than 22 °C and with a precipitation regime more intense in the coldest season. According to Pinna (1978), within Sicilian context, it is possible to distinguish different typologies of climate: mild subtropical , mild-hot , mild subcoastal, mild subcontinental, mild-cool. In particular, the average annual temperature varies between 11°C and 20°C, depending on the considered zone. On the other hand, the total annual precipitation varies between 400 and 1200 mm, with an average values equal to 700 mm (period 1921–2004). Precipitations are concentrated in the winter period while the July–August months are usually rainless.

4.2 Analysis of data: precipitation and temperature dataset

The rainfall and temperature dataset used in this study has been provided by OA-ARRA (*Osservatorio delle Acque - Agenzia Regionale dei Rifiuti e delle Acque* the

formerly *Ufficio Idrografico Regionale* - UIR.This datasets belong to two types: the total monthly data (rainfall and temperature), measured in each gauging station and in each operation year and the total annual data (rainfall). In order to make these data usable to the application of the methods used in this work, it is necessary to calculate the average monthly and annual values of the rainfall and temperature for each gauging station; then the data have to be processed. The above mentioned average methods are computed in rigorouse way, i.e. using only the years when the single gauging station has been operated for all year. The first step is the setup of the consistency tables (in Annex 1), where, for each year, the months of operation of each station are indicated. For the precipitation data, the gauging stations that have operated from 1916 to 2004 are 711 in the whole Sicily region; among these, 157 have worked for a period less than one entire year. For the temperature datathe gauging stations that have operated from 1924 to 2006 are 303 in the whole Sicily region.

Unfortunately the source data may suffer from several errors due to incorrect recorded of data or measurement error. It was necessary, therefore, to check the recorded data and test their reliability. In particular, a validation procedure, according to the methodology developed by Campisano et al. (2002), has been applied to the original database to identify the "uncertain data" in the dataset. This validation used procedure is based on a dual control: a control *at site* and a control in the *spatial pattern* or *space* control.

The first control assesses whether a data measured at a station falls within a certain range of probability, that is a function of the history of that station. Campisano et al. (2002), for the building of confidence bands for the *at site* control, have tested the goodness of fit of normal distributions, lognormal and normal to the cube roots at the time series of precipitations and temperatures monthly average data of the stations in Sicily. The estimation of the parameters of the distributions was made using the method of maximum likelihood and the goodness of fit of the distributions was assessed by the Pearson test with a significance level of 5%. In the case of precipitation and temperatures the authors have chosen normal distribution, taking into account that the precipitation, alike temperature, can have zero values with a non-zero probability. In this study, the values of monthly average precipitation and monthly average temperatures have been created using Matlab scripts, for each station. Then, a control has been made to verify if the single monthly value fell within the range of probability to 95% and 99 % for each station. The data did not comply with this control were classified as outliers. In Table 4.1 the identification code for each temperature station, the name of the station, the number of the "uncertain data",

the total number of recorded data in that station, and the percentage with probability of 95% and 99%.

For the *at site*and *spatial pattern*controls, only the tables related to the data from temperature station are shown, for the sake of brevity.

The task of spatial control, however, is to identify whether the data measured at a station is congruent with contemporary measurements of neighboring stations. For the construction of confidence bands in the space control, we have examined, for each station, the two most related to each other, choosing among the five closest stations. This is due to the fact that not always the nearest stations are also the most related, since the morphology and other micro-climatic conditions may cause changes in precipitation and temperature even at small scale. The correlation was made on the basis of the common years of operation, to prevent a station during the year, could be correlated with a station for a few months and for several months with another. Once this information is obtained, it was possible to implement the space control with Matlab.

A linear relationship between the considered station and the contemporaneous measures in the two reference stations has been assumed, such as:

$$Z_{i,j} = b_j^{(0)} + b_j^{(1)} Z_{i,j}^{(1)} + b_j^{(2)} Z_{i,j}^{(2)} \tag{4.1}$$

where $Z_{i,j}$, *i = 1, 2, ...,n* is the vector of observations in n operation years of a station for a generic month j, (1) and (2) identifies the two neighboring stations, $Z_{i,j}^{(1)}$, $Z_{i,j}^{(2)}$ are the measurements in the two reference stations for fixed months j, $b_j^{(0)}$, $b_j^{(1)}$, $b_j^{(2)}$ are the parameters of linear regression.

In particular, called the vector $\overline{Y_j}$ of the station taken into account as:

$$\overline{Y_j} = [Z_{1,j}, Z_{2,j}, \ldots\ldots, Z_{n,j}] \tag{4.2}$$

and the matrix of contemporary reference measurements $\overline{X_j}$ of size $(n * (k+1))$ as:

$$\begin{bmatrix} 1 & Z_{1,j}^{(1)} & Z_{1,j}^{(2)} \\ 1 & Z_{2,j}^{(1)} & Z_{2,j}^{(2)} \\ \vdots & \vdots & \vdots \\ 1 & Z_{n,j}^{(1)} & Z_{n,j}^{(2)} \end{bmatrix} \tag{4.3}$$

it is possible get the vector of coefficients through the matrix relationship proposed by Kottegoda and Red (1997). In particular, we have:

At site control							
CODE	NAME	95%			99%		
		Uncertain data	N. data	perc.	Uncertain data	N. data	perc.
3150	ACIREALE	42	824	5,10	11	824	1,33
2900	ADRANO	16	375	4,27	1	375	0,27
1810	AGRIGENTO	49	980	5,00	9	980	0,92
1100	BIRGI NUOVO	9	223	4,04	1	223	0,45
1480	BIVONA	42	827	5,08	17	827	2,06
3050	CALTAGIRONE	33	721	4,58	4	721	0,55
2020	CALTANISSETTA	46	786	5,85	7	786	0,89
2025	CALTANISSETTA (GenioCivile)	6	213	2,82	0	213	0,00
620	CAMPOFELICE DI FITALIA	7	183	3,83	1	183	0,55
1840	CANICATTI'	10	270	3,70	0	270	0,00
1050	CAPO S.VITO	35	548	6,39	11	548	2,01
320	CARONIA	11	231	4,76	2	231	0,87
1570	CASE MUSTIGARUFI	11	251	4,38	1	251	0,40
1190	CASTELVETRANO	18	520	3,46	7	520	1,35
3160	CATANIA (Ist. Agrario)	29	714	4,06	8	714	1,12
3155	CATANIA (Osservatorio)	16	322	4,97	3	322	0,93
2980	CATENANUOVA	13	225	5,78	2	225	0,89
470	CEFALU'	36	674	5,34	14	674	2,08
2690	CESARO'	26	505	5,15	4	505	0,79
660	CIMINNA	41	835	4,91	18	835	2,16
1270	CORLEONE	40	845	4,73	7	845	0,83
2420	COZZO SPADARO	34	690	4,93	9	690	1,30
2120	DELIA	10	251	3,98	0	251	0,00
2230	DIGA COMUNELLI	12	237	5,06	1	237	0,42
2260	DIGA DISSUERI	9	366	2,46	0	366	0,00
3010	DIGA DON STURZO	12	238	5,04	1	238	0,42
1520	DIGA FANACO	14	267	5,24	0	267	0,00
2140	DIGA GIBBESI	8	278	2,88	0	278	0,00
1240	DIGA MAGANOCE	13	228	5,70	2	228	0,88
2310	DIGA RAGOLETO	14	290	4,83	4	290	1,38
1070	DIGA RUBINO	16	354	4,52	1	354	0,28
1960	ENNA	38	771	4,93	9	771	1,17
990	ERICE	7	253	2,77	1	253	0,40
720	FICUZZA	35	739	4,74	6	739	0,81
3210	FLORESTA	40	843	4,74	13	843	1,54
1930	GANGI	25	385	6,49	11	385	2,86
3400	GANZIRRI	29	619	4,68	12	619	1,94
2240	GELA	21	587	3,58	3	587	0,51
570	GIOIA	22	420	5,24	3	420	0,71
930	ISOLA DELLE FEMMINE	22	444	4,95	3	444	0,68
1770	LAGO GORGO	22	261	8,43	7	261	2,68
2580	LENTINI	24	822	2,92	12	822	1,46
1490	LERCARA FRIDDI	36	781	4,61	6	781	0,77
2200	LICATA	37	733	5,05	9	733	1,23
3120	LINGUAGLOSSA	35	805	4,35	8	805	0,99
1120	MARSALA	36	819	4,40	11	819	1,34
1140	MAZARA DEL VALLO	31	756	4,10	17	756	2,25
2210	MAZZARINO	40	701	5,71	10	701	1,43
3380	MESSINA (Ist. Geofisico)	35	672	5,21	7	672	1,04
3370	MESSINA (Osservatorio)	45	840	5,36	5	840	0,60
3060	MINEO	42	844	4,98	8	844	0,95
800	MONREALE	32	713	4,49	3	713	0,42
2290	MONTEROSSO ALMO	25	512	4,88	8	512	1,56
3090	NICOLOSI	30	761	3,94	8	761	1,05
870	PALERMO (Ist. Zootecnico)	30	612	4,90	6	612	0,98
910	PALERMO (Ist.Castelnuovo)	31	507	6,11	5	507	0,99
880	PALERMO (Oss.Astronomico)	43	800	5,38	7	800	0,88
920	PALERMO (Piazza Verdi)	31	718	4,32	17	718	2,37
1180	PARTANNA	47	794	5,92	15	794	1,89
850	PARTINICO	39	749	5,21	7	749	0,93
1860	PETRALIA SOTTANA	36	720	5,00	13	720	1,81
1210	PIANA DEGLI ALBANESI	8	220	3,64	1	220	0,45
1420	PIANO DEL LEONE	25	671	3,73	8	671	1,19
1580	PIANO FALZONE	8	225	3,56	0	225	0,00
2250	PIAZZA ARMERINA	31	572	5,42	7	572	1,22
3140	PIEDIMONTE ETNEO	20	446	4,48	3	446	0,67
1740	PIETRANERA	12	297	4,04	0	297	0,00
1700	RACALMUTO	36	563	6,39	10	563	1,78
2370	RAGUSA	38	802	4,74	7	802	0,87
1460	RIBERA	10	293	3,41	2	293	0,68
740	RISALAIMI	16	394	4,06	1	394	0,25
310	S. FRATELLO	22	445	4,94	2	445	0,45
950	S. GIUSEPPE JATO	42	793	5,30	16	793	2,02
1350	S. MARGHERITA BELICE	17	248	6,85	3	248	1,21
1400	SCIACCA	42	868	4,84	15	868	1,73
520	SCILLATO	18	284	6,34	3	284	1,06
2540	SIRACUSA	41	746	5,50	9	746	1,21
3310	TAORMINA	38	829	4,58	6	829	0,72
140	TINDARI	24	578	4,15	6	578	1,04
1030	TRAPANI	54	972	5,56	22	972	2,26
3130	VIAGRANDE	43	756	5,69	8	756	1,06
2530	VILLASMUNDO	9	238	3,78	1	238	0,42
2350	VITTORIA	47	742	6,33	8	742	1,08
3110	ZAFFERANA ETNEA	23	541	4,25	7	541	1,29

Table 4.1: Results of the *at site* control. Temperature station

$$\overline{b_j} = \frac{\overline{X_j}^T \overline{Y_j}}{\overline{X_j}^T \overline{X_j}} \tag{4.4}$$

Once the vector of coefficients has been obtained, it is possible to derived for each month and each year an estimated value of monthly average precipitation and monthly average temperature at the station by the equation 4.1, in fact we have $\hat{Z}_{i,j} = \overline{Z}_{i,j} * \overline{b}_j$. Then a range of probabilities of the single monthly mean temperature has been obtained through the following relationship:

$$l_{\alpha/2}, l_{1-\alpha/2} = \hat{Z}_{i,j} \pm t_{n-p,\alpha/2} * \hat{\sigma} * \sqrt{1 + \frac{\overline{Z}_{i,j} \overline{Z}_{i,j}^T}{\overline{X_j}^T \overline{X_j}}} \tag{4.5}$$

where σ is the variance of regression residuals and $t_{n-p,\alpha/2}$ is the quantile of a t-student variable corresponding to a probability of no exceedance equal to $\alpha/2$.

Also for the space control probability intervals of 95 and 99% has been made for both variables (precipitation and temperature). In Table 4.2 the identification code for each temperature station, the name of the station, the number of the "uncertain data", the total number of recorded data in that station and the percentage with probability of 95% and 99%.

In Figures 4.1 and 4.2, the number of the operating stations for each years are shown for precipitation and temperature, respectively. It should be observed that in the period of the second war world and subsequently up to 1950 the number of the operating function is extremely low.

Moreover, the World Meteorological Organization (WMO) recommends to use monthly and annual average values determinated by stations with at least 30 years of operation. The database, here used without the "uncertain data" coming from the stations that had been worked for more of 30 years are 247 for the raingauging station and 60 for the temperature stations. In the application of interpolation methods, 84 temperature stations are used to have an adequately and well spatial distributed data set. The latter stations have worked for a period of at least 20 years. Finally the considered gauging stations are show in Figure 4.3 for the precipitation dataset (247 gauging stations) and in Figure for the temperature dataset (84 gauging stations) by graduate symbol maps.

Space control							
CODE	NAME	95%			99%		
		Uncertain data	N. data	perc.	Uncertain data	N. data	perc.
3150	ACIREALE	50	824	6.07	15	824	1,82
2900	ADRANO	4	375	1,07	1	375	0,27
1810	AGRIGENTO	19	980	1,94	2	980	0,20
1100	BIRGI NUOVO	4	223	1,79	0	223	0,00
1480	BIVONA	22	827	2,66	16	827	1,93
3050	CALTAGIRONE	25	721	3,47	6	721	0,83
2020	CALTANISSETTA	0	786	0.00	0	786	0,00
2025	CALTANISSETTA (GenioCivile)	0	213	0,00	0	213	0,00
620	CAMPOFELICE DI FITALIA	3	183	1,64	0	183	0,00
1840	CANICATTI'	7	270	2,59	0	270	0,00
1050	CAPO S.VITO	26	548	4,74	9	548	1,64
320	CARONIA	4	231	1,73	0	231	0,00
1570	CASE MUSTIGARUFI	4	251	1,59	0	251	0,00
1190	CASTELVETRANO	24	520	4,62	4	520	0,77
3160	CATANIA (Ist. Agrario)	24	714	3,36	8	714	1,12
3155	CATANIA (Osservatorio)	3	322	0,93	0	322	0,00
2980	CATENANUOVA	5	225	2,22	0	225	0,00
470	CEFALU'	28	674	4,15	8	674	1,19
2690	CESARO'	22	505	4,36	5	505	0,99
660	CIMINNA	29	835	3,47	16	835	1,92
1270	CORLEONE	37	845	4,38	5	845	0,59
2420	COZZO SPADARO	26	690	3,77	5	690	0,72
2120	DELIA	5	251	1,99	0	251	0,00
2230	DIGA COMUNELLI	2	237	0,84	0	237	0,00
2260	DIGA DISSUERI	5	366	1,37	0	366	0,00
3010	DIGA DON STURZO	9	238	3,78	0	238	0,00
1520	DIGA FANACO	6	267	2,25	3	267	1,12
2140	DIGA GIBBESI	1	278	0,36	0	278	0,00
1240	DIGA MAGANOCE	5	228	2,19	1	228	0,44
2310	DIGA RAGOLETO	8	290	2,76	2	290	0,69
1070	DIGA RUBINO	10	354	2,82	1	354	0,28
1960	ENNA	31	771	4,02	7	771	0,91
990	ERICE	5	253	1,98	1	253	0,40
720	FICUZZA	11	739	1,49	2	739	0,27
3210	FLORESTA	31	843	3,68	15	843	1,78
1930	GANGI	14	385	3,64	9	385	2,34
3400	GANZIRRI	29	619	4,68	12	619	1,94
2240	GELA	14	587	2,39	1	587	0,17
570	GIOIA	23	420	5,48	8	420	1,90
930	ISOLA DELLE FEMMINE	33	444	7,43	15	444	3,38
1770	LAGO GORGO	14	261	5,36	1	261	0,38
2580	LENTINI	22	822	2,68	11	822	1,34
1490	LERCARA FRIDDI	27	781	3,46	15	781	1,92
2200	LICATA	21	733	2,86	3	733	0,41
3120	LINGUAGLOSSA	28	805	3,48	4	805	0,50
1120	MARSALA	34	819	4,15	10	819	1,22
1140	MAZARA DEL VALLO	30	756	3,97	11	756	1,46
2210	MAZZARINO	24	701	3,42	16	701	2,28
3380	MESSINA (Ist. Geofisico)	20	672	2,98	5	672	0,74
3370	MESSINA (Osservatorio)	20	840	2,38	7	840	0,83
3060	MINEO	26	844	3,08	2	844	0,24
800	MONREALE	25	713	3,51	2	713	0,28
2290	MONTEROSSO ALMO	24	512	4,69	16	512	3,13
3090	NICOLOSI	33	761	4,34	11	761	1,45
870	PALERMO (Ist. Zootecnico)	21	612	3,43	2	612	0,33
910	PALERMO (Ist.Castelnuovo)	19	507	3,75	5	507	0,99
880	PALERMO (Oss.Astronomico)	26	800	3,25	12	800	1,50
920	PALERMO (Piazza Verdi)	23	718	3,20	19	718	2,65
1180	PARTANNA	24	794	3,02	6	794	0,76
850	PARTINICO	23	749	3,07	6	749	0,80
1860	PETRALIA SOTTANA	17	720	2,36	11	720	1,53
1210	PIANA DEGLI ALBANESI	1	220	0,45	0	220	0,00
1420	PIANO DEL LEONE	37	671	5,51	14	671	2,09
1580	PIANO FALZONE	4	225	1,78	3	225	1,33
2250	PIAZZA ARMERINA	6	572	1,05	0	572	0,00
3140	PIEDIMONTE ETNEO	19	446	4,26	4	446	0,90
1740	PIETRANERA	11	297	3,70	1	297	0,34
1700	RACALMUTO	4	563	0,71	0	563	0,00
2370	RAGUSA	31	802	3,87	2	802	0,25
1460	RIBERA	9	293	3,07	3	293	1,02
740	RISALAIMI	16	394	4,06	4	394	1,02
310	S. FRATELLO	16	445	3,60	7	445	1,57
950	S. GIUSEPPE JATO	26	793	3,28	5	793	0,63
1350	S. MARGHERITA BELICE	11	248	4,44	0	248	0,00
1400	SCIACCA	36	868	4,15	8	868	0,92
520	SCILLATO	4	284	1,41	0	284	0,00
2540	SIRACUSA	31	746	4,16	11	746	1,47
3310	TAORMINA	29	829	3,50	6	829	0,72
140	TINDARI	4	578	0,69	1	578	0,17
1030	TRAPANI	14	972	1,44	2	972	0,21
3130	VIAGRANDE	40	756	5,29	20	756	2,65
2530	VILLASMUNDO	9	238	3,78	0	238	0,00
2350	VITTORIA	31	742	4,18	4	742	0,54
3110	ZAFFERANA ETNEA	22	541	4,07	5	541	0,92

Table 4.2: Results of the *space* control. Temperature stations.

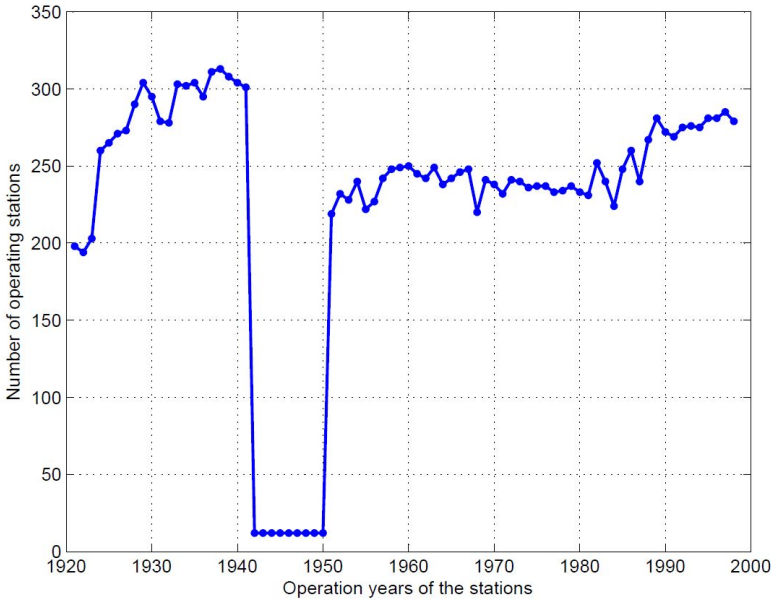

Figure 4.1: Operation years of raingauging stations

Figure 4.2: Operation years of temperature stations

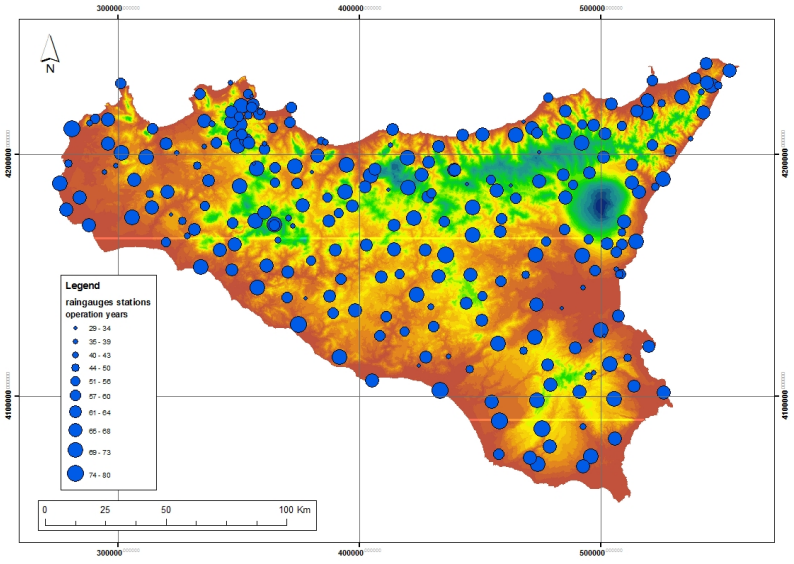

Figure 4.3: Raingauge station used in the study and their operation years

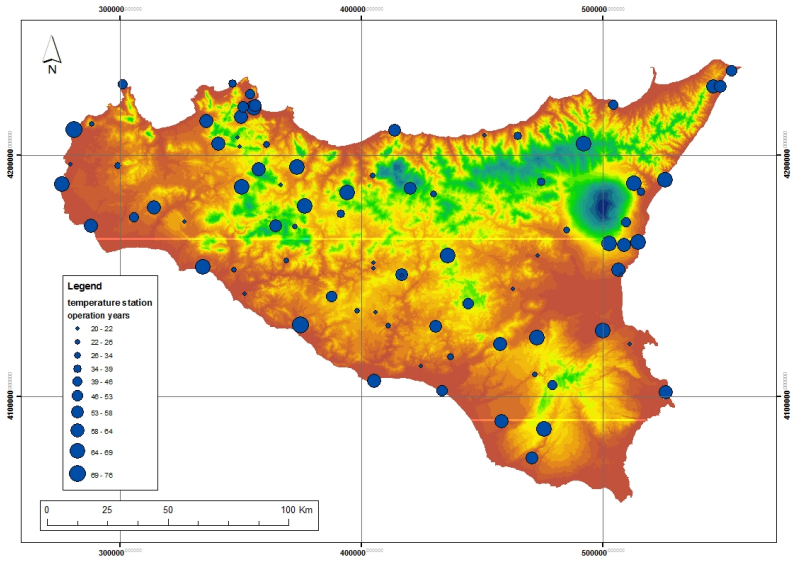

Figure 4.4: Temperature station used in the study and their operation years

4.3 Analysis of data: runoff dataset

In this study, the runoff dataset used are provided by OA-ARRA as well.The Hydrological Annals, in section C (streamflow and water balances), provides a table for each station and for each year that consists of several parts. First, the characteristics of the station and its catchment area (area of the basin, the percentage of permeable area, maximum and average altitude, elevation of the hydrometric zero and distance from the outlet or tributary), the daily hydrometric levels, the maximum and minimum streamflow of the entire period of the observations are reported.

The dataset used in this study come from 105 hydrometric stations distributed throughout Sicily.

In order to obtain a reliable database of runoff, in the process of data retrieval, the monthly and annual streamflows recorded in the period ranging from 1923 to 2002, for a total of 80 years observed , have been minutely examined. In Table 4.3 a list of the hydrometric stations distributed over the Sicily is reported.

In order to ensure greater reliability for further analysis, the instructions of the Bulletin 17B (*Guidelines for Determining Flood Flow Frequency drawn from 'Department of the Interior U.S. Geological Survey*) have been followed. The gauge stations that have worked for less than 10 years have been removed; so the initially available

ID	Station	ID	Station	ID	Station
1	Rosmarino at Passo Gallo	38	Belice at Belice	75	Tellaro at Mandrevecchie
2	Pollina at Aquileia	39	Magazzolo at Corvo	76	Cassibile at Manghisi
3	Elicona at Falcone	40	Sosio at S.Carlo	77	Anapo at Diddino
4	Timeto at Murmari	41	Belici at Marionopoli scalo	78	Anapo at passo di Siracusa
5	Torrente dei Mulini at Guglielmotto	42	Belici at Bruciato	79	Irminio at S.Rosalia
6	Torto at Roccapalumba Scalo	43	Carboi at Arancio	80	Asianaro at Noto
7	Isnello at ponte grande	44	Verdura at Sosio	81	Anapo at S.Nicola
8	Castelbuono at ponte vecchio	45	Verdura at Poggidiana	82	Trigona at Rappis
9	Imera settentrionale at Scillato	46	Platani at Passofonduto 1	83	Zena at Reina
10	Torto at Bivio Cerda	47	Platani at Passofonduto 2	84	Cutrò at Vitalone
11	S. Leonardo at Vicari	48	Platani at Ganzeria	85	Martello at Petrosino
12	S. Leonardo at Vecchio	49	Platani at Platani	86	Saraceno at Chiusitta
13	S. Leonardo at Monumentale	50	Ipsas at S.Anna	87	Troina di sopra at Serravalle
14	Milicia at Milicia	51	S. Biagio at Mandorleto	88	Simeto at Biscari
15	Eleuterio at lupo	52	Palma at Mandranova	89	Simeto at Don Gennaro
16	Valle dell"acqua at Serena	53	Grancifone at La Loggia	90	Simeto at Giarretta
17	Eleuterio at Risalaimi 1	54	Alberi at Irosa	91	Simeto at ponte Maccarone
18	Eleuterio at Risalaimi 2	55	Castello at Castello	92	Salso at ponte Gagliano
19	Eleuterio at Rossella	56	Imera meridionale at Imera	93	Girgia at case Celso
20	Oreto at Parco	57	Imera meridionale at Petralia	94	Dittaino at Bozzetta
21	Nocella at Zucco	58	Imera meridionale at Cinque Archi	95	Dittaino at Stempato
22	Jato at Taurro	59	Imera meridionale at Capodarso	96	Crisà at case Carella
23	Jato at Fellamonica	60	Gangi at regiovanni	97	Sciaguana at Torricchia 1
24	Freddo at alcamo scalo	61	Salso at Raffo	98	Sciaguana at Torricchia 2
25	Forgia at Lentina	62	Salso at Monzanaro	99	Simeto at Sommaruga
26	Baiata at Sapone	63	Gibbesi at Donnapaola	100	Gornalunga at Secreto
27	Fastaia at La Chinea 1	64	Imera meridionale at Drasi	101	Simeto at Gornalunga
28	Fastaia at La Chinea 2	65	Imera meridionale at Besero	102	Gornalunga at Libertini
29	Chitarra at Rinazzo	66	Gattano at Zai	103	Simeto at Monaci
30	Birgi at Chinisia	67	Cimia at Cerasaro	104	Flascio at Zarbata
31	Delia at Pozzillo	68	Gela at Dissueri	105	Flascio at Acquasanta
32	Modione at S.Elia	69	Ficuzza at S. Pietro	106	Flascio at ponte flascio
33	Belice destro at sparacia	70	Valle torta at Cubaitaro	107	Alcantara at Moio
34	Corleone at piano scala	71	Para para at Mazzaronello 1	108	Alcantara at Alcantara
35	Belice sinistro at casa Balate	72	Para para at Mazzaronello 2	109	Alcantara at S.Giacomo 1
36	Senore at Senore	73	Dirillo at Dirillo	110	Alcantara at S.Giacomo 2
37	Senore at Finocchiara	74	Tellaro at Castelluccio	111	Forza D'Agrò at Ranciara

Table 4.3: Hydrometric stations distributed throughout Sicily region

ID	Station	ID	Station	ID	Station
1	Pollina at Aquileia	24	Fastaia at La Chinea	47	Asianaro at Noto
2	Elicona at Falcone	25	Chitarra at Rinazzo	48	Anapo at S.Nicola
3	Timeto at Murmari	26	Birgi at Chinisia	49	Trigona at Rappis
4	Torrente dei Mulini at Guglielmotto	27	Delia at Pozzillo	50	Zena at Reina
5	Torto at Roccapalumba Scalo	28	Belice destro at sparacia	51	Martello at Petrosino
6	Isnello at ponte grande	29	Belice sinistro at casa Balate	52	Saraceno at Chiusitta
7	Castelbuono at ponte vecchio	30	Senore at Finocchiara	53	Troina di sopra at Serravalle
8	Imera settentrionale at Scillato	31	Belice at Belice	54	Simeto at Biscari
9	Torto at Bivio Cerda	32	Belici at Marionopoli scalo	55	Simeto at Don Gennaro
10	S. Leonardo at Vicari	33	Belici at Bruciato	56	Simeto at Giarretta
11	S. Leonardo at Monumentale	34	Platani at Passofonduto 1	57	Salso at ponte Gagliano
12	Milicia at Milicia	35	S. Biagio at Mandorleto	58	Girgia at case Celso
13	Eleuterio at lupo	36	Castello at Castello	59	Dittaino at Bozzetta
14	Valle dell'acqua at Serena	37	Imera meridionale at Petralia	60	Crisà at case Carella
15	Eleuterio at Risalaimi 1	38	Imera meridionale at Cinque Archi	61	Sciaguana at Torricchia
16	Eleuterio at Rossella	39	Imera meridionale at Capodarso	62	Gornalunga at Secreto
17	Oreto at Parco	40	Gangi at regiovanni	63	Flascio at Zarbata
18	Nocella at Zucco	41	Salso at Raffo	64	Alcantara at Moio
19	Jato at Taurro	42	Salso at Monzanaro	65	Alcantara at Alcantara
20	Jato at Fellamonica	43	Gibbesi at Donnapaola	66	Alcantara at S.Giacomo
21	Freddo at alcamo scalo	44	Imera meridionale at Drasi		
22	Forgia at Lentina	45	Ficuzza at S. Pietro		
23	Baiata at Sapone	46	Cassibile at Manghisi		

Table 4.4: Hydrometric station with operation years greater or equal to 10 years

hydrographic information has reduced from 105 to 67 stations. It is reported in Table 4.4 a list of the 67 hydrometric stations.

To the application of the interpolation method that will be used for the runoff estimation, a fundamental step is the derivation of chatchment areas, i.e. the spatial supports of runoff data. From the geographical coordinates of the hydrometric stations outlet, the water catchment areas are marked off through the use of GIS thecniques (Geographic Information System). A DEM (Digital Elevation Model) of the Sicily region, with 100 m resolution (Figure 4.5) has been use.

In order to obtain data that are as homogeneous as possible, a check on the history of the stations has been carried out to verify the presence of possible inhomogeneity which have affected the series. From that review it was possible to obtain all the information below.

In 1991 in the ELICONA at FALCONE there was a lowering of the hydrometric zero equal to 21 m a.l.s. without substantial effect on the treatment of the series.

The area of the Eleuterio at Risalaimi station was of 79.50 km^2 , but with the entry into operation of the Scanzano reservoir (1964) the area was reduced to 52.9 km^2. For this reason, the data are separated and treated separately in the following periods: 1961-1964 and 1965-1990.

Since 1930 the area of the Oreto at Parco station turns 76 km^2 while other characteristics remain unchanged; since 1934 there is an increase in the hydrometric zero of 12.58 m a.l.s..

In 1972, the station Fastaia at La Chinea has been moved upstream, following the

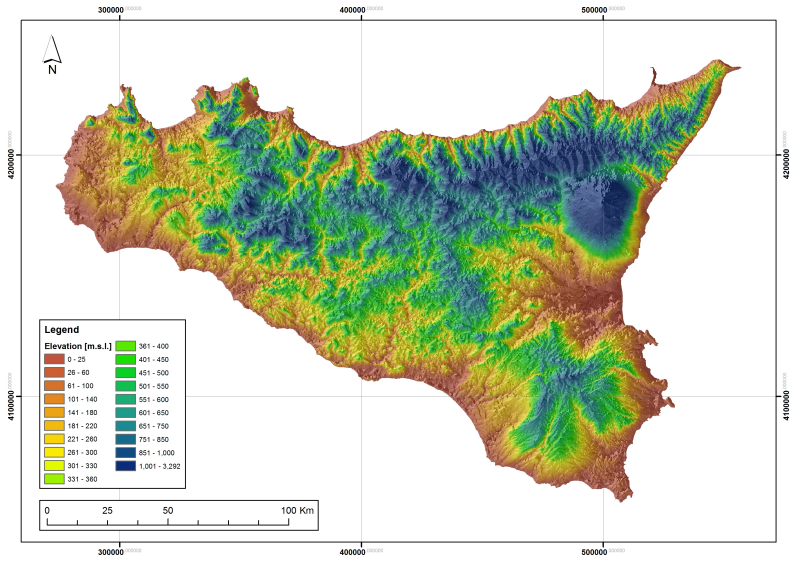

Figure 4.5: DEM of Sicily region

entry into operation of the Rubino reservoir (1972), so the basin area was reduced to 22.6 km^2. For this reason the data is divided and treated in different way in the periods 1962-1971 and 1972-1997.

In 1972, the Birgi at Chinisia station has been reduced to 292.60 km2 for the presence of the Rubino reservoir (1971). The data are treated uniformly excluded from the series after the year 1971.

Since 1985, the Platani at Passofonduto station has been reduced to 1186 km^2 for the construction of the Fanaco reservoir; a division of data in the periods 1956-1984 and 1985-1994 has been realized.

Since 1983, theS. Biagio at Mandorleto station has been reducedfrom 80.8 to 74 km^2;

Since 1983, the Simeto a Biscari station was moved upstream 2.5 km; with the entry into operation of Nicoletti reservoir (1964) the area was reduced to49 km2 o; we therefore decided to exclude from the analysis the data after 1983.

In 1974, the Sciaguana a Torricchia station there was a lowering of the hydrometric zero to 5 m s. m. In 1975 the station was moved to the valley and the basins area increases. The data from the 1974 are delete and divided into two distinct sets characterized by two different surface domain: 1969-1973 and 1975-1989.

In 1983 the Alcantara at San Giacomo station was moved to the valley of 0.5 km

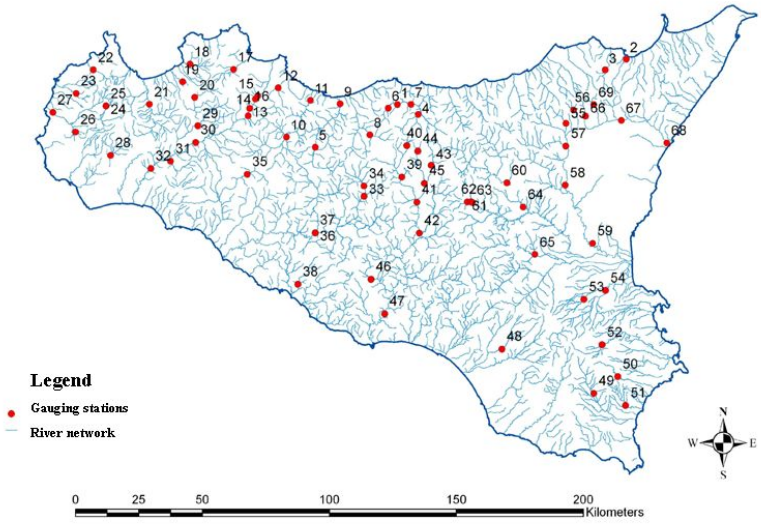

Figure 4.6: River network of Sicily with streamgauges

with consequent increase in the area of the basin, it is appropriate to distinguish two distinct sets of data in the periods 1926-1982 and 1983-1997.

The number of stations is therefore 69 and not more than 67, since the series of Fastaia at La Chinea and Platani at Passofonduto stations have been divided into two independent series. In Figure 4.6 the hydrographic network is shown, in Figure 4.7 the location of the catchment is shown , while the list of the 69 stations with information about the main river, the years of operation, the year of the beginning and end of the observations and the coordinates of the stations are shown in Table 4.5 and 4.6.

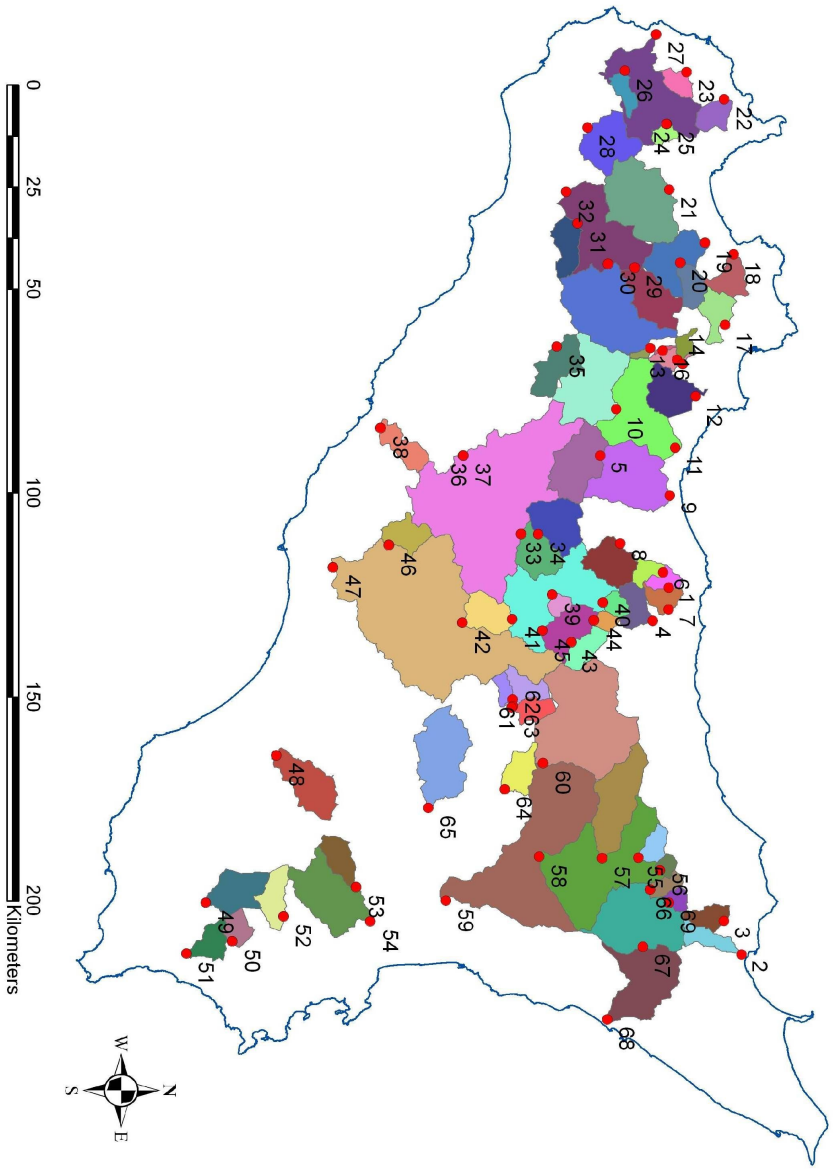

Figure 4.7: Cathcments areas

Legend:
1. Pollina at Aquileia;
2. Elicona at Falcone;
3. Timeto at Murmari;
4. Torrente Mulini at Guglielmotto;
5. Torto at Roccapalumba Scalo;
6. Isnello at Ponte Grande;
7. Castelbuono at Ponte Vecchio;
8. Imera Settentrionale at Scillato;
9. Torto at Bivio Cerda;
10. S. Leonardo at Vicari;
11. S. Leonardo at Monumentale;
12. Milicia at Milicia;
13. Eleuterio at Lupo;
14. Valle dell'Acqua at Serena;
15. Eleuterio at Risalaimi;
16. Eleuterio at Rossella;
17. Oreto at Parco;
18. Nocella at Zucco;
19. Jato at Taurro;
20. Jato at Fellamonica;
21. Freddo at Alcamo Scalo;
22. Forgia at Lentina;
23. Baiata at Sapone;
24. Fastaia at La Chinea 1;
25. Fastaia at la Chinea 2;
26. Chitarra at Rinazzo;
27. Birgi at Chinisia;
28. Delia at Pozzillo;
29. Belice Destro at Sparacia;
30. Belice Sinistro at Casa Balate;
31. Senore at Finocchiara;
32. Belice at Belice;
33. Belici at Marionopoli Scalo;
34. Belici at Bruciato;
35. Verdura at Sosio;
36. Platani at Passofonduto 1;
37. Platani at Passofonduto 2;
38. S. Biagio at Mandorleto;
39. Castello at Castello;
40. Imera Meridionale at Petralia;
41. Imera Merid. at Cinque Archi;
42. Imera Meridionale at Capodarso;
43. Gangi at Re Giovanni;
44. Salso at Raffo;
45. Salso at Monzanaro;
46. Gibbosi at Donnapaola;
47. Imera Meridionale at Drasi;
48. Ficuzza at S. Pietro;
49. Tellaro at Castelluccio;
50. Cassibile at Manghisi;
51. Asinaro at Noto;
52. Anapo at S. Nicola;
53. Trigona at Rappis;
54. Zena at Reina;
55. Martello at Petrosino;
56. Saraceno at Chiusitta;
57. Troina di Sopra at Serravalle;
58. Simeto at Biscari;
59. Simeto at Giarretta;
60. Salso at Ponte Gagliano;
61. Girgia at Case Celso;
62. Dittaino at Bozzetta;
63. Crisà at Case Carella;
64. Sciaguana at Torricchia; 2
65. Gornalunga at Secreto;
66. Flascio at Zarbata;
67. Alcantara at Moio;
68. Alcantara at Alcantara;
69. Alcantara at S. Giacomo.

ID	Station	River	Operation years	Y start	Y end
1	Pollina at Aquileia	Pollina	10	1952	1961
2	Elicona at Falcone	Elicona	18	1976	1996
3	Timeto at Murmari	Timeto	11	1976	1995
4	Torrente dei Mulini at Guglielmotto	Pollina	13	1983	1997
5	Torto at Roccapalumba Scalo	Torto	15	1983	1997
6	Isnello at ponte grande	Isnello	14	1984	1997
7	Castelbuono at ponte vecchio	Castelbuono	18	1978	1997
8	Imera settentrionale at Scillato	Imera settentrionale	21	1976	1997
9	Torto at Bivio Cerda	Torto	17	1969	1989
10	S. Leonardo at Vicari	S. Leonardo	27	1923	1987
11	S. Leonardo at Monumentale	S. Leonardo	55	1928	1984
12	Milicia at Milicia	Milicia	19	1976	1997
13	Eleuterio at lupo	Eleuterio	49	1936	1995
14	Valle dell"acqua at Serena	Eleuterio	35	1961	1996
15	Eleuterio at Risalaimi 2	Eleuterio	24	1965	1990
16	Eleuterio at Rossella	Eleuterio	14	1936	1957
17	Oreto at Parco	Oreto	75	1923	1997
18	Nocella at Zucco	Nocella	37	1958	1997
19	Jato at Taurro	Giancaldara	13	1955	1968
20	Jato at Fellamonica	Jato	16	1973	1997
21	Freddo at alcamo scalo	Freddo	10	1972	1987
22	Forgia at Lentina	Forgia	25	1971	1996
23	Baiata at Sapone	Baiata	23	1968	1997
24	Fastaia at La Chinea 1	Fastaia	10	1962	1971
25	Fastaia at La Chinea 2	Fastaia	24	1972	1997
26	Chitarra at Rinazzo	Chiatarra	17	1972	1988
27	Birgi at Chinisia	Birgi	20	1971	1997
28	Delia at Pozzillo	Delia	20	1959	1978
29	Belice destro at sparacia	Belice destro	33	1955	1987
30	Belice sinistro at casa Balate	Belice sinistro	26	1955	1980
31	Senore at Finocchiara	Belice	26	1961	1986
32	Belice at Belice	Belice	29	1955	1994
33	Belici at Marionopoli scalo	Belici	11	1984	1997
34	Belici at Bruciato	Belici	20	1972	1994
35	Verdura at Sosio	Verdura	13	1930	1942

Table 4.5: List of gauging stations - first part

Another important step that have to be made is the analysis of the presence of artificial reservoirs, as dams or diversion dams.

Sicily is one of the Italian regions with the largest number of dams built to meet the severe water shortages, to the fragility of the infrastructure system that oversees procurement and distribution of the water resource in Sicily. Below the river basins affected by the presence of dams and / or diversion dams are reported:

In the San Leonardo basin there is the Rosamarina artificial reservoir built in twenty years 1972-1992. During the construction period it does not intercept water, so the data are processed without any modifications;

In the Eleuterio at Risalaimi station, with the entry into operation of the Scanzano dam (1964), the original area equal to 79.50 $km^2$was reduced to 52.9 km^2. For this

ID	Station	River	Operation years	Y start	Y end
36	Platani at Passofonduto 1	Platani	22	1956	1980
37	Platani at Passofonduto 2	Platani	10	1985	1994
38	S. Biagio at Mandorleto	S. Biagio	26	1968	1997
39	Castello at Castello	Castello	15	1983	1997
40	Imera meridionale at Petralia	Imera meridionale	24	1971	1997
41	Imera meridionale at Cinque Archi	Imera meridionale	16	1960	1988
42	Imera meridionale at Capodarso	Imera meridionale	44	1923	1996
43	Gangi at regiovanni	Gangi	16	1978	1996
44	Salso at Raffo	Salso	16	1979	1997
45	Salso at Monzanaro	Salso	14	1983	1997
46	Gibbesi at Donnapaola	Gibbesi	15	1971	1992
47	Imera meridionale at Drasi	Imera meridionale	36	1960	1997
48	Ficuzza at S. Pietro	Ficuzza	17	1974	1994
49	Tellaro a Castelluccio	Tellaro	17	1974	1997
50	Cassibile at Manghisi	Cassibile	17	1974	1997
51	Asianaro at Noto	Asinaro	12	1984	1997
52	Anapo at S.Nicola	Anapo	14	1973	1997
53	Trigona at Rappis	Trigona	26	1972	1997
54	Zena at Reina	Zeina	11	1972	1984
55	Martello at Petrosino	Martello	10	1972	1981
56	Saraceno at Chiusitta	Saraceno	14	1981	1995
57	Troina di sopra at Serravalle	Troina	15	1982	1997
58	Simeto at Biscari	Simeto	20	1975	1997
59	Simeto at Giarretta	Simeto	38	1923	1967
60	Salso at ponte Gagliano	Simeto	21	1975	1997
61	Girgia at case Celso	Simeto	21	1958	1980
62	Dittaino at Bozzetta	Simeto	18	1950	1968
63	Crisà at case Carella	Simeto	25	1958	1986
64	Sciaguana at Torricchia	Simeto	11	1975	1989
65	Gornalunga at Secreto	Gornalunga	10	1957	1966
66	Flascio at Zarbata	Flascio	16	1981	1997
67	Alcantara at Moio	Alcantara	34	1939	1995
68	Alcantara at Alcantara	Alcantara	32	1934	1995
69	Alcantara at S.Giacomo	Alcantara	15	1983	1997

Table 4.6: List of gauging stations - second part

reason the data is used only the data belonging to the period 1965-1990;

In the Birgi at Chinisia station there is an artificial reservoir completed in 1970; since 1972, a modification of the data has been made taking into account the 41 km^2basin directly underlies and the 40% of the catchment area of 34 km^2 fastened;

In the Belice at Belice there is the Garcia dam built in 1985; the data recorded in a subsequent period to completion of the work are corrected by a correction factor that considers the basin (294 km^2) where the reservoir basin is present and the 40% of the area where the diversion dam is present (16 km^2).

In the Imera Meridionale at Drasi station there is the Olivo reservoir completed in 1982; since 1983, therefore it is necessary to modify the values of the runoff, taking into account the variation of the area.

Chapter 5

Comparative analysis of different spatial interpolation techniques of rainfall and temperature data

The availability of good and reliable rainfall data is fundamental for most of the hydrological analyses and for the design and management of water resources systems. However, in practice, precipitation or temperature records often suffer from missing data values mainly due to malfunctioning of gauging stations for a specific time period. This is an important issue in practical hydrology because it affects the continuity of climatic variables data and ultimately influences the results of hydrologic studies which use these as input. Many methods to estimate missing rainfall data have been proposed in literature and, among these, the most are based on spatial interpolation algorithms. In this chapter different spatial interpolation algorithms have been evaluated to produce a reasonably good continuous data set bridging the gaps in the historical series. The used algorithms are deterministic methods as inverse distance weighting, simple linear regression, multiple regression, geographically weighted regression and artificial neural network, and geostatistical models as ordinary kriging and residual ordinary kriging. In some of these methods, the elevation information, provided by a Digital Elevation Model, has been added to improve estimation of

missing data. These algorithms have been applied to the mean annual and monthly rainfall data of Sicily (Italy), measured at 247 raingauges and to the mean annual and monthly temperature data of Sicily (Italy), measured at 84 temperature stations.

Optimization of different settings of the various interpolation methods has been carried out using a subset of the available rainfall dataset (modelling set) while the remaining subset (validation set) has been used to compare the results by the different algorithms (*jack-knife* validation method, (Efron, 1982)).

5.1 Spatial interpolation approach for rainfall and temperature

The problem here analyzed is to provide the estimate $\hat{z}$ of the rainfall variable z at an ungauged location $\boldsymbol{x_0}$ using rainfall data at gauged sites. Denoting with $\{z(\boldsymbol{x_i}),\ i = 1, 2, ...N\}$ the precipitation dataset measured at the N sites $\boldsymbol{x_i}$, two different classes of interpolators have been here used: univariate methods and variables-aided interpolation (VAI) methods. While the former take into account only the data and spatial coordinates $\boldsymbol{x_i}$, the latter use also supplementary data as elevation $q(\boldsymbol{x_i})$, or other morphometric variables, of gauged sites in order to improve the estimation at the same ungauged sites. The univariate methods used in this work are radial basis function, inverse distance weighting and ordinary kriging while, as VAI methods, the linear regression between elevation and precipitation and between elevation and temperature, the multiple regression between temperature and morphometric variables, the geographically weighted regression, the artificial neural network and the residual kriging methods have been used. These methods were described in chapter 2.

This study has been carried out for the largest island in the Mediterranean Sea: Sicily, Italy which extends over an area of 25,700 km^2. The mean annual precipitation over Sicily is about 715 mm (period 1921-2004) with rainfall concentrated in the winter period. The July-August months are usually rainless. There is a strong spatial variability of precipitation, ranging from an average of 400 mm in the South-Eastern part to an average of 1300 mm in the Northern-Eastern part.

Concerning to the mean annual temperature over Sicily, it is about 17 °C (period 1924-2006) with with maximum temperatures in the period from June to October.

Precipitation and temperature dataset used in this study has been provided by OA-ARRA (*Osservatorio delle Acque - Agenzia Regionale dei Rifiuti e delle Acque*) and comes from 247 raingauge stations for the case of precipitation and 84 temperature

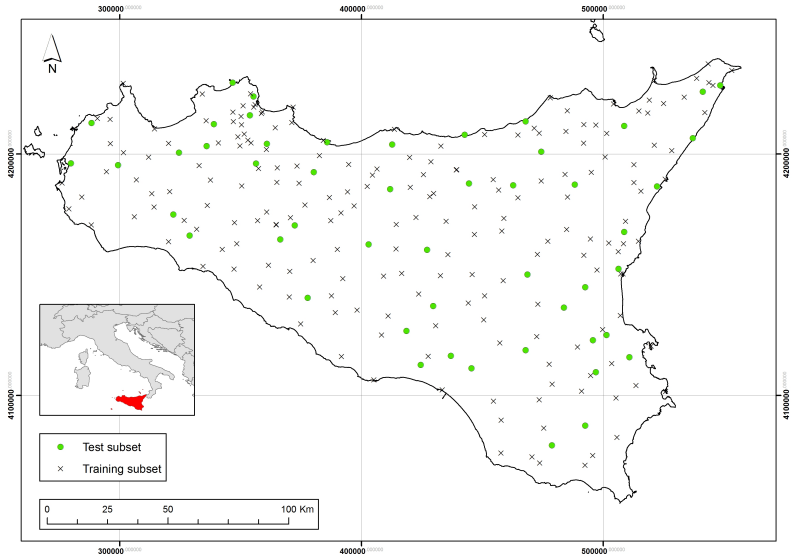

Figure 5.1: Location of the study area and position of the raingauge stations

stations for the case of temperature. The locations are shown in Figure 5.1 and in Figure 5.2 and the entire dataset has been divided randomly in two subsets: 1) *modelling subset*, used to calibrate the different interpolation methods, contains about the 80% of the entire dataset and 2) *validation subset*, used to validate the validate calibration results, contains about the remaining 20% of rainfall and temperature data. The division of the dataset into two subsets has been carried out also taking into account the empirical elevation distribution of the raingauges, i.e., the empirical *cdf's* of raingauges elevation in test and validation subsets are almost the same. The monthly and annual values of the precipitation dataset have been averaged, as said in chapter 4, over the period Jan 1921 - Dec 2004 and their basic sample statistics have been reported in Table 5.1. The monthly and annual values of the temperature dataset have been averaged, as said in chapter 4, over the period Jan 1924 - Dec 2006 and their basic sample statistics have been reported in Table 5.2.

The ancillary information regarding the elevation has been embedded through the use of a Digital Elevation Model (DEM) of the entire Sicily having a horizontal resolution of 100 m. From the Figure 5.3, showing the DEM of Sicily, it can be observed that about 62% of the surface is characterized by a hilly morphology, while about 24% can be ascribed to a mountainous morphology and the remaining part to plains. The most extended plain is the one around Catania in the Eastern part,

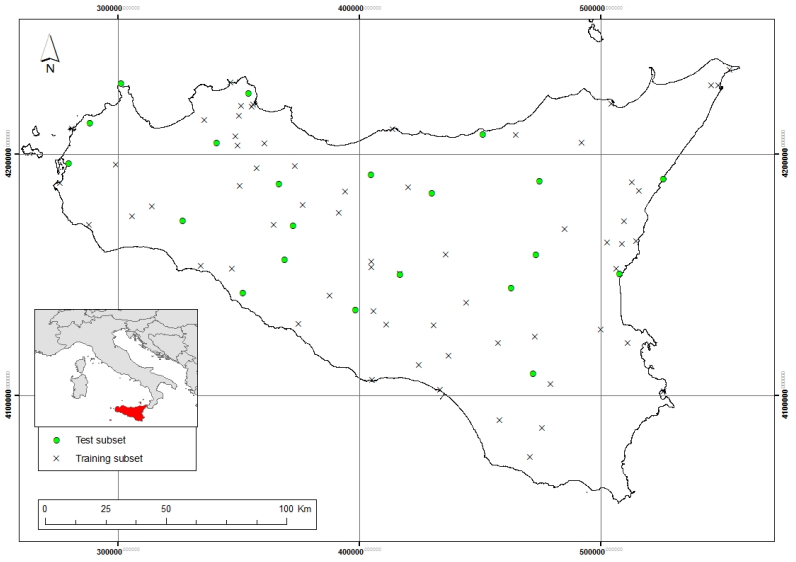

Figure 5.2: Location of the study area and position of the temperature stations

Period	Rainfall [mm]			
	Mean	Std. dev.	Min	Max
Jan	102.3	30.8	46.3	214.2
Feb	79.7	24.9	39.8	176.2
Mar	72.2	23.6	35.1	167.6
Apr	50.5	15.3	21.1	110.7
May	30	8.9	9.8	64.1
Jun	13.4	6	2.8	39.9
Jul	7.1	4.2	1.1	26
Aug	16.9	6	5.6	41.6
Sep	46.1	11.7	18.6	91.6
Oct	89.3	22	27	185.7
Nov	96.6	26	42.4	199.7
Dec	108.3	29.2	61.7	220.7
Annual	712.4	187.3	379.5	1346.8

Table 5.1: Statistics for the monthly and annual rainfall data (247 raingauge stations)

Period	Temperature [°C]			
	Mean	Std. Dev.	Min	Max
Jan	10.9	2.1	4.2	14.1
Feb	9.4	2.1	2.7	12.7
Mar	9.8	2.0	3.2	13.0
Apr	11.5	1.8	5.3	14.6
May	14.1	1.6	8.2	17.1
Jun	18.3	1.4	12.8	20.5
Jul	22.6	1.2	17.5	24.2
Aug	25.4	1.2	20.2	27.1
Sep	25.7	1.2	20.3	27.5
Oct	22.7	1.5	16.7	24.7
Nov	18.6	1.8	12.5	21.3
Dec	14.4	2.0	7.9	17.4
Annual	17.0	1.6	11.0	19.4

Table 5.2: Statistics for the monthly and annual temperature data (84 temperature stations)

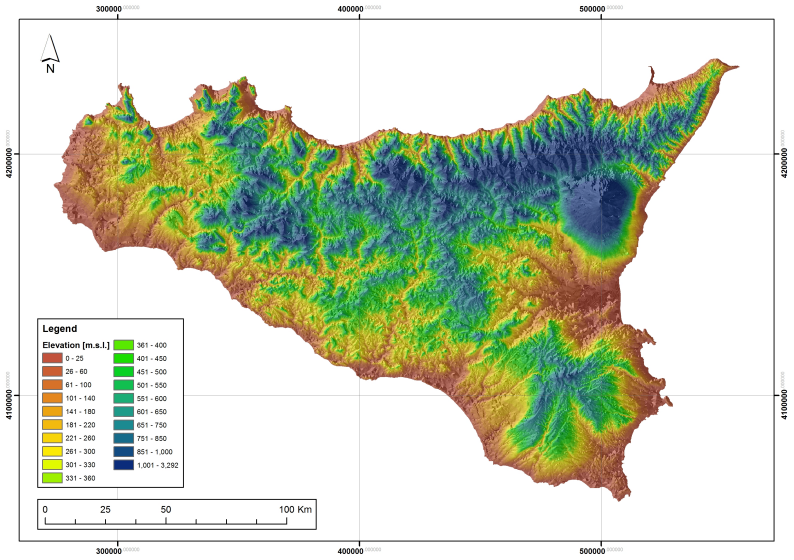

Figure 5.3: DEM 100 m of Sicily

while the Mount Etna with its 3,323 m is the highest mountain of Sicily. Along the Northern coast, from East to West, are situated the Peloritani, Nebrodi and Madonie Mountains, some of their peaks reaching 2,000 m. A series of high plateaus, which constitute Hyblaean Plateau, characterize South-Eastern part of Sicily.

The performances of the different interpolation methods have been assessed and compared using different indexes starting from the *validation subset*. It was not possible to use the classical geostatistical cross-validation (Isaaks and Srivastava, 1990) because of the presence of global interpolation methods such as LR and GWR. The comparison criteria are based on the following different indexes used to measure the strength of the statistical relationship between the estimated $\hat{z}(\boldsymbol{x_i})$ and measured $z(\boldsymbol{x_i})$ rainfall values in the N_v points of validation subset.

1. MSE, *mean square error* of prediction which measures the average square difference between the true rainfall and its estimate in the validation points (the root of MSE, RMSE, has been used as well):

$$MSE = \frac{1}{N_v} \sum_{i=1}^{N_v} \left[z(\boldsymbol{x_i}) - \hat{z}(\boldsymbol{x_i}) \right]^2 \tag{5.1}$$

2. MBE, *mean bias error* or simply bias

$$MBE = \frac{1}{N_v}\sum_{i=1}^{N_v}\left[z\left(\boldsymbol{x_i}\right) - \hat{z}\left(\boldsymbol{x_i}\right)\right] \tag{5.2}$$

3. MAE, *mean absolute error*

$$MAE = \frac{1}{N_v}\sum_{i=1}^{N_v}\left|z\left(\boldsymbol{x_i}\right) - \hat{z}\left(\boldsymbol{x_i}\right)\right| \tag{5.3}$$

4. s-MSE, *scaled mean square error* of prediction which measures the average square difference between the observed rainfall $z\left(\boldsymbol{x_i}\right)$ and its estimate $\hat{z}\left(\boldsymbol{x_i}\right)$ divided by the observed rainfall $z\left(\boldsymbol{x_i}\right)$ in the N_v validation points:

$$s - MSE = \frac{1}{N_v}\sum_{i=1}^{N_v}\left[\frac{z\left(\boldsymbol{x_i}\right) - \hat{z}\left(\boldsymbol{x_i}\right)}{z\left(\boldsymbol{x_i}\right)}\right]^2 \tag{5.4}$$

   This index has been used in the analysis of monthly results.

5. CC, linear correlation coefficient

$$CC = \frac{\sum_{i=1}^{N_v}\left[z\left(\boldsymbol{x_i}\right) - z_m\right]\left[\hat{z}\left(\boldsymbol{x_i}\right) - \hat{z}_m\right]}{\sigma_e \sigma_o N_v} \tag{5.5}$$

   where z_m and $\hat{z}_m$ are respectively the mean of measured and estimated rainfall values and σ_e and σ_o are are respectively the standard deviation of measured and estimated rainfall values.

5.2 Annual analysis: Precipitation

In this section the results obtained using different interpolation methods are analyzed and compared. This comparison is initially carried out using the average annual rainfall data. The results provided from this kind of analysis will be used to limit the number of trials carried out on the dataset concerning the average monthly precipitation. Finally, on the basis of the best average annual and monthly estimation methods, respectively, the reconstruction of rainfall data corresponding to the gaps present in the Hydrological Annals will be carried out for each year and for each month.

5.2.1 Univariate methods

The first univariate method here used has been the RBF with thin plate spline. The application of this method has been done through the use of the ESRI^{TM} ArcGIS Geostatistical Analyst tool. The strategy used in this application involves the choice of the minimum number of points N. Table 5.3 shows the different trials carried out in order to get the best results concerning statistical indexes calculated for the validation set. One can observe that all the statistical indexes worsen when N increases up to 12-15, then they get better for higher values of N. The best result, in terms of RMSE and CC, is obtained for N=40, although a satisfactory result is also achieved for N=10 (lowest value of MBE). The smoothing parameter c that appears in the radial basis function (eq.**??**), optimized by the software for each trial, maintains its value almost constant for each trial.

The estimate made by the second method, IDW, depends on the selection of the exponent r (eq. **??**) and the neighborhood search strategy. As can be seen in Table 5.3 several attempts have been made in order to get the optimal combination of such parameters. These trials have been carried out fixing the exponent r and changing the minimum number of points N or the search radius R. It is simple to observe that as the exponent r increases, the influence of the choice of N (or R) becomes more and more negligible and the trend of the RMSE values, is approximately constant. In fact, as the exponent r increases more and more, the IDW becomes the nearest neighbor method that is totally not-dependent from N (or R). Moreover one can observe that fixing N (or R) and increasing r, the bias decreases. Here the best performance is achieved for the exponent r equal to 3 and the minimum number of points N equal to 5 even if the bias relative to this trial is slightly greater (19.58) than the bias obtainable to the trials with $r = 5$. All the bias values related to RBF and IDW are positive and this points out that these univariate methods overestimate the precipitation at validation sites.

With regard to OK application, particular attention has been paid to the derivation of experimental and theoretical semivariograms; this derivation allows to solve the system of linear equations shown in eq. (**??**) in order to obtain the weights needed to estimate the z value at ungauged sites. As mentioned before, the experimental semivariogram is a function of both the distance and direction, and so it can account for direction dependent variability (anisotropic pattern). The exploration of the experimental semivariogram along several directions has shown the existence of typically anisotropic pattern. In particular, it has been observed that the sill varies

				MSE [mm^2]	RMSE [mm]	MBE [mm]	MAE [mm]	CC
RBF			N = 10	12,650	112.47	27.82	81.77	0.831
			N = 12	13,285	115.26	31.49	85.16	0.824
			N = 15	13,233	115.04	32.27	84.35	0.827
			N = 20	12,750	112.91	29.41	83.48	0.831
			N = 25	12,684	112.62	29.65	82.57	0.832
			N = 30	12,594	112.22	29.33	82.22	0.833
			N = 35	12,449	111.57	29.03	81.91	0.835
			N = 40	12,421	111.45	29.05	81.78	0.835
IDW	Number of points	r=2	N = 5	15,174	123.18	22.45	93.51	0.774
			N = 10	16,709	129.26	27.26	100.65	0.751
			N = 20	18,097	134.53	29.33	107.30	0.729
			N = 30	18,573	136.28	30.03	108.59	0.722
		r=3	N = 5	15,133	123.02	19.58	93.30	0.773
			N = 10	15,628	125.01	22.42	95.58	0.766
			N = 20	16,173	127.17	23.97	98.25	0.758
			N = 30	16,384	128.00	24.67	99.01	0.755
		r=5	N = 5	15,267	123.56	18.22	92.23	0.771
			N = 10	15,285	123.63	18.99	92.39	0.770
			N = 20	15,355	123.91	19.35	92.68	0.769
			N = 30	15,377	124.00	19.50	92.76	0.769
	Radius search	r=2	R = 20 Km	16,812	129.66	25.15	98.86	0.748
			R = 40 Km	18,539	136.16	29.27	108.20	0.721
			R = 60 Km	19,102	138.21	30.09	110.38	0.714
			R = 100 Km	19,772	140.61	32.29	113.90	0.710
		r=3	R = 20 Km	15,764	125.56	21.12	94.96	0.763
			R = 40 Km	16,370	127.95	24.46	98.83	0.755
			R = 60 Km	16,539	128.61	25.20	99.50	0.753
			R = 100 Km	16,657	129.06	26.07	100.34	0.752
		r=5	R = 20 Km	15,372	123.98	18.48	92.43	0.769
			R = 40 Km	15,375	124.00	19.47	92.78	0.769
			R = 60 Km	15,382	124.02	19.58	92.80	0.769
			R = 100 Km	15,383	124.03	19.62	92.80	0.769
OK				12,831	113.27	19.30	84.68	0.812

Table 5.3: Comparison of interpolation accuracy of the univariate applied to mean annual precipitation methods based on the five different statistical indexes

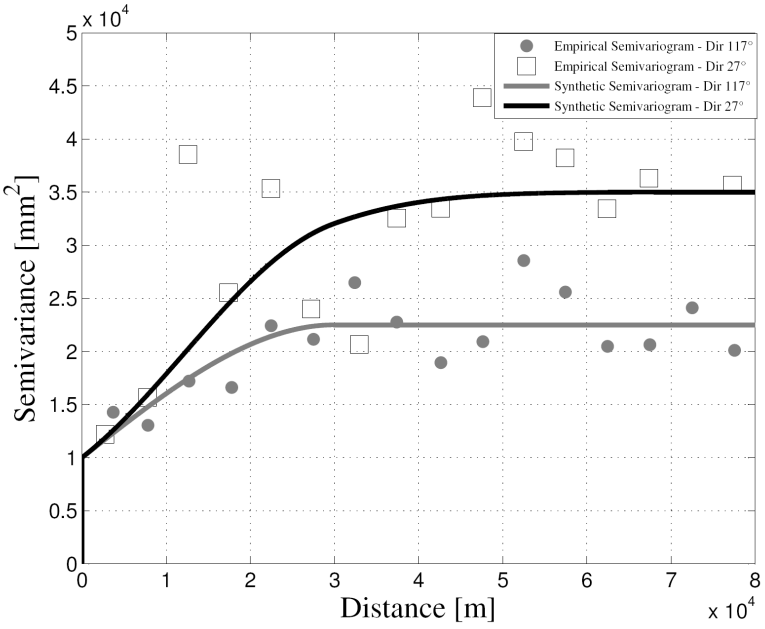

Figure 5.4: Empirical and synthetic semivariograms of average annual precipitation in the zonal direction and in the principal direction (zonal anisotropy)

along different directions determining the phenomenon known as *zonal anisotropy*. In this work, taking as reference a north-eastern system measured in degrees clockwise from positive Y (north), the zonal direction (direction of maximum variability) has been found at 27 degrees, while the isotropy direction (i. e. the perpendicular to zonal direction) has been found at 117 degrees (Figure 5.4).

In order to fit a theoretical semivariogram to the empirical one, taking into account the detected zonal anisotropy, the GSTAT software (http://www.gstat.org/) has been used. This software allows to create zonal anisotropy semivariogram starting from the concept of geometric anisotropy (i.e. anisotropy in the range). In particular GSTAT uses the following notation to model the zonal anisotropy:

$$\gamma = n\,Nug() + c_1\,Mod(a_1) + c_2\,Mod(a_2,\,p,\,s) \tag{5.6}$$

where $n\,Nug()$ is the nugget, c_1 is the partial sill of the theoretical semivariogram at the principal direction, c_2 is the value that has been summed to c_1 to obtain partial sill in the zonal direction, Mod is the model type, a_1 is the maximum range of the

variogram model, a_2 is the maximum range amplified, p is the angle for the principal direction, s is the anisotropy ratio (i.e. the ratio between the minor range and the maximum range). Through eq. (5.6), the theoretical zonal anisotropic semivariogram has been obtained as the sum of a isotropic semivariogram, given by the first two terms, and an anisotropic semivariogram obtained by defining a geometric anisotropy with large anisotropy ratio. For the average annual rainfall data the eq. (5.6) assumes the following form:

$$\gamma = 10000\, Nug() + 12500\, Sph(30000) + 12500\, Gau(1.25 * 10^8,\, 117,\, 0.0001) \quad (5.7)$$

The parameter values of the eq. (5.7) have been here chosen by observing to the experimental semivariogram. In particular the nugget value, $nNug()$ is equal to *10000* mm^2, c_1 and c_2 are equal to *12500* mm^2; a_1 is equal to *30000* m, a_2 is equal to $1.25{*}10^8$ m and p is equal to *117* degrees. The model types chosen for this application are spherical (Sph) and gaussian (Gau) models (Kitadinis, 1997) and their selection depends on the behavior of the semivariograms close to the origin.

With regard to the application of the OK, one can observe that its RMSE and CC values are comparable to the RMSE and CC values obtained with the RBF while the MBE value is comparable to the MBE value obtained with the IDW. For these reasons, the OK has been here considered the best method among the univariate methods in terms of accuracy and bias.

5.2.2 VAI methods

The first VAI method here applied is the LR. It uses the dependence of rainfall from elevation, in order to estimate the rainfall. In particular, different relationships between z and q have been tested, considering that in scientific literature is demonstrated that the spatial distribution of annual rainfall depth z seldom follows a log-normal or quadratic root distribution (Revfeim, 1990 ; Suhaila and Jemain, 2007). With regard to the regression coefficient estimation, two different statistical methods have been used: the OLS and the ROB regression. By observing table 5.4 it is possible to argue that the linear regression method with the OLS gives satisfactory results in terms of MBE, as one can expect by underlying hypotesis of this methods (i.e. expected value of errors is equal to zero on modelling set). The trials performed with ROB regression give good results in terms of RMSE and MBE as well. The values of these statistical indexes are slightly better for the linear regression with OLS but the ROB regression

			MSE [mm^2]	RMSE [mm]	MBE [mm]	MAE [mm]	CC
LR	OLS	z = f(q)	29470	171.67	20.63	136.04	0.477
		log(z) = f(q)	28563	169.01	3.23	134.07	0.49
		sqrt(z) = f(q)	28941	170.12	11.64	134.83	0.483
	ROB	z = f(q)	29206	170.90	-0.37	135.46	0.477
		log(z) = f(q)	28619	169.17	-3.77	134.20	0.49
		sqrt(z) = f(q)	28891	169.97	-0.74	134.89	0.483
GWR		N=10	20306	142.50	3.30	90.85	0.682
		N=15	19794	140.69	12.53	99.17	0.686
		N=20	20580	143.46	6.04	102.09	0.67
		N=25	20322	142.56	1.72	102.57	0.667
		N=30	21240	145.74	0.50	105.95	0.648
		N=35	22982	151.60	2.45	110.45	0.61
ANN		MLP 3-4-1	19075	138.11	29.34	98.89	0.713
		MLP 3-6-1	18327	135.38	10.33	99.14	0.741
		MLP 3-8-1	14847	121.85	7.87	83.56	0.781
		MLP 3-10-1	17773	133.32	7.14	98.09	0.728
		MLP 3-12-1	19736	140.58	16.77	97.63	0.714
RK		LR-ROB; log(z)=f(q)	7659	87.52	14.32	65.43	0.894
		GWR; N=15	14155	118.98	15.63	79.27	0.801
		ANN-MLP 3-4-1	10074	100.37	12.07	71.33	0.86

Table 5.4: Comparison of interpolation accuracy of the VAI methods applied to mean annual precipitation based on the five different statistical indexes

method has been preferred in order to avoid the influence of outliers on the rainfall estimation. In particular, the best performance obtained with ROB regression is provided by the relationship between the log-transformation of depth $z(\boldsymbol{x}_i)$ and $q(\boldsymbol{x}_i)$ (Figure 5.5).

With regard the second VAI method (GWR), the adaptive spatial kernel method has been here applied. In particular a minimum number of points has been chosen in order to have at least this number into the considered kernel. Analyzing the statistical indexes computed for the validation set, shown in Table 5.4, one can observe that the best performance, in terms of accuracy, is achieved for a number of points N equal to 15; unfortunately this trial is, at the same time, the most accurate (lowest value of RMSE) and the most biased (highest value of MBE).

The third VAI method is an ANN using MLP structure. In order to define such an architecture, the training algorithm and the learning rate are fixed during the training phase. The ANN has been trained by the scaled conjugate gradient algorithm (Moeller, 1993). The initial weights have been extracted randomly from a standardized normal distribution and the value of the learning rate, initially equal to 0.04, decreases according to an exponential law. This technique, called *early stopping* (Bishop, 1995), has been used in order to decrease the risk of overfitting and

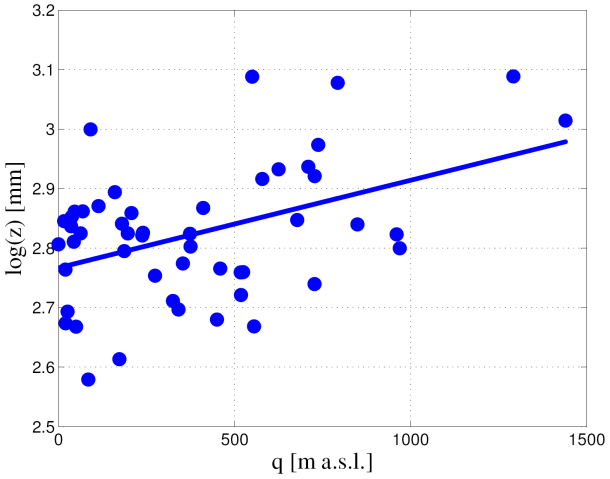

Figure 5.5: Relationship between the logarithm of average annual precipitation and raingauge elevation for the test subset (R^2=0.45); the trend line has been obtained with ROB method

training is stopped after only 100 iterations. Five network architectures, having different number of neurons in the hidden layer, have been compared and, on the basis of the RMSE and MBE performances (Table 5.4), the best one is the network with one hidden layer having 8 neurons with hyperbolic tangent activation function (MLP 3-8-1) both in terms of RMSE and MBE values. Moreover, the correlation coefficient between estimated and measured data for the validation set is equal to 0.79, confirming a good agreement between estimates and observations. The correlation coefficient between residuals and measured data of the validation set, equal to -0.47, shows that not all the spatial data correlation has been explained by the ANN, also implying that overfitting has been avoided.

Another evaluation of the network's performances has been completed by computing the Akaike's FPE (*final prediction error*) (Larsen et al., 1994). This is a validation index that uses the information from all the second order derivatives of the error function to remove unimportant weights from a trained network (*pruning technique*) allowing to identify the best network architecture. In order to find the optimal MLP architecture, taking into account generalization error, the pruning technique called *optimal brain surgeon* (OBS, Hassibi and Stork., 1993) has been applied. It allows to improve the generalization capability of the MLP determining the significance for each weight, which is an estimate of the variation in the FPE index if the

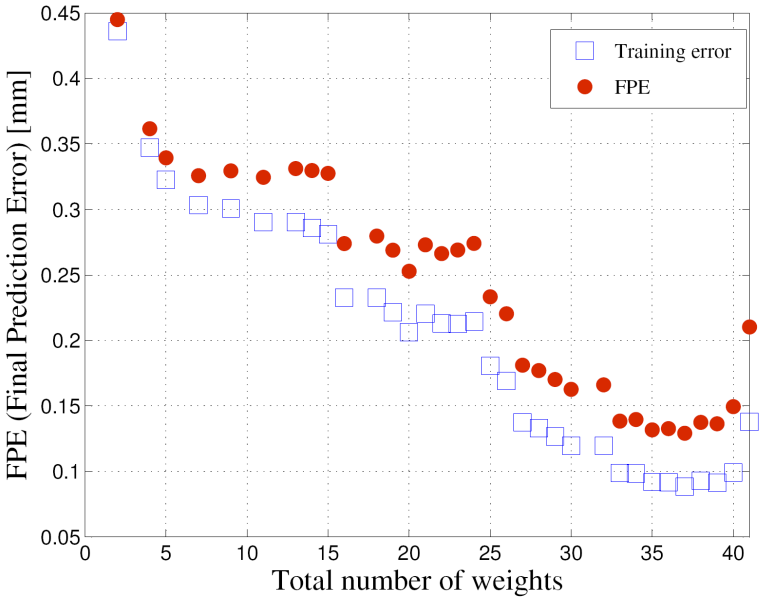

Figure 5.6: Training error and FPE plots as a function of the network parameters (weights). The abscissa from right to left points out the number of pruned weights

corresponding weights is pruned. The goal of OBS is the pruning of the weights with the smallest importance and the optimal network is the one whose weights yield the smallest FPE. Figure 5.6 shows that the minimum FPE is obtained after pruning only 3 weights and this confirms that the MLP architecture is well chosen.

Taking into account the values of the correlation coefficients between residuals and measured data of the ROB regression and GWR, respectively equal to -0.91 and -0.64, one can observe that not all the spatial data correlations have been explained by these above mentioned methods. The same observation can be done for the MLP with 8 neurons in the hidden layer whose correlation coefficient between residuals and measured data, as said above, is -0.47. In particular the ROB regression does not allow to explain definitely the deterministic component. This is demonstrated by the fact that the obtained residuals have a great spatial correlation, very similar to the one of the original data. On the contrary, the MLP permits to accurately explain the deterministic component providing a good estimate of z, as can be noted from the correlation coefficient between residuals and measured data value. As the table

5.4 shows, MLP is characterized, among the three analyzed VAI methods (LR, GWR and ANN), by the lowest value of RMSE, by the highest CC value and by MBE value comparable to the other methods; its perfomances are then comparable, on the whole, to the those provided by OK .

Starting from these considerations, the fourth VAI method (the residual kriging, RK) has been used and tested. The basic idea of RK is, as mentioned in section 3.2.4, that to estimate the residuals coming from a VAI method (LR, GWR and ANN) at an ungauged site and then to sum them to the estimate previously obtained by underlying deterministic interpolators.

In order to obtain the best results in terms of estimate by application of RK approach, a undersized MLP has been chosen. In this way, MLP will carry out a simple de-trending of data and leaving a significant spatial variability to the stochastic component that will be processed by OK. In particular a MLP 3-4-1 has been chosen and its training procedure is the same of that explained in the previous section, except for the learning rate which starts from 0.01. The correlation coefficient between residuals coming from this MLP and measured data is equal to -0.63 pointing out that is more spatial information has been left in the residuals than to the case relative to the MLP 3-8-1. The OK is applied to the residuals derived from the LR, GWR and ANN (MLP 3-4-1). In this case, an isotropic behavior in the residuals coming from the three different methods has been noticed through the derivation of experimental semivariograms.

As described in Section 3.2.4, the estimated values of residuals have been added to the LR, IDW and ANN, respectively, according to eq. (18). In table 5.4 one can observe that most of the performance of these three methods are better than the univariate and the other VAI methods. The best method, in terms of RMSE and CC, appears to be the RK-LR while the method that has the lowest MBE value is the RK-ANN (MLP 3-4-1).

Figures 5.7 and 5.8 show the performances of the best methods respectively in terms of RMSE and MBE for both univariate and VAI methods. The best univariate method, in terms of RMSE, is the RBF with N=40 even if a high bias can be noticed, making this method the most biased one. The IDW behavior is opposite to the RBF one and is characterized by a lower MBE and a higher RMSE. The OK gives good results with a MBE value lower than that provided by the RBF and it can be considered, as mentioned before, the best analyzed univariate method. However all univariate methods are very biased providing a systematic overestimation of z. In order to improve the results in terms of MBE, it is necessary to switch from univariate

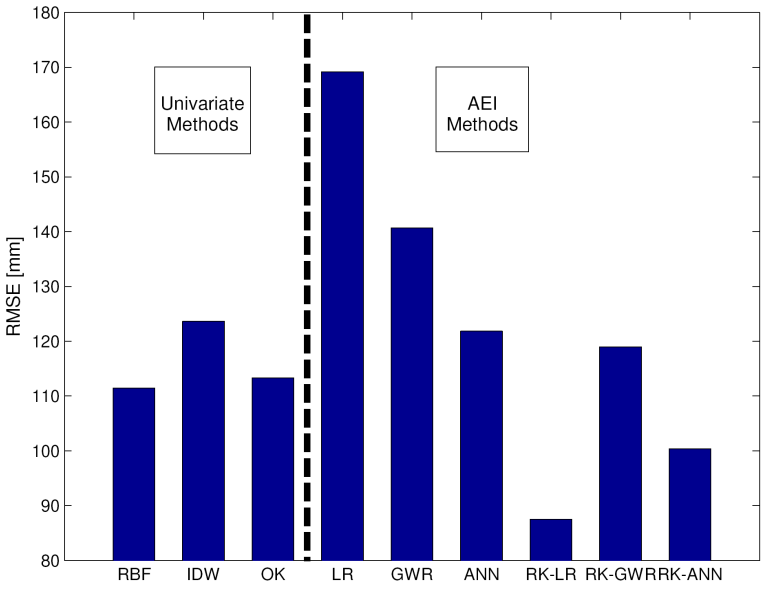

Figure 5.7: RMSE of best univariate and VAI interpolation methods

methods to the VAI ones, by introducing the elevation information. For example, the LR, as shown in Figures 5.7 and 5.8, is a method with a low bias but very high value of RMSE. The GWR and ANN give better results with acceptable value of RMSE and MBE. The best methods are however the RK and in particular RK-LR, that gives the lowest value of RMSE, the highest CC and an acceptable value of MBE, and RK-ANN (MLP 3-4-1) with acceptable values of RMSE and CC and the lowest value of MBE.

A deep analysis of Figures 5.7 and 5.8 and table 5.4 shows that the application of ordinary kriging to the residuals coming from any VAI methods decreases in substantial manner the RMSE values of the underlying VAI method (from 140.69 to 118.98 for GWR; from 169.17 to 87.52 for LR-ROB; from 138.11 to 100.37 for MLP 3-4-1) but, at the same time, increases the bias of the same method (from 12.53 to 15.63 for GWR; from -3.77 to 14.32 for LR-ROB), except for MLP 3-4-1 (from 29.34 to 12.07).

Here the RK-LR method has been chosen as the best method even if one can observe that a RK-LR presents value of MBE denoting a systematic overestimation of z; this overestimation can be also observed from the plot shown in Figure 5.9 which compares the observed average annual precipitation z_{obs} with the average annual precipitation $z_{est\,(RK-LR)}$ estimated with the RK-LR method. Figure 5.10 shows the

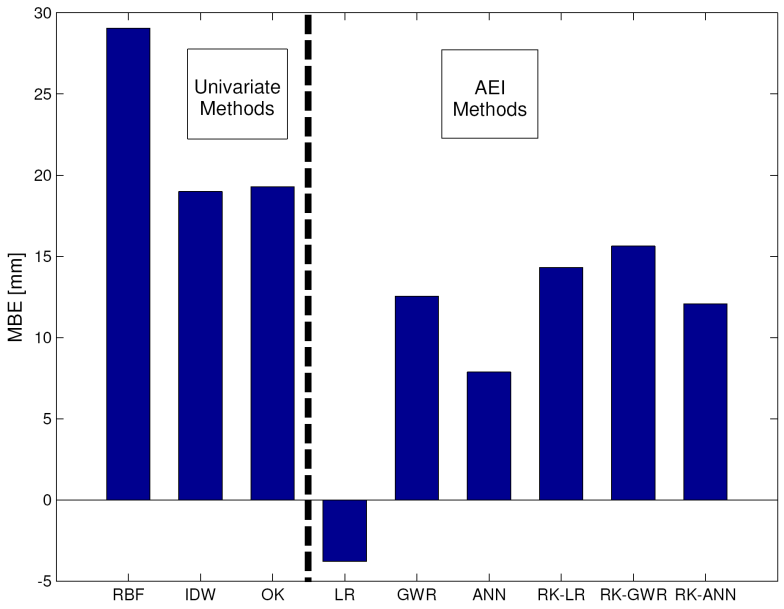

Figure 5.8: MBE of best univariate and VAI interpolation methods

mean annual rainfall map produced by RK-LR [ROB with $\log(z)$-q] method.

5.3 Monthly analysis: Precipitation

The interpolation methods previously described have been also applied to the average monthly rainfall data, taking into account the results obtained in the case of average annual rainfall data. Particular attention has been paid to the estimation of the semivariograms of both the average monthly data and the residuals coming from the use of three VAI methods (i.e. LR-ROB, GWR and MLP 3-4-1). Also at monthly scale, the presence of a zonal anisotropy has been observed mainly in the average monthly data. Furthermore the observation of the empirical semivariograms derived from original data for the different months points out the presence of a seasonal trend of the sills in the zonal and isotropic directions, shown in figure 5.11.

In table 5.5 an overview of the results obtained for the statistics is shown. The method leading to the best result is indicated for each month together with the corresponding value of the relative statistical indexes. With regard to the RMSE values, it can be observed that the best values of this index are always obtained through the application of a VAI method. In particular, in most of the months

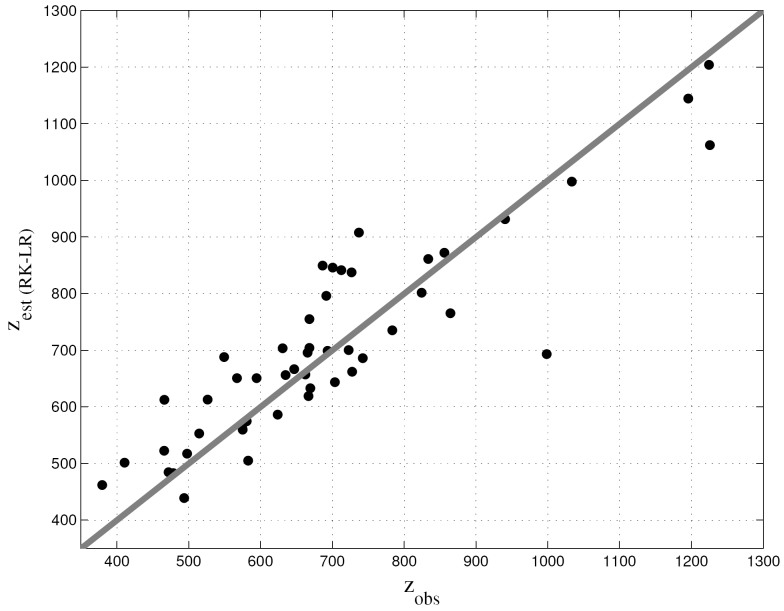

Figure 5.9: Scatterplot between observed annual precipitation and annual precipitation estimated with the RK-LR method

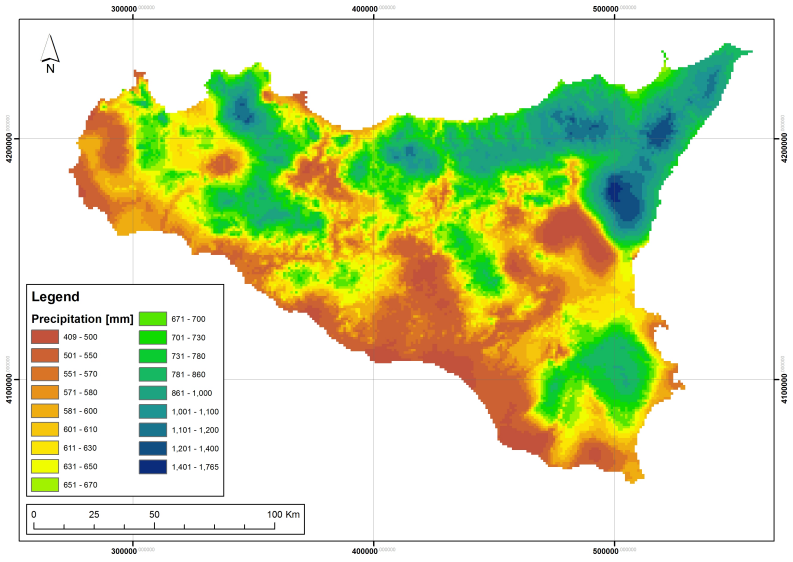

Figure 5.10: Mean annual precipitation interpolated using RK-LR log(z)-q method

Indexes / Months	MSE [mm^2]	RMSE [mm]	MBE [mm]	MAE [mm]	s-RMSE	CC
Jan	RK-LR *309.26*	RK-LR *17.59*	LR-ROB *2.60*	RK-LR *13.39*	RK-LR *0.03*	LR-ROB *0.84*
Feb	RK-LR *210.29*	RK-LR *14.50*	LR-ROB *1.85*	RK-LR *10.73*	RK-LR *0.04*	LR-ROB *0.81*
Mar	RK-LR *158.13*	RK-LR *12.57*	LR-ROB *-2.30*	RK-LR *9.27*	RK-LR *0.03*	LR-ROB *0.85*
Apr	RK-LR *68.70*	RK-LR *8.29*	LR-ROB *-3.23*	RK-LR *6.33*	RK-LR *0.03*	LR-ROB *0.88*
May	RK-LR *22.06*	RK-LR *4.70*	LR-ROB *-1.37*	RK-LR *3.58*	RK-LR *0.03*	IDW *0.88*
Jun	GWR *17.78*	GWR *4.22*	LR-ROB *-0.53*	GWR *2.86*	IDW *0.14*	LR-ROB *0.72*
Jul	RK-LR *6.62*	RK-LR *2.57*	LR-ROB *-1.23*	RK-LR *1.87*	IDW *0.22*	GWR *0.84*
Aug	RK-LR *16.59*	RK-LR *4.07*	LR-ROB *-0.55*	RK-LR *3.11*	OK *0.06*	LR-ROB *0.82*
Sep	RK-LR *55.89*	RK-LR *7.48*	LR-ROB *-2.80*	RK-GWR *5.75*	RK-LR *0.02*	LR-ROB *0.83*
Oct	RK-ANN *234.67*	RK-ANN *15.32*	LR-ROB *-3.41*	RK-ANN *11.30*	RK-LR *0.03*	LR-ROB *0.85*
Nov	RK-LR *279.76*	RK-LR *16.73*	LR-ROB *-0.11*	RK-LR *12.63*	RK-LR *0.03*	LR-ROB *0.81*
Dec	RK-LR *325.14*	RK-LR *18.03*	LR-ROB *2.15*	RK-LR *13.74*	RK-LR *0.03*	LR-ROB *0.81*

Table 5.5: Overview of monthly results. Each box reports the model leading to the best result and the corresponding index value.

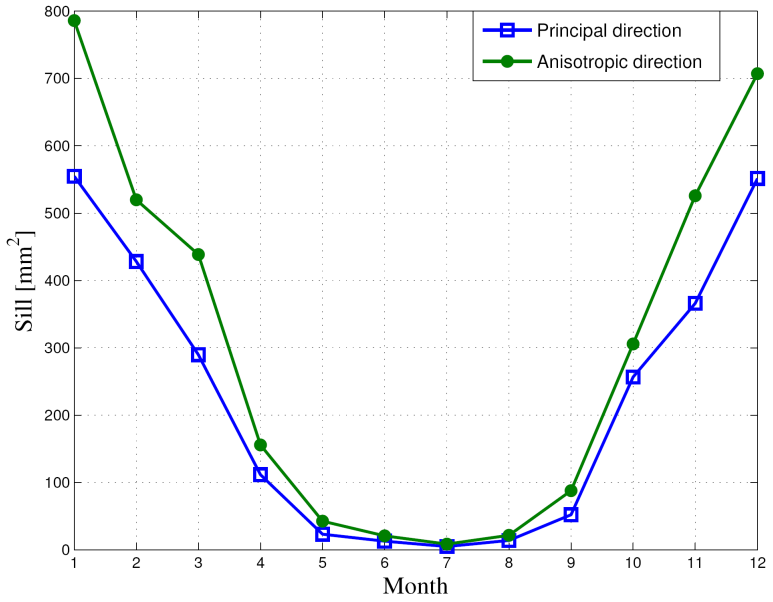

Figure 5.11: Trend of sills of the average monthly rainfall data in the two explored direction (zonal and principal direction)

the lowest value of RMSE is achieved using the VAI methods with residual kriging application, as RK-LR. Only for two months, the best value of RMSE is given by the other methods application: for June the best performance is provided by the GWR method, while for October the best method is the RK-ANN. The same observations can be approximately made analyzing the index MAE that is characterized by the best values for the RK-LR application except for June and September (GWR and RK-GWR respectively). With regard to the unbiasedness of the different methods, the best results are instead achieved with LR-ROB application that, as already mentioned, is the least biased of the all methods. The s-RMSE is characterized by very low values except for the months of June and July; this could be due to the influence of low observed rainfall values in summer months on the s-RMSE behavior.

The results obtained at annual and monthly scale have been used to fill the gaps in the precipitation dataset when a monthly value or data of an entire year are missing for a gauge. Missing rainfall data are estimated through spatial interpolation with the best method among those previously exposed. In particular the RK-LR method, both for annual and monthly case, has been chosen also for those months in which the application of the same method does not lead to the best values of the statistical

1921														
	Station	Jan	Feb	Mar	Apr	May	Jun	Jul	Aug	Sept	Oct	Nov	Dec	Annual
10	CALVARUSO	74	82	120	61	35.2	108.5	30.5	17	63	141	141	117	990.2
20	S. SABA	[37.8]	[75.4]	[105.6]	[58.8]	[23.4]	[89.4]	[23.4]	[27.8]	[58.8]	[112.8]	[143.3]	[97.9]	[854.4]
30	ZIRIO'	[186.6]	[78.3]	[122.5]	[85.8]	[45.9]	[105.6]	[31.8]	[40.4]	[74.7]	[108.5]	[130.4]	[110.7]	[1121.2]
40	MONFORTE S.GIORGIO	77.1	69.9	87.8	77.8	23.3	107.7	16.9	38.2	81	124.3	151.2	119.5	974.7
50	S. LUCIA DEL MELA	107.7	96.9	46.8	77.3	38.9	144.2	22.1	51.7	120.2	77	87.9	124.9	995.6
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
3390	TREMESTIERI	231.8	99	39.4	47.6	24.8	2.1	-	-	1.4	71.6	127.7	67.5	712.9
3400	GANZIRRI	258.4	85.4	39.6	75.4	27.6	10	-	-	0.5	68.7	53.3	61.4	680.3
3410	ALI" TERME	[163.5]	[165.5]	[6.3]	[31.5]	[20.0]	[3.1]	[-]	[-]	[0.7]	[26.4]	[94.8]	[76.3]	[588.1]

Table 5.6: Reconstruction of part of Sicilian Hydrological Annals (year 1921)

indexes (June and October).

The most common case observed in the rainfall dataset is the simultaneous lack of monthly and annual data. In this case the first step consists of the estimation of the annual precipitation for the missing year followed by the estimation at monthly scale. Both these estimations are carried out considering all the available raingauges for the examined year. In the second step the monthly estimations have been corrected to make them congruent with the annual estimation, here, considered more reliable than the monthly estimates. This correction simply consists in the assessment of a corrective coefficient $\psi_{i,j}$ to apply to all the months of i-th year and for the j-th gauge; this corrective coefficient is given by the following relationship:

$$\psi_{i,j} = \frac{P_{ann,i,j}}{M_{i,j}} \tag{5.8}$$

where $P_{ann,i,j}$, is the annual estimation for the i-th year and for the j-th gauge and $M_{i,j}$ is the sum of monthly estimation $\mu_{i,j,k}$:

$$M_{i,j} = \sum_{k=1}^{12} \mu_{i,j,k}. \tag{5.9}$$

The estimated precipitation data together with the observed data have been collected in a relational data-base which allows the creation of reports with the same form of Hydrological Annals. Figure 5.6 shows a share of Hydrological Annals (with the observed data and reconstructed data in square brackets).

5.4 Annual analysis: Temperature

In this section the results obtained using different interpolation methods are analyzed and compared, using the same criteria employed for the case of precipitation (Section 1.2). For the case of temperatures a different method, not previuosly considered for the precipitation estimation, has been used, i.e. the multiple regression, described in

Chapter 2.

5.4.1 Univariate methods

The first univariate method here used has been the RBF with thin plate spline. As in the case of precipitation, the application of this method has been done through the use of the ESRI^{TM} ArcGIS Geostatistical Analyst tool. The strategy used in this application involves the choice of the minimum number of points N. Table 5.7 shows the different trials carried out in order to get the best results concerning statistical indexes calculated for the validation set. One can observe that the RMSE value decreases for values of Nranging from 10 to 30, then it remains constant for values of Nranging from 30 to 40. The MBE worsen when Nincreases up to 10-20, then it gets better for higher values of N. The best result, in terms of RMSE and CC, is obtained for N=30. The smoothing parameter c that appears in the radial basis function (eq....., in section 2.1.3), optimized by the software for each trial, maintains its value almost constant for each trial.

The estimate made by the second method, IDW, depends on the selection of the exponent r(eq. ...., section 2.1.1) and the neighborhood search strategy. As can be seen in Table 5.7 several attempts have been made in order to get the optimal combination of such parameters. These trials have been carried out fixing the exponent r and changing the minimum number of points N or the search radius R. For the temperature, the best performance is achieved for the exponent $r$equal to 2 and the search radius R equal to 40 km, both in terms od RMSE and bias. The best result, in terms of CC, is instead obtain for $r = 2$ and $N = 5$. As for the results obtained in the applications of precipitation, all the bias values related to RBF and IDW are positive and this points out that these univariate methods overestimate the precipitation at validation sites.

With regard to OK application, particular attention has been paid to the derivation of experimental and theoretical semivariograms; As mentioned before, the experimental semivariogram is a function of both the distance and direction, and so it can account for direction dependent variability (anisotropic pattern). The exploration of the experimental semivariogram along several directions has shown the existence of typically anisotropic pattern, also in the experimental semivariogram of the annual mean temperature. In particular, it has been observed that the sill varies along different directions determining the phenomenon known as *zonal anisotropy*. In this case , taking as reference a north-eastern system measured in degrees clockwise from

				MSE [°C^2]	RMSE [°C]	MBE [°C]	MAE [°C]	CC
RBF			N = 10	1.7956	1.3400	-0.1961	1.0211	0.6883
			N = 12	1.8378	1.3556	-0.2205	1.0218	0.6964
			N = 15	1.7485	1.3223	-0.2355	0.9902	0.7146
			N = 20	1.7264	1.3139	-0.2296	0.9752	0.7157
			N = 25	1.6961	1.3023	-0.2231	0.9655	0.7203
			N = 30	1.6596	1.2883	-0.2117	0.9527	0.7250
			N = 35	1.6585	1.2878	-0.2117	0.9528	0.7250
			N = 40	1.6585	1.2878	-0.2117	0.9528	0.7250
IDW	Number of points	r=2	N = 3	2.4079	1.5518	-0.4396	1.1974	0.5490
			N = 5	2.2966	1.5154	-0.3744	1.1578	0.5526
			N = 10	2.3839	1.5440	-0.3173	1.1936	0.4990
			N = 20	2.3680	1.5388	-0.2515	1.2020	0.4840
		r=3	N = 3	2.4056	1.5510	-0.4311	1.1905	0.5497
			N = 5	2.4344	1.5603	-0.3926	1.1893	0.5205
			N = 10	2.4202	1.5557	-0.3582	1.1855	0.5114
			N = 20	2.4203	1.5557	-0.3419	1.1858	0.5055
		r=5	N = 3	2.6211	1.6190	-0.5010	1.2637	0.5293
			N = 5	2.6205	1.6188	-0.4891	1.2680	0.5212
			N = 10	2.6117	1.6161	-0.4828	1.2667	0.5202
			N = 20	2.6114	1.6160	-0.4816	1.2669	0.5196
	Radius search	r=2	R = 20 Km	3.2091	1.7914	-0.5241	1.3502	0.4114
			R = 40 Km	2.2638	1.5046	-0.3018	1.1520	0.5348
			R = 60 Km	2.4169	1.5546	-0.2492	1.2088	0.4737
			R = 100 Km	2.3792	1.5425	-0.1678	1.2127	0.4639
		r=3	R = 20 Km	3.1948	1.7874	-0.5299	1.3585	0.4204
			R = 40 Km	2.3505	1.5331	-0.3893	1.1610	0.5452
			R = 60 Km	2.4142	1.5538	-0.3586	1.1778	0.5158
			R = 100 Km	2.4029	1.5501	-0.3272	1.1822	0.5056
		r=5	R = 20 Km	3.1838	1.7843	-0.5399	1.3718	0.4286
			R = 40 Km	2.5908	1.6096	-0.4891	1.2580	0.5299
			R = 60 Km	2.6055	1.6142	-0.4818	1.2638	0.5221
			R = 100 Km	2.6092	1.6153	-0.4805	1.2662	0.5197
OK				2.247078	1.499026	-0.24529	1.105024	0.534335

Table 5.7: Comparison of interpolation accuracy of the univariate applied to mean annual precipitation methods based on the five different statistical indexes

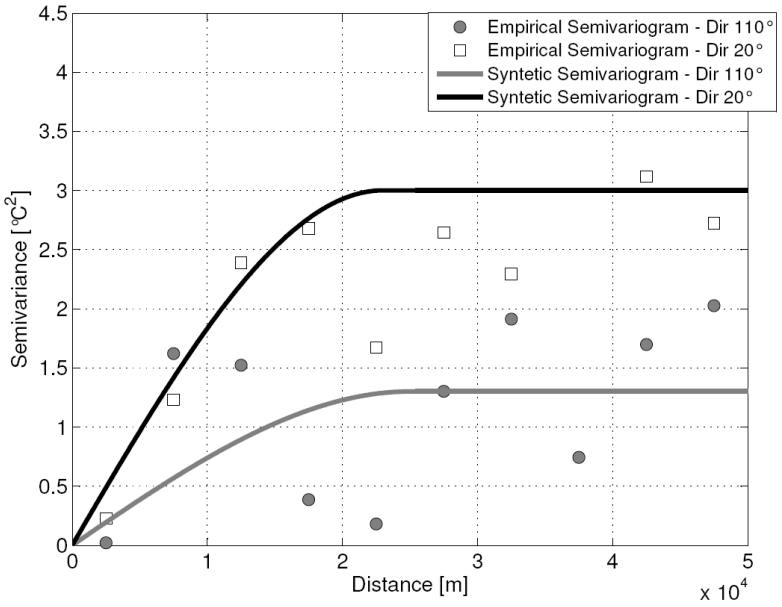

Figure 5.12: Empirical and synthetic semivariograms of average annual precipitation in the zonal direction and in the principal direction (zonal anisotropy)

positive Y (north), the zonal direction (direction of maximum variability) has been found at 20 degrees, while the isotropy direction (i. e. the perpendicular to zonal direction) has been found at 110 degrees (Figure 5.12).

In order to fit a theoretical semivariogram to the empirical one, taking into account the detected zonal anisotropy, the GSTAT software (http://www.gstat.org/) has been used.

Through eq. (5.6), defined before, the theoretical zonal anisotropic semivariogram has been obtained as the sum of a isotropic semivariogram, given by the first two terms, and an anisotropic semivariogram obtained by defining a geometric anisotropy with large anisotropy ratio. For the average annual temperature data the eq. (5.6) assumes the following form:

$$\gamma = 0\,Nug() + 1.25\,Sph(25000) \; + \; 1.75\,Gau(1.25 * 10^4,\, 110,\, 0.0001) \qquad (5.10)$$

The parameter values of the eq. (5.10) have been here chosen by observing to the experimental semivariogram. In particular the nugget value, $nNug()$ is equal to *0*, c_1 is equal to 1.25 $^{\circ}C^2$ and c_2 are equal to *1.75* $^{\circ}C^2$; a_1 is equal to *25000* m, a_2 is equal to $1.25 * 10^4$ $m$and p is equal to *110*degrees. The model types chosen for this application are spherical (Sph) and gaussian (Gau) models (Kitadinis, 1997) and their selection depends on the behavior of the semivariograms close to the origin.

With regard to the application of the OK, one can observe that its RMSE value is better than the RMSE value obtained with IDW and worse than the RMSE value obtained with RBF. The MBE and CC values are better than the MBE and CC obtained with IDW, while the MBE and CC values are comparable to the MBE and CC values obtained with the RBF.

Finally, the RBF has been here considered the best method among the univariate methods in terms of accuracy and bias.

5.4.2 VAI methods

The first VAI method here applied is the LR. It uses the dependence of temperature from elevation, in order to estimate the temperature. In particular, as in the case of rainfall, different relationships between z and q have been tested, as log-normal or quadratic root distribution. With regard to the regression coefficient estimation, two different statistical methods have been used: the OLS and the ROB regression. By observing table 5.8 it is possible to argue that the linear regression method with the OLS and with the original depth $z(\boldsymbol{x}_i)$ gives the best results in terms of MBE, as one can expect by underlying hypotesis of this methods (i.e. expected value of errors is equal to zero on modelling set) and in terms of RMSE.

With regard the second VAI method (GWR), the adaptive spatial kernel method has been here applied. In particular a minimum number of points has been chosen in order to have at least this number into the considered kernel. Analyzing the statistical indexes computed for the validation set, shown in Table 5.8, one can observe that the best performance, in terms of accuracy, is achieved for a number of points N equal to 20; in this trial a MBE equal to 0.028 is obtained. It is not the lowest value of MBE but it is comparable with the lowest MBE values of the other trials.

The third VAI method is the multiple linear regression by means of a stepwise procedure.It uses the dependence of temperature from geographic and morphologic parameters, in order to estimate the temperature. According to the criteria suggested

			MSE [°C^2]	RMSE [°C]	MBE [°C]	MAE [°C]	CC
LR	OLS	z = f(q)	0.2388	0.4886	-0.0014	0.376	0.9598
		log(z) = f(q)	0.2882	0.5369	-0.0039	0.4196	0.9445
		sqrt(z) = f(q)	0.2614	0.5113	-0.0028	0.3954	0.9535
	ROB	z = f(q)	0.241	0.4909	-0.0098	0.3784	0.9598
		log(z) = f(q)	0.2941	0.5423	-0.0058	0.4218	0.9452
		sqrt(z) = f(q)	0.2668	0.5165	-0.0012	0.4	0.9537
ML	SW	$T_s = 18.68 - 0.0057\ q - 8*10^{-6}\ D_s$	0.251	0.501	0.0159	0.3617	0.9572
GWR		N = 5	0.5422	0.7363	0.1176	0.5366	0.9113
		N = 10	0.289	0.5376	-0.0064	0.4081	0.9515
		N = 15	0.3119	0.5585	-0.0026	0.4509	0.9469
		N = 20	0.2496	0.4996	0.028	0.3939	0.9588
		N = 25	0.2633	0.5131	0.0298	0.4132	0.9562
		N = 30	0.2679	0.5176	0.0762	0.4179	0.9592
		N = 35	0.2569	0.5068	0.0688	0.4034	0.9595
ANN		MLP 3 - 3 - 4	0.1992	0.4464	-0.0153	0.3478	0.9663
		MLP 3 - 4 - 3	0.1758	0.4193	-0.0158	0.3161	0.9703
		MLP 3 - 6 - 2	0.2351	0.4849	-0.0186	0.3455	0.9607
		MLP 3 - 8 - 1	0.2601	0.51	-0.005	0.3689	0.9556
		MLP 3 - 10 - 1	0.3403	0.5833	-0.0816	0.4592	0.9435
		MLP 3 - 15 - 1	0.3975	0.6304	-0.127	0.4984	0.9348
RK		GWR: N=20	0.1291	0.3592	-0.0582	0.274	0.9692
		LR-OLS: z = f(q)	0.1089	0.33	-0.0124	0.1098	0.9797
		ANN-MLP 3 - 4 - 1	0.1336	0.3655	0.015	0.2505	0.9779
		SW	0.191	0.347	0.0171	0.2677	0.9676

Table 5.8: Comparison of interpolation accuracy of the VAI methods applied to mean annual precipitation based on the five different statistical indexes

by Prudhomme and Reed (1999), the topographic and geographic factors that can influence the spatial distribution of mean temperatures are:

- Minimum absolute distance from the sea, d_{min};
- Angle α formed by the minimum distance vector and the South;
- Distance d_i from the sea in the eight cardinal directions, i ;
- Azimuthal angle β_i of the horizon in the eight cardinal directions.

The latter variable plays the role of an obstruction factor, according to the original definition of Faulkner and Prudhomme 1998, and is defined as the angle subtended by the highest topographical barrier in the ith direction, such that $\tan \beta_i = \Delta H / \Delta X$, with ΔHas the difference of elevation between the barrier and the station, and X as the distance between the two points. Note that d_{min} is not necessarily one of the eight distances di taken in the cardinal directions, because it represents the minimum distance between the station point and the coastline. The above-mentioned variables were computed for each of the considered stations see Fig. 5.2 using a geographical representation of the Sicilian coastal line and a digital elevation model DEM with a resolution equal to 100 m.The previously described variables were subsequently

transformed and averaged in different ways, producing three parameters:a distance measure D_s, an aspect variable A_s, and a concavity index C.

The distance measure D_s geometric mean of the distance from the sea in the eight cardinal directions is given by

$$D_s = \sqrt[8]{d_1 * d_2 * \ldots * d_8}. \tag{5.11}$$

The second parameter is a combined measure of aspect orientation and sea proximity:

$$A_s = \frac{10\, cos\alpha}{1 + d_{min}} \tag{5.12}$$

The third index is a concavity index, obtained by weighting the azimuthal angle $i$in the eight directions:

$$C = \sqrt[8]{\prod_{i=1}^{8} 10^{2\, tan\, \beta_i}} \tag{5.13}$$

Only a few authors have considered the effect of orographic barriers on average temperatures. Gentilli 1959 considered the shape of the terrain and noticed that, at the same elevation, prediction in areas with concave topography resulted in negative temperature anomalies i.e., cooler terrain because of cold air stagnation in the concave areas. He defined a qualitative topographic index in an attempt to improve temperature-elevation regressions. Faulkner and Prudhomme 1998, and less explicitly Ninyerola et al. 2000, referred to "obstruction" factors with respect to wind directions in their analyses, but no explicit consideration of a concavity effect on air temperature was found in the literature.

Once the indices has been obtained, the mean annual temperature for the training set was related to the morphological parameters and elevation and latitude, by means of a stepwise regression procedure. This procedure produces multiple regression models of increased complexity, according to the number of independent variables considered. Apart from elevation and Ds, the other value of geographic information in the explanation of mean annual temperature is questionable. Finally the the best model for mean annual temperature is that based only on elevation q and Ds (geometric mean of the distance from the sea in the eight cardinal directions), as follows:

$$\hat{T}_a = 18.68 - 0.0057\,q + 8 * 10^{-6} D_s \tag{5.14}$$

The indeces provided by the methods are satisfactory, both in terms of RMSE and MBE. The CC value equal to 0.96 confirms a good agreement between estimates and observations.

The fourth VAI method is an ANN using MLP structure. In order to define such an architecture, the training algorithm and the learning rate are fixed during the training phase. The ANN has been trained by the scaled conjugate gradient algorithm (Moeller, 1993). The initial weights have been extracted randomly from a standardized normal distribution and the value of the learning rate, initially equal to 0.04, decreases according to an exponential law. This technique, called *early stopping* (Bishop, 1995), has been used in order to decrease the risk of overfitting and training is stopped after only 100 iterations. Six network architectures, having different number of neurons in the hidden layer, have been compared and, on the basis of the RMSE and MBE performances (Table 5.8), the best one is the network with one hidden layer having 4 neurons with hyperbolic tangent activation function (MLP 3-4-1) both in terms of RMSE and MBE values. Moreover, the correlation coefficient between estimated and measured data for the validation set is equal to 0.97, confirming a good agreement between estimates and observations. The correlation coefficient between residuals and measured data of the validation set, equal to -0.27, shows that not all the spatial data correlation has been explained by the ANN, also implying that overfitting has been avoided.

Another evaluation of the network's performances has been completed by computing the Akaike's FPE (*final prediction error*) (Larsen et al., 1994). This is a validation index that uses the information from all the second order derivatives of the error function to remove unimportant weights from a trained network (*pruning technique*) allowing to identify the best network architecture. In order to find the optimal MLP architecture, taking into account generalization error, the pruning technique called *optimal brain surgeon* (OBS, Hassibi and Stork., 1993) has been applied also in this case. It allows to improve the generalization capability of the MLP determining the significance for each weight, which is an estimate of the variation in the FPE index if the corresponding weights is pruned. The goal of OBS is the pruning of the weights with the smallest importance and the optimal network is the one whose weights yield the smallest FPE. Figure 5.13 shows that the minimum FPE is obtained after pruning only 1 weights and this confirms that the MLP architecture is well chosen.

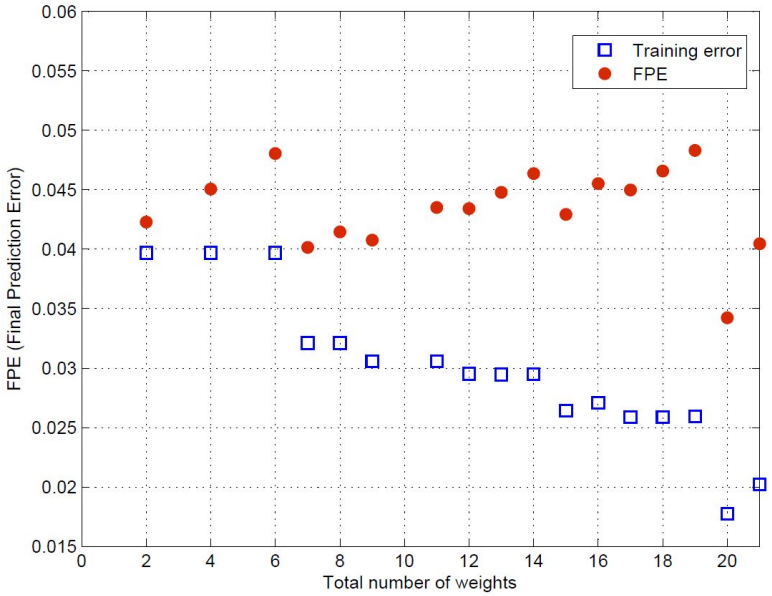

Figure 5.13: Training error and FPE plots as a function of the network parameters (weights). The abscissa from right to left points out the number of pruned weights

Taking into account the values of the correlation coefficients between residuals and measured data of the OLS regression and GWR, respectively equal to-0.40 and -0.17 one can observe that,even if in a lighter way respect to the case when the same methods are applied to the rainfall,not all the spatial data correlations have been explained by these above mentioned methods. The same observation can be done for the MLP with 4 neurons in the hidden layer whose correlation coefficient between residuals and measured data, as said above, is -0.30. In particular the OLS regression is the methods that less than other explain the deterministic component. This is demonstrated by the fact that the obtained residuals have a modest spatial correlation. On the contrary, the GWR permits to accurately explain the deterministic component providing a good estimate of z, as can be noted from the correlation coefficient between residuals and measured data value.

Starting from these considerations, the fifth VAI method (the residual kriging, RK) has been used and tested. The basic idea of RK is, as mentioned in section 2.3, that to estimate the residuals coming from a VAI method (LR, GWR, SW and ANN) at an ungauged site and then to sum them to the estimate previously obtained by underlying deterministic interpolators.

In this case, an undersized MLP has not been chosen, since the neural network taken into account has a fairly limited number of neurons.

The OK is applied to the residuals derived from the LR, GWR, SW and ANN (MLP 3-4-1). In this case, as for the rainfall, an isotropic behavior in the residuals coming from the three different methods has been noticed through the derivation of experimental semivariograms.

As described inSection 2.3, the estimated values of residuals have been added to the LR, GWR, SW and ANN, respectively, according to eq. (.....). In table 5.8 one can observe that most of the performance of these three methods are better than the univariate and the other VAI methods. The best method, in terms of RMSE, MBE and CC, appears to be the RK-LR (with OLS).

Figures 5.14 and 5.15 show the performances of the best methods respectively in terms of RMSE and MBE for both univariate and VAI methods. The best univariate method, in terms of RMSE and MBE, is the RBF with N=30. The IDW behavior is opposite to the RBF one and is characterized by higher RMSE and MBE. The OK gives RMSE and MBE values higher than that provided by the RBF. However all univariate methods are very biased providing a systematic underestimation of z. In order to improve the results in terms of MBE, it is necessary to switch from univariate methods to the VAI ones, by introducing the elevation information, in the case of LR, GWR and ANN methods, and by introducing the elevation and morphologic information, un the case of SW. For example, also in this case, the LR, as shown in Figures 5.14 and 5.15, is a method with a low bias but high value of RMSE. The GWR, ANN and SW give better results with acceptable value of RMSE and MBE. Except for the RK-SW application, the best methods are however the RK and in particular RK-LR, that gives the lowest value of RMSE, the highest CC and an acceptable value of MBE, and RK-ANN (MLP 3-4-1) with acceptable values of RMSE and CC and the lowest value of MBE.

A deep analysis of Figures 5.14 and 5.15 and table 5.8 shows that the application of ordinary kriging to the residuals coming from any VAI methods decreases in substantial manner the RMSE values of the underlying VAI method (from 0.50 to 0.36 for GWR; from 0.49 to 0.33 for LR-OLS; from 0.50 to 0.43 for SW; from 0.42 to 0.36 for MLP 3-4-1) but, at the same time, increases the bias of the same method (from 0.028 to -0.058 for GWR; from -0.0014 to -0.0124 for LR-OLS; from 0.015 to 0.017 for), except for MLP 3-4-1 (from -0.016 to 0.015).

Here the RK-LR (with OLS) method has been chosen as the best method even if one can observe that a RK-LR presents value of MBE denoting a systematic un-

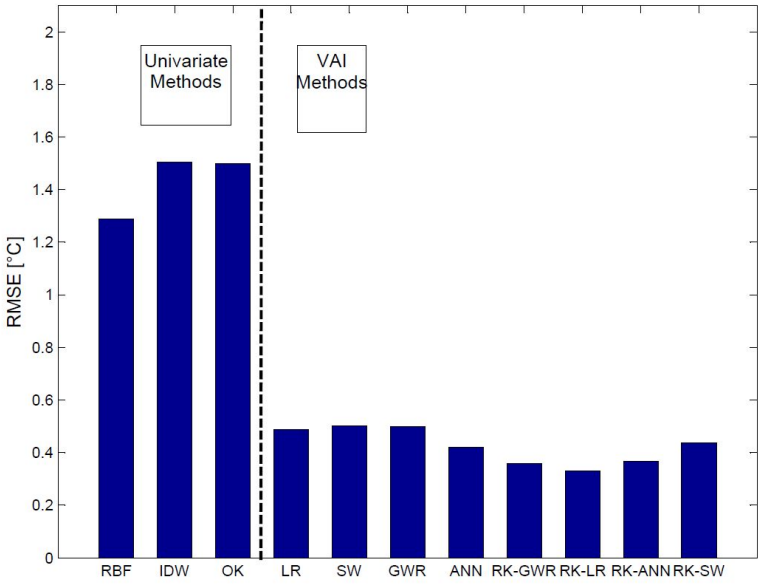

Figure 5.14: RMSE of best univariate and VAI interpolation methods

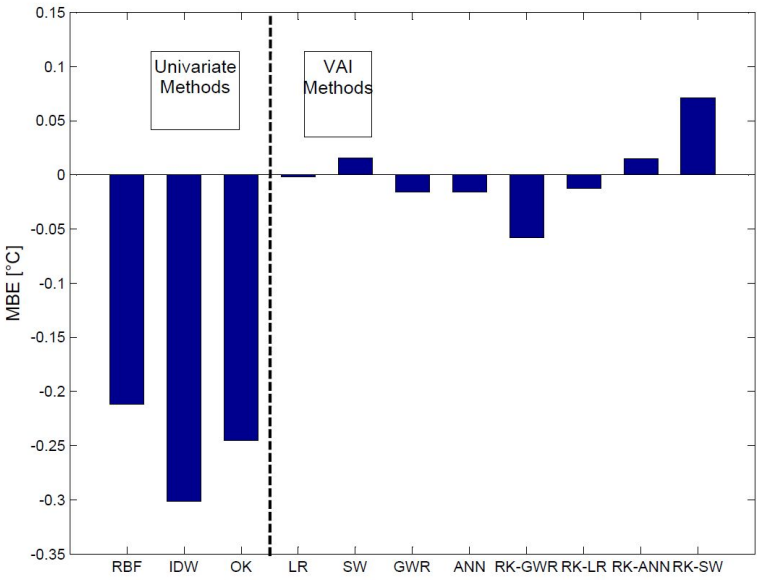

Figure 5.15: MBE of best univariate and VAI interpolation methods

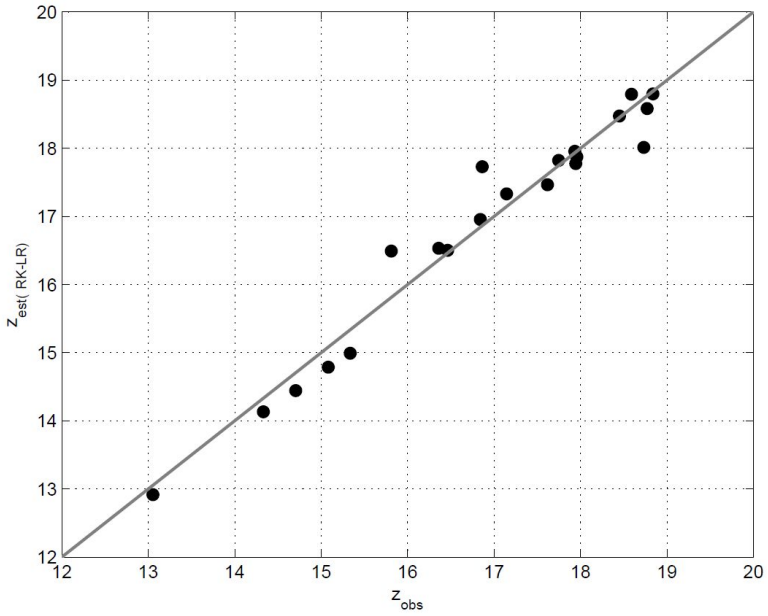

Figure 5.16: Scatterplot between observed annual temperature and annual temperature estimated with the RK-LR method

Figure 5.17: Mean annual temperature interpolated using RK-LR z -q method

derestimation of z; this underestimation can be also observed from the plot shown in Figure 5.16 which compares the observed average annual precipitation z_{obs} with the average annual temperature $z_{est\,(RK-LR)}$ estimated with the RK-LR method. Figure 5.17 shows the mean annual temperature map produced by RK-LR (OLS) method.

5.4.3 Monthly analysis

The interpolation methods previously described have been also applied to the average monthly temperature data for individual months, taking into account the results obtained in the case of average annual temperature data.

In order to study the temperature regime, another method is suggested. In this method a curve of the 12 monthly mean temperatures within the year is taking into account to estimate the average monthly temperature for each months. A detailed description of this method is presented in the next subsection.

5.4.3.1 Fitting mean monthly temperature by Fourier series: study of the temperature regime

The temperature regime is meant to be the curve of the 12 monthly mean temperatures within the year. Rather than estimating the monthly temperature normals by separate regression equations e.g., Zheng and Basher 1996 the whole parametrized curve of the temperature regime is estimated here at each grid point in the Sicilian territory. In this way it try to build a single model that can demonstrate possible causal relations between some geographical predictors and the within-year variability of temperatures. Such relations are not evident when regression models are built for individual months.

The sequence of 12 monthly temperature normals $T(j)$ can be well approximated by means of sinusoidal curves obtained by Fourier series

$$T(j) = A_0 + \sum_{i=1}^{N} A_i cos\left(\frac{2\pi}{\tau_i} j + \frac{\tau}{\tau_i}\phi_i\right) \tag{5.15}$$

where j are the months of the year (1–12); A_0 is the mean of $T(j)$; τ (=12) is the fundamental period of the cycle; N is the number of the harmonics; A_i is the amplitude, i is the phase and i is the period of the ith harmonic. For the estimation of the Fourier series parameters the cosine argument in Eq. 5.15 can be decomposed to a polynomial form:

$$T(j) = A_0 + \sum_{i=1}^{N} \left[b_i * cos\left(\frac{2\pi}{12} j\right) + c_i * sin\left(\frac{2\pi n_i}{12} j\right)\right] \tag{5.16}$$

where $b_i = A_i cos(n_i \phi_i)$; $c_i = -A_i cos(n_i \phi_i)$; and A_0 is the parameter that can be estimated by least squares, and where $n_i = \frac{\tau}{\tau_i}$. The amplitude and phase of the ith harmonic can then be obtained as:

$$A_i = \frac{b_i}{cos(n_i \phi_i)} \tag{5.17}$$

$$\phi_i = arctan\left(-\frac{c_i}{b_i}\right) \tag{5.18}$$

Having selected a model to estimate the spatial variability of the mean annual temperatures, only the monthly deviations from the annual mean were analyzed, considering two alternatives:

- A nondimensional temperature regime $t(j)$, where

Station	q (m a.s.l.)	L (deg)	T_a (°C)
ISOLA DELLE FEMMINE	0	38.20	19.42
SIRACUSA	0	37.06	18.67
CAPO S.VITO	0.2	38.19	18.86
CATANIA	13.08	37.50	18.55
LICATA	71.96	37.10	18.20
AGRIGENTO	184.04	37.31	17.98
CASTELVETRANO	201.4	37.69	17.96
CIMINNA	476.72	37.89	16.09
CESARO'	1119.68	37.85	12.42
FLORESTA	1266.48	37.99	10.98

Table 5.9: Characteristics of the temeprature station considered in Fig. .....

$$t(j) = T(j)/T_a \tag{5.19}$$

with T_a as the mean annual temperature. From Eq. 5.19 one obtains that in Eq. 5.16 $A_0 = 1$ see Fig. 5.18a.

- A zero-mean temperature regime, where $t(j)$ is obtained as

$$t(j) = T(j) - T_a \tag{5.20}$$

The above-presented relation produces $A_0 = 0$ see Fig. 5.18b.

Ten stations that have substantially different mean temperature values and different geographic features within Sicily described in Table 5.9 were selected to compare these two alternatives. The shape of the zero-mean temperature regime curve Alternative 2 is less variable, from one station to another, than that of the other curve. Giordano 2002 showed that the model for representation of Alternative 2 was more accurate than the one for the first alternative. The assessment was made on the quality of reconstruction of the regime curves in all of the considered stations. Results from Alternative 2 were better in terms of R2 of the estimation of the Fourier coefficients and of the RMSE computed with the estimated curves. So, the zero-mean temperature regime Alternative 2 was used in the subsequent analyses.

The first attempt to reconstruct the zero-mean temperature regime was made with a one-harmonic Fourier series (F1H), with $\tau_1 = 12$ months.

Parameters A_1 and ϕ_1 was estimated for each temperature station minimizing this equation:

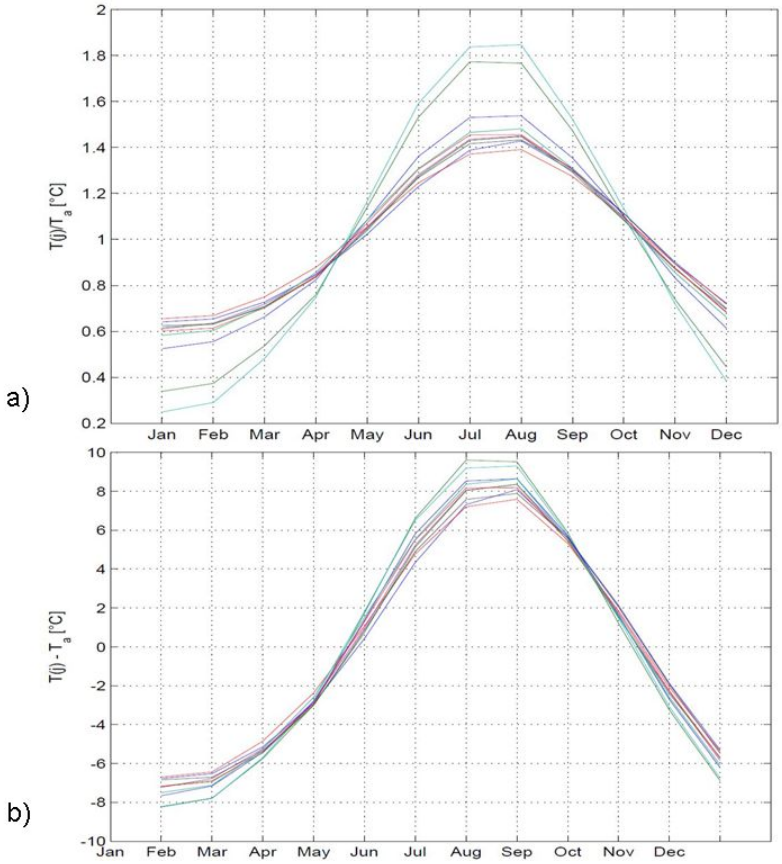

Figure 5.18: Diagrams for the ten stations reported in Table 5.9: a) nondimensional temperature regime; b) zero-mean temperature regime

$$\left(A_0 + \sum_{i=1}^{N} A_1 cos\left(\frac{2\pi}{\tau_i} j + \frac{\tau}{\tau_i}\phi_1\right) - T_{obs}(j)\right)^2 = 0 \tag{5.21}$$

and estimate values $t(j)$is obtained by

$$t(j) = A_0 + A_1 cos\left(\frac{2\pi}{12} j + \phi_1\right) \tag{5.22}$$

The parameters were then correlated with the stations' geographical and morphological parameters (q, L, D_s, A_s, C), before defined, using the stepwise procedure. The most efficient models found for amplitude and phase are

$$A_1 = -0.0011\, q + 0.3024\, L + 0.000116\, D_s - 18.69 \tag{5.23}$$

$$\phi_1 = -0.0000588\, q + 0.000000759\, D_s - 0.8385 \tag{5.24}$$

Even though the model results are fair, significant errors occur in correspondence to the highest and lowest values of the curve, in particular when these values persist for two or more consecutive months. To improve the representation, a second Fourier wave (F2H) with τ_2 =6 months was introduced

$$t(j) = A_0 + A_1 cos\left(\frac{2\pi}{12} j + \phi_1\right) + A_2 cos\left(\frac{4\pi}{12} j + 2\phi_2\right) \tag{5.25}$$

In this case, least-squares regressions of the four parameters on all of the stations produce estimates of A_1 and ϕ_1 identical to those obtained considering only one harmonic. The analysis on the second harmonic parameters shows that A_2 and ϕ_2 can be considered constant in space, with an average value of 0.1158 and 2.0293 respectively:

$$A_2 = 0.1158 \tag{5.26}$$

$$\phi_2 = 2.0293 \tag{5.27}$$

The final model for the temperature regime in Italy is then represented by Eqs. 5.20 and 5.25, with the four parameters obtained respectively by Eqs. 5.23, 5.24, 5.26, 5.27. Comparing the table 5.10 a) and table 5.10 b), can be seen that the statistics values improve for the case of two harmonic Fourier series (F2H). The index are obtained taking into account the estimate values of each month within a year for

each station belonging to the test set.

Monthly mean temperatures are reconstructed using the following combined model:

$$\hat{T}(j) = T_a + t(j) = T_a + A_1 cos\left(\frac{2\pi}{12}j + \phi_1\right) + A_2 cos\left(\frac{4\pi}{12}j + 2\phi_2\right) \tag{5.28}$$

with T_a are the mean annual temperature.

5.4.4 Results monthly analysis

As before said, the interpolation methods with the best performance, described for the case of average annual temperature data, have been also applied to the average monthly temperature. Particular attention has been paid to the estimation of the semivariograms of both the average monthly data and the residuals coming from the use of three VAI methods (i.e. LR-ROB, GWR and MLP 3-4-1).

At monthly scale, the presence of a zonal anisotropy has not been observed in the average monthly data.

In table 5.11 an overview of the results obtained for the statistics is shown. The method leading to the best result is indicated for each month together with the corresponding value of the relative statistical indexes. With regard to the RMSE values, it can be observed that the best values of this index are always obtained through the application of a VAI method. In particular, in six of the months the lowest value of RMSE is achieved using the two step VAI method, i.e. the Fourier series application with the estimation of parametres by a SW approach. In particular the best performance in each months is achieved by the Fourier series with two harmonic (F2H). And for other six months, the best value of RMSE is given by RK-LR (OLS), RK-MLP and RK-GWR. The same observations can be approximately made analyzing the index MAE. With regard to the unbiasedness of the different methods, the best results are instead achieved with univariate applications, in particular with OK methods. The s-RMSE did not give meaning informations in this case, so here it is not shown.

The results obtained at annual and monthly scale have been used to fill the gaps in the temperature dataset when a monthly value or data of an entire year are missing for a gauge. Missing temperature data are estimated through spatial interpolation with the best method among those previously exposed. In particular the RK-LR (OLS) method for annual case has been chosen while for the monthly case F2H has been chosen also for those months in which the application of the same method does not lead to the best values of the statistical indexes.

a)

Station	MSE [°C^2]	RMSE [°C]	MBE [°C]	MAE [°C]	s-MSE
CARONIA	0.218	0.466	-0.026	0.413	0.001
SCILLATO	0.264	0.514	0.065	0.455	0.001
CAMPOFELICE DI FITALIA	0.410	0.640	-0.025	0.537	0.004
PALERMO	0.188	0.434	-0.027	0.399	0.001
S. GIUSEPPE JATO	0.359	0.599	-0.061	0.512	0.001
ERICE	0.419	0.648	0.087	0.538	0.002
CAPO S.VITO	0.337	0.581	0.056	0.485	0.001
BIRGI NUOVO	0.200	0.447	0.019	0.383	0.001
S. MARGHERITA BELICE	0.289	0.537	0.094	0.507	0.001
DIGA FANACO	0.409	0.640	0.135	0.532	0.004
PIETRANERA	0.319	0.565	0.134	0.457	0.002
LAGO GORGO	0.201	0.448	0.025	0.390	0.001
CANICATTI'	0.391	0.625	0.039	0.543	0.002
GANGI	0.394	0.628	0.068	0.533	0.003
CALTANISSETTA	0.412	0.642	0.017	0.541	0.002
DIGA RAGOLETO	0.340	0.583	0.042	0.531	0.002
CESARO'	0.516	0.718	0.031	0.625	0.010
CATENANUOVA	0.379	0.616	0.005	0.537	0.002
DIGA DON STURZO	0.366	0.605	0.067	0.526	0.002
CATANIA	0.365	0.604	0.039	0.520	0.001
TAORMINA	0.348	0.590	-0.067	0.492	0.001

b)

Station	MSE [°C^2]	RMSE [°C]	MBE [°C]	MAE [°C]	s-MSE
CARONIA	0.154	0.393	-0.026	0.336	0.001
SCILLATO	0.197	0.444	0.065	0.391	0.001
CAMPOFELICE DI FITALIA	0.330	0.574	-0.025	0.484	0.003
PALERMO	0.127	0.357	-0.027	0.322	0.001
S. GIUSEPPE JATO	0.285	0.534	-0.061	0.447	0.001
ERICE	0.342	0.585	0.087	0.488	0.002
CAPO S.VITO	0.259	0.509	0.056	0.412	0.001
BIRGI NUOVO	0.141	0.375	0.019	0.314	0.001
S. MARGHERITA BELICE	0.220	0.469	0.094	0.442	0.001
DIGA FANACO	0.335	0.579	0.135	0.475	0.003
PIETRANERA	0.252	0.502	0.134	0.400	0.001
LAGO GORGO	0.145	0.381	0.025	0.325	0.001
CANICATTI'	0.312	0.559	0.039	0.490	0.001
GANGI	0.320	0.565	0.068	0.480	0.003
CALTANISSETTA	0.334	0.578	0.017	0.476	0.002
DIGA RAGOLETO	0.267	0.517	0.042	0.466	0.001
CESARO'	0.432	0.657	0.031	0.572	0.009
CATENANUOVA	0.300	0.547	0.005	0.473	0.001
DIGA DON STURZO	0.290	0.538	0.067	0.461	0.001
CATANIA	0.288	0.537	0.039	0.464	0.001
TAORMINA	0.277	0.526	-0.067	0.434	0.001

Table 5.10: Index values obtained taking into account the estimate values of each month within a year for each station belonging to the test set; a) one harmonic Fourier series (F1H); b) two harmonic Fourier series (F2H)

indexes / months	MSE [°C^2]	RMSE [°C]	MBE [°C]	MAE [°C]	CC
Jan	F2H *0.15*	F2H *0.39*	OK *-0.39*	RK-GWR *0.29*	RK-GWR *0.99*
Feb	RK-OLS *0.19*	RK-OLS *0.44*	RBF *-0.40*	RK-GWR *0.25*	F2H *0.99*
Mar	F2H *0.07*	F2H *0.27*	OK *-0.29*	F2H *0.23*	F2H *0.99*
Apr	RK-MLP *0.17*	RK-MLP *0.41*	IDW *-0.25*	RK-OLS *0.27*	F2H *0.99*
May	RK-OLS *0.16*	RK-OLS *0.40*	OK *-0.21*	RK-OLS *0.28*	RK-OLS *0.96*
Jun	F2H *0.07*	F2H *0.26*	IDW *-0.15*	F2H *0.21*	F2H *0.99*
Jul	RK-GWR *0.20*	RK-GWR *0.44*	F2H *-0.46*	RK-OLS *0.32*	F2H *0.98*
Aug	RK-OLS *0.21*	RK-OLS *0.46*	F2H *-0.65*	RK-OLS *0.33*	RK-OLS *0.94*
Sep	F2H *0.03*	F2H *0.19*	OK *-0.35*	F2H *0.16*	RK-OLS *0.96*
Oct	RK-OLS *0.19*	RK-OLS *0.44*	RBF *-0.38*	RK-OLS *0.27*	F2H *0.99*
Nov	F2H *0.13*	F2H *0.37*	OK *-0.53*	RK-MLP *0.28*	RK-MLP *0.98*
Dec	F2H *0.10*	F2H *0.32*	OK *-0.47*	F2H *0.25*	F2H *0.99*

Table 5.11: Overview of monthly results; each box reports the model leading to the best result and the corresponding index value

The most common case observed also in the temperature dataset (as in the rainfall dataset) is the simultaneous lack of monthly and annual data. In this case the first step consists of the estimation of the annual temperature for the missing year followed by the estimation at monthly scale. Both these estimations are carried out considering all the available temperature station for the examined year. In the second step the monthly estimations have been corrected to make them congruent with the annual estimation, here, considered more reliable than the monthly estimates. This correction simply consists in the assessment of a corrective coefficient $\psi_{i,j}$ to apply to all the months of i-th year and for the j-th gauge; this corrective coefficient is given by the following relationship:

$$\psi_{i,j} = \frac{T_{ann,i,j}}{M_{i,j}} \tag{5.29}$$

where $T_{ann,i,j}$, is the annual estimation for the i-th year and for the j-th gauge and $M_{i,j}$ is the mean of monthly estimation $\mu_{i,j,k}$:

$$M_{i,j} = \frac{1}{2}\sum_{k=1}^{12} \mu_{i,j,k}. \tag{5.30}$$

The estimated temperature data together with the observed data have been collected in a relational data-base which allows the creation of reports with the same form of Hydrological Annals.

5.5 Conclusions

The comparison between the methods was done through the use of jack-knife validation method. From this comparison, it has been observed that, among the univariate methods for the precipitations, the best performance has been obtained with the ordinary kriging method. In fact the geostatistical methods, as kriging, unlike the simpler methods, such as inverse distance weighting, takes into account most of the spatial pattern which can be observed for rainfall data, providing, in this way, acceptable results in terms of accuracy and unbiasedness. This results are in agreement with the results of the previous studies that highlight that the best results for the estimation of precipitation can be reached by using geostatistic interpolation methods and not with deterministic techniques, since they ignore the pattern of spatial dependences prevalently observed for rainfall data.For the temperature, indeed, the best performance, in terms of accuracy and unbiasedness, has been obtained with the RBF method. The

spatial distribution of temperature in a region is characterized by greater uniformity than that of rainfall, so it is possible obtaining good results also with the application of a methods that ignore the pattern of spatial dependences of temperature.

Subsequently, the application of VAI (variables-aided interpolation) methods, which take into account the elevation of raingauges, in the case of precipitations, and the geographic and morphologic parameters of the temperature stations, in the case of temperatures, has been investigated starting from the simpler deterministic methods (i.e., linear regression and geographically weighted regression) and proceeding to more sophisticated ones. In particular, the application of ordinary kriging to the residuals, coming from to the deterministic methods, has shown that the introduction of the elevation information improves the performances of the VAI methods. Then it has been pointed out that, for regions characterized from a really complex morphology (Goovaerts, 1999, Diodato and Ceccarelli, 2005) as Sicily, it is really important to take into account the elevation information to carry out a reliable variables estimate.

On the whole, the best results have been achieved through the application of a combination of a VAI deterministic methods and a geostatistical method: the residual kriging of linear robust regression (ordinary least square regression, for temperature) or, alternatively, artificial neural network. This matter demonstrates that not all the deterministic component is definitely explained by the simple VAI methods and that the obtained residuals have still a spatial correlation.

The linear regression is the least sophisticated method among all the VAI used methods and it carries out a slight detrending of the deterministic component and provides residuals with a strong spatial correlation that can be still explained using the ordinary kriging method. Moreover, it can be point out that, while the residual kriging application improves the accuracy of the underlying deterministic methods, the application of the same method increases, unfortunately, the bias of the deterministic ones (Prudhomme and Reed, 1999).

Regarding monthly mean temperatures, the spatial variation of the within-year pattern (temperature regime) is found to depend linearly ongeometric mean of the distance from the sea in the eight cardinal directions (D_s) and latitude in addition to the usual elevation predictors. The parameters of the two-harmonics Fourier (F2H) series reproducing the regime have been related to the above-mentioned predictors through a linear multivariate model. This made it unnecessary to have 12 different models, one for the mean temperature of each month. However, statistical indices, obtained with this application, are calculated in the same way as for the other methods (i.e, per month) to compare them each other. The best results in terms of accuracy

and unbiasedness, for the mean monthly temperature data estimation, are obtained by the F2H application.

Chapter 6

Analysis of results: Runoff estimate maps

This chapter describes the results achieved by the interpolation method proposed in the work of Sauquet et al. (2000) and explained in detail in the section 3.4. The points of departure are the considerations about the characteristics of the runoff and the stochastic interpolation procedure developed by Gottschalk et al. (1993 a,b). The presented approach is based on a disaggregation of the mean annual streamflow measured at the outlet of several basins finalized to the estimation of annual runoff on a target partition of these basins defined by the superimposition of a regular grid with a certain resolution. The technique is based on geostatistical interpolation procedure improved by a global constraint of water balance. The methodology is applied to 23 Sicilian basins constituted by 58 sub-basins. All the main basins taking into account have been previously grouped in three homogeneous sub-zone, as suggested by Cannarozzo et al. (1995).

A cross validation is performed, in order to validate the procedure applied in each homogeneous sub-zone. Finally six gridded maps are derived by applying the disaggregation twice to asses runoff on a increasingly finer grid mesh. The global constraint of water balance is applied to each element of a coarser mesh in order to give estimates of annual runoff within the finer one.

6.1 Case of study

In this study, an annual scale analysis has been performed to obtain the gridded maps of estimated average annual runoff.

The runoff observed data, used for the application of the interpolation scheme described in section 3.4, has been provided by OA-ARRA (*Osservatorio delle Acque - Agenzia Regionale dei Rifiuti e delle Acque*) (the former *Ufficio Idrografico Regionale* - UIR). For the application of this method, the everage annual runoff, the area of basins and the operation years of the gauging station are taking in account.

As said in chapter 4, the initially available hydrographic information has been reduced from 105 to 67 stations. This reduction has been made taking into account the instructions of the Bulletin 17B (Guidelines for Determining Flood Flow Frequency drawn from Department of the Interior U.S. Geological Survey), that suggest to eliminate gauging stations that have worked for less than 10 years, and taking into account the hystory of the stations (section 4.3). An analysis of variance of the total data set shows high annual variability between basins and a less strong annual variability between years. Nevertheless, if the total data set is used with records of different lengths there is an obvious risk that the runoff map may rather reflect temporal variability. In order to considered the period of years in which intermittence working of the hydrometric stations is to a lesser extend, a more limited window of time is taken into account. This period includes the years from 1960 to 2002 (instead of 1923-2002). Despite that, the dataset, used for this method, is characterized by a different leghths of records and such of this results inconsistent with the structure of basins, as will be discussed in the following.

The 67 selected gauging stations represent watersheds ranging from 10 km^2 (*Eleuterio at Lupo*) to 1782 km^2 (*Imera Meridionale at Drasi*). Mean annual runoff for the 67 study watersheds ranges from 39 to 885 mm, with an average of 204 mm. The overall area of the basins is equal to 15,050.2 km^2 that extents for all lenght of Sicily.

Because of the high exention of basins and because of heterogeneity that characterizes the climate and morphology in Sicily, a subdivision of the Sicily region has been made. In particular, this analysis has been performed dividing the island into three sub-Zones, as summarized in Figure 6.1, using the homogeneous regions suggested by Cannarozzo et al. (1995):

1. Zone 1. The most of the catchments (32) belongs to the sub-Zone 1, which is the Northwestern part of the island where the mean annual rainfall is around 680 mm, close to the regional value. The average area of the basins in this area

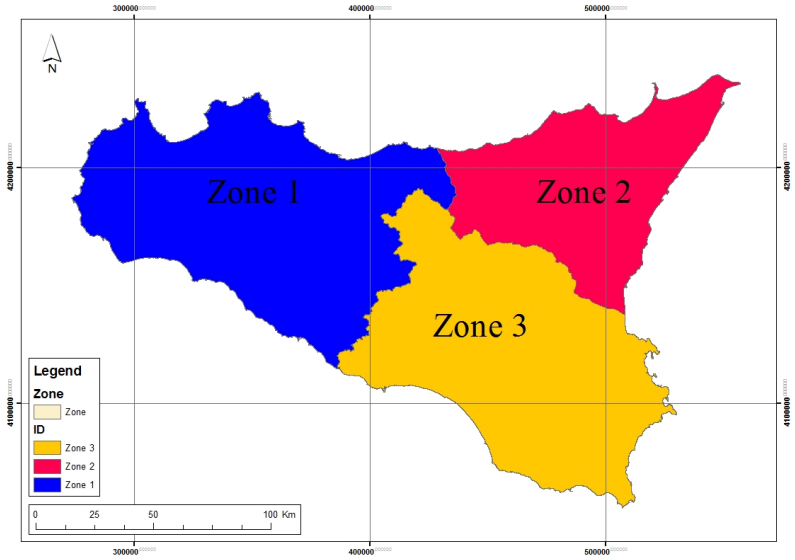

Figure 6.1: Sicily region subdivided in thre zone.

is 200 Km^2, ranging from 10 up to 1186 Km^2.

2. Zone 2. The sub-Zone 2 has the lower number of stations (12), but it is also the smallest sub-area. The mean annual rainfall is around 900 mm, higher than the regional value and the basins inside this zone are characterized by relatively small size and steep slopes, especially in the Northeastern part.

3. Zone 3. The sub-Zone 3 is located in the South- East part of the island and contains (14) stations. The average annual rainfall equal to 620 mm, is lower than the regional value and the average size of the considered basins is about 300 Km^2.

The homogeneity of these regions has been tested in terms of annual streamflow (Cannarozzo *et al.*, 2009) using the homogeneity test of Hosking and Wallis (1997).

In this study, the *Ficuzza* at *S. Pietro, Tellaro* at *Castelluccio, Cassibile* at *Manghisi, Asinaro* at *Noto, Anapo* at *S. Nicola* and *Lentini* basins are excluded from the set of basins taken into account. In fact, considering basins located far from the rest of the basins within the same zone could be affected the final estimation of local runoff, obtained with the application of this method. So the basins used for this application are 23 constituted by 58 sub-basins. In Figure 6.2 the considered basins within the previous mentioned zone are shown.

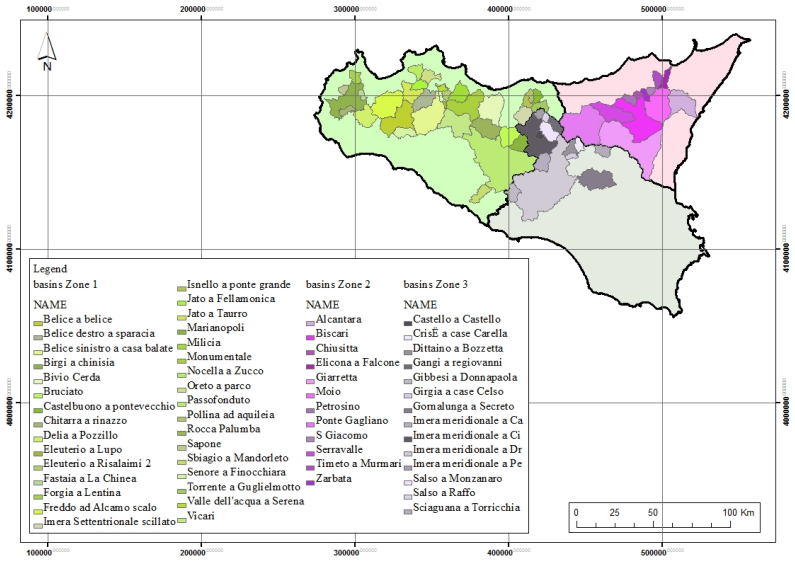

Figure 6.2: Catchments location for each zone

Moreover, the interpolation scheme requires a well-defined hierarchical structure of the river network to estimate the distance between the drainage basins and the aggregate covariance. From the geographical coordinates of the hydrometric stations outlet, the water catchment areas are marked off through the use of GIS thecniques (Geographic Information System).In this study, the DEM of Sicily with cell size of 100 meters has been used.

6.2 Procedure

As described in chapter 3, when applying this methodology, the process will be based on subsequent levels of hierarchization. The number of levels in this hierarchy is determined mainly by the amount of available observations, which also indicates the level of detail that can be achieved (size and number of fundamental units of the map). The first level in a larger drainage basin is usually already well defined by existing observation stations in the main rivers constituting the first level of sub-basins (*first level of hierarchization*). These basins are in their turn divided into a second level of sub-basins (or grid cells) (*second level of hierarchization*), and observation stations with appropriate basin scales are chosen as the background for the interpolation. The interpolation procedure guarantees that the water balance equation is satisfied so that

the sum of runoff from this second level of basins is equal to that of the first order basin accommodating them.

The first step is the denesting or disaggregation procedure to obtain the "*first level of hierarchization*" (section 3.4). In this case, the method is applied to the nested basins belonging to the three zone above defined (Zone 1, Zone 2 and Zone 3). Moreover, more precise information is needed to do for the case of study.

For the application of the denesting procedure (section 3.3), here it is necessary considering the presence of dams and diversion dams in the Sicily basins. In fact, the presence of these artificial infrastructures within a basin modifies the runoff regime. It is clear that, when a dam is present within a drainage basin, the runoff is arrested in the outlet section of the dam. The same consideration can be done for the diversion dams. In this case, only a part of the whole runoff, i.e., the 40%, arbitrarily fixed, is captured. Therefore, when the variation of the basin area, due to the realization of a dam, is not recorded in the Hydrological Annals, equation 3.8 (chapter 3) must be modified. For sake of clarity the equation 3.8 of the chapter 3, is here quoted:

$$\hat{x}(A_1) = \frac{(Ax(A) - A_2 x(A_2))}{A_1} \tag{6.1}$$

.

A corrective coefficient has been used in presence of a dam; in particular, the area of basin is divided for this coefficient, so that the reduction of surface, where the runoff is evaluated, is taken into account. The equation 6.2 will be, in this case:

$$A_1 \hat{x}(A_1) = \frac{(Ax(A) - A_2 x(A_2))}{\left(\frac{A_1}{A_d + A_{dd}}\right)} \tag{6.2}$$

where A_d is the surface at the dam and A_{dd} is the surface at the diversion dam.

Figure 6.3 shows the basin related to the station *Belice at Belice*, belonging to the Zone 1, where there are the artificial *Garcia* dam and a diversion dam.

Therefore Nsub-basins A_j contained in the group of considerated zone are obtained, with $j = 1, 2 \ldots, N$. The value of runoff for each sub-basin (now, non-overlapping basins) was obtained taking into account the network structure of the basin.In Figures 6.4, ... and ....,the average annual runoff estimated by the disaggregation procedure for different areas is shown. The yellow squared markers represent the gauging station in the outlet sections of the major drainage basins, while the red circular ones represent the other gauging stations (sub-basins of the major drainage basins). Moreover, the thick black lines are used to delineate the nested basin.

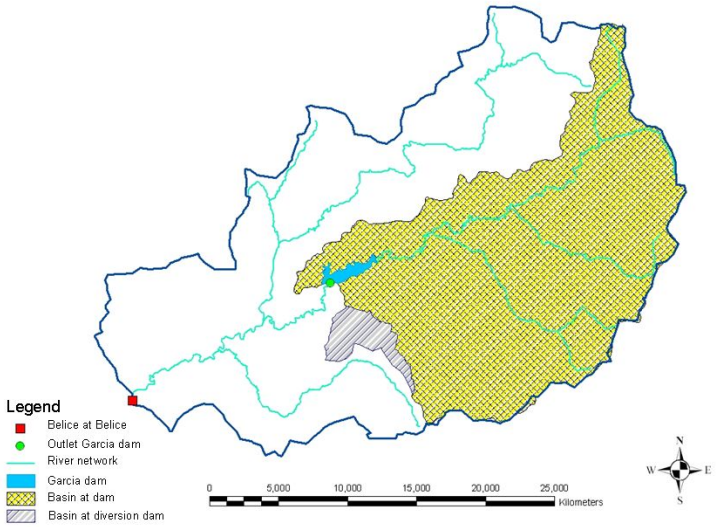

Figure 6.3: Belice at Belice

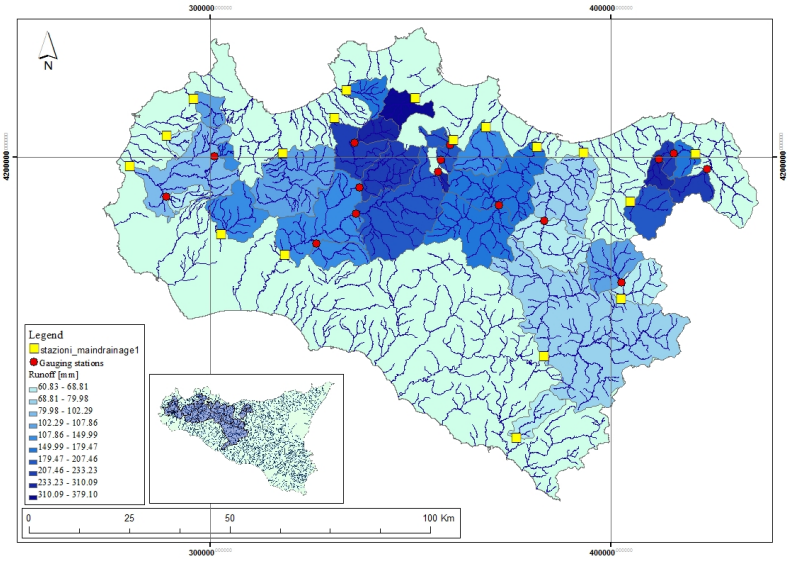

Figure 6.4: The average annual runoff estimated by the disaggregation procedure for the Zone 1. The yellow squared markers represent the gauging stations in the outlet sections of the major drainage basins, whereas the red circular ones represent the other gauging stations (sub-basins of the major drainage basins). The thick black lines are used to delineate the nested basin.

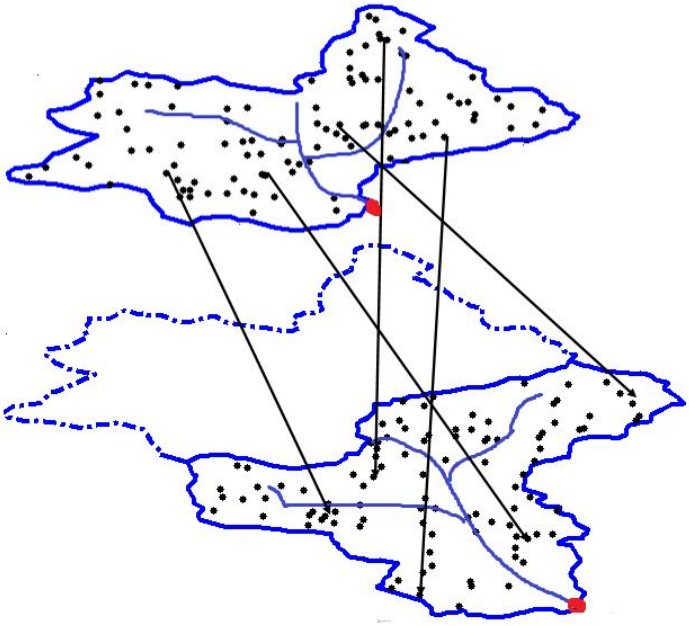

Figure 6.5: Schematic stream network and catchment boundaries with point pairs shown.

Applying the disaggregation method to all sicilian basins, the precence of negative runoff values has been observed in such basins. Such circumstance clearly shows an error in data capture and data processing delivered by UIR. Because of this, the data are not reliable and sometimes not useful for hydrological modelling. In this studied case, when the runoff values, obtained by the methods previously explained, are negative they are eliminated.

Then, the calculatation of the ghosh distance and the theorical and empirical covariograms (section 3.4) is made. In practice, for the ghosh distance calculation, 100 random point for each basins were produced and the euclidean distance between all random point in the area, taken pairwise, was made. Figure 6.5 shows a schematic of two nested catchments, the random points produced for each area and the distances between the points within the catchments.

Concerning to thechoice of the best therorical covariogram model, in order to fit the experimental covariogram, it is necessary carry out some consideration. For the runoff measured data used in this study a very singular correlation structure of spatially dependent observations is found. The classical theoretical covariance

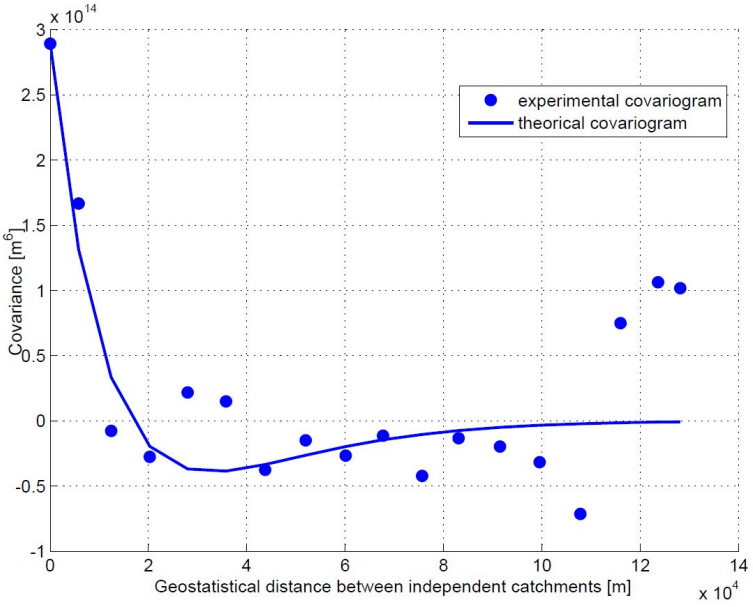

Figure 6.6: Experimental and theoric covariogram: Zone 1

functions have not been able to model this trend. Because of this, a novel and non-parametric equation of the covariogram function (Ploner and Dutter, 2000) is here suggested:

$$Cov(d) = C(0) * \left[\left(\left(1 - \frac{d}{T}\right) * exp\left(-\frac{d}{T}\right)\right)\right] \tag{6.3}$$

This model function can be fitted by minimizing the target function F for the parameter T:

$$F(T) = \left(Cov(d) - C(0) * \left[\left(\left(1 - \frac{d}{T}\right) * exp\left(-\frac{d}{T}\right)\right)\right]\right)^2 \tag{6.4}$$

where $C(0)$ is the covariance value at zero distance (variance).

The theoretical covariogram is represented with a lag equal to 8000 m. One covariogram for each zone has been found, as can be seen inFigures 6.6, 6.7 and 6.8.

In order to illustrate the principle of the disaggregation procedure, the interpolation scheme is applied to a target partition defined by the superimposition of a regular 8 x 8 km grid (64 km^2) over the catchments boundariesbelonging to the three different zones taken into account ("*second level of hierarchization*") (Figures 6.9, 6.10

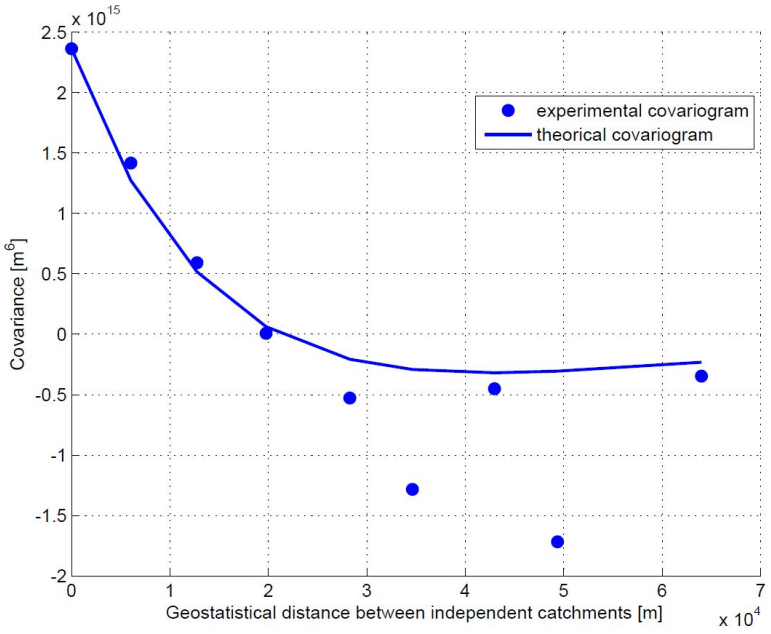

Figure 6.7: Experimental and theoric covariogram: Zone 2

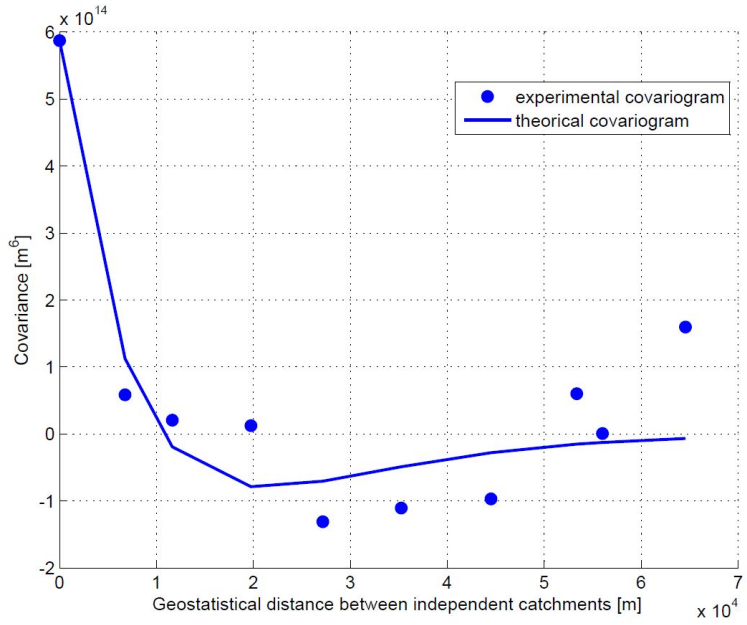

Figure 6.8: Experimental and theoric covariogram: Zone 3

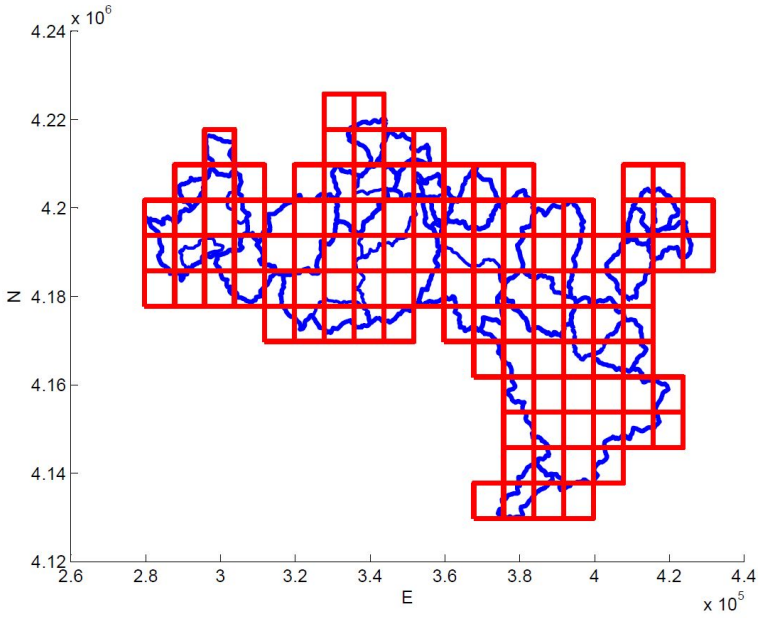

Figure 6.9: Total area A_T(basins belonging to the Zone 1) subdivided into M non-overlapping areas ΔA_i

and 6.11). In this case the point of departure is given from more than one drainage basin A_T. Now A_{T_k}, for each zone, are taking into account, where k is the number of outlet sections of the major drainage basins (rectangural markers in Figure 6.4, ... and .....).The procedure is the same that described in section 3.4. One difference is on the constraint in the equation 3.23 in chapter 3. Now there is a constraint for all major drainage basins in the considerate area ($n_{T_k} * q_{T_k}$) and k systems of equations 3.24 for each k constraints will be applied .

The interpolation procedure to assess runoff on 2 x 2 km cells (4 km^2) ("*third level of hierarchization*") is applied to the full data set and the interpolation constraint is kept within each 8 x 8 km cell so that the sum of streamflow from the smaller cells equals the streamflow from this larger one. The sum of streamflow from the larger cells is, in its turn, equals the streamflow from the total drainage basins, i.e. the sum of the streamflow of the basins belonging to the zone taken into account. The expected pattern of runoff structure is reproduced on the six maps (two for each zone).

The Figures (from 6.12 to 6.17) in the next show the obtained maps .

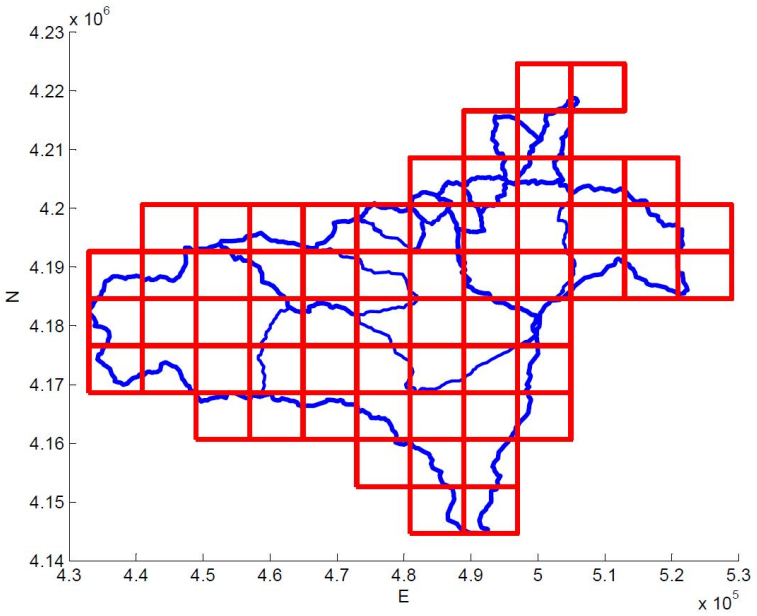

Figure 6.10: Total area A_T(basins belonging to the Zone 2) subdivided into M non-overlapping areas ΔA_i

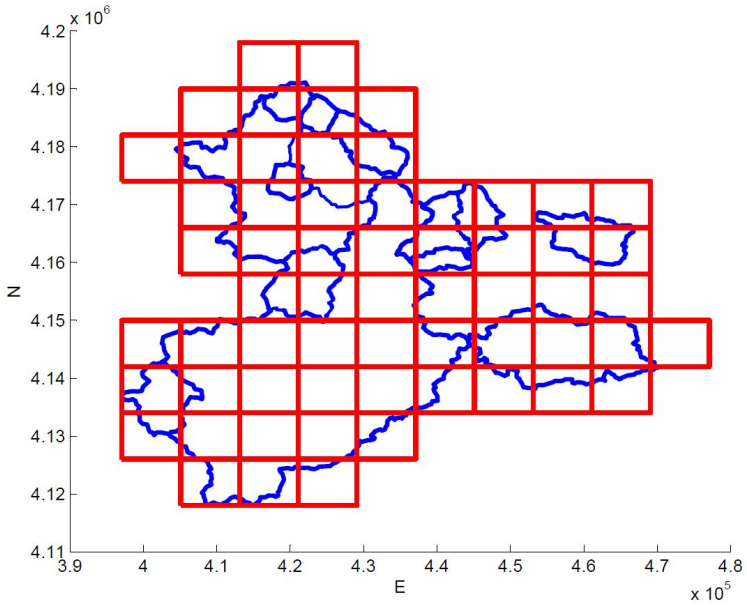

Figure 6.11: Total area A_T(basins belonging to the Zone 3) subdivided into M non-overlapping areas ΔA_i

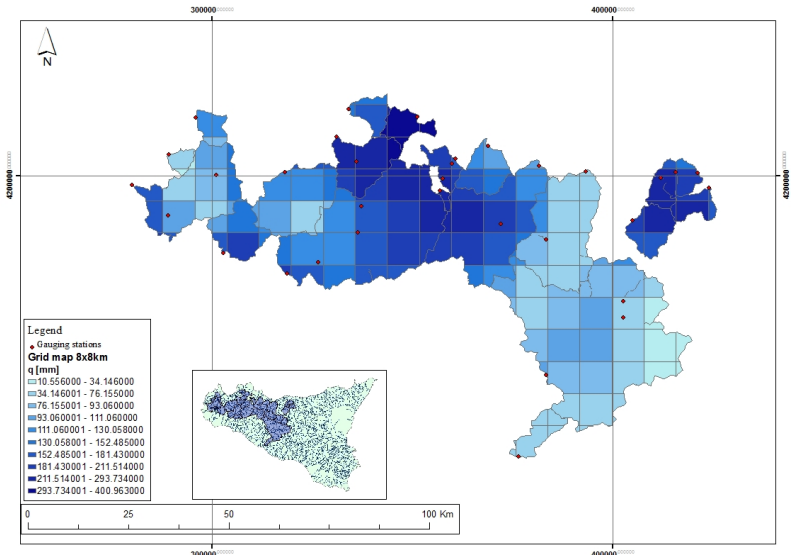

Figure 6.12: Gridded maps of average annual runoff with 8 x 8 km resolution. (Zone 1)

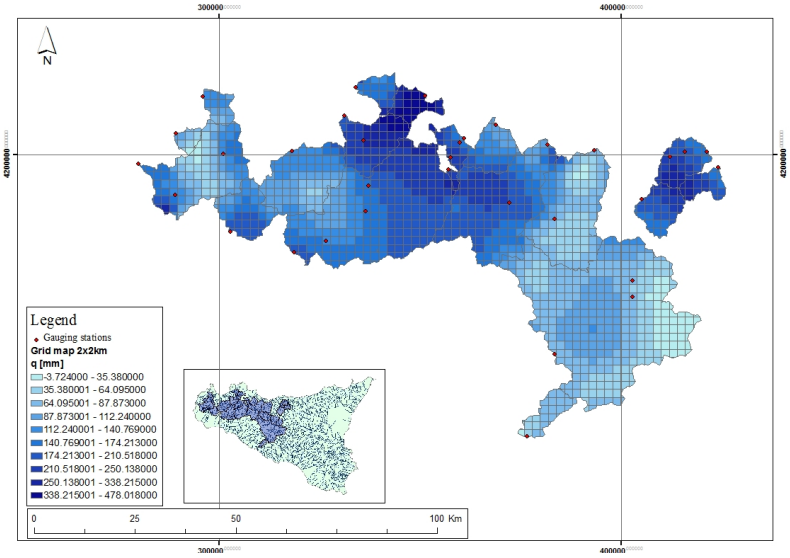

Figure 6.13: Gridded maps of average annual runoff with 8 x 8 km resolution. (Zone 1)

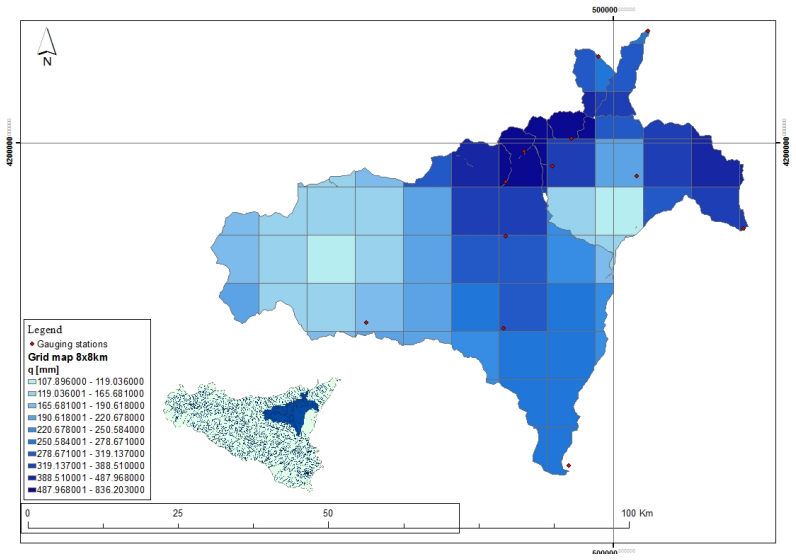

Figure 6.14: Gridded maps of average annual runoff with 8 x 8 km resolution. (Zone 2)

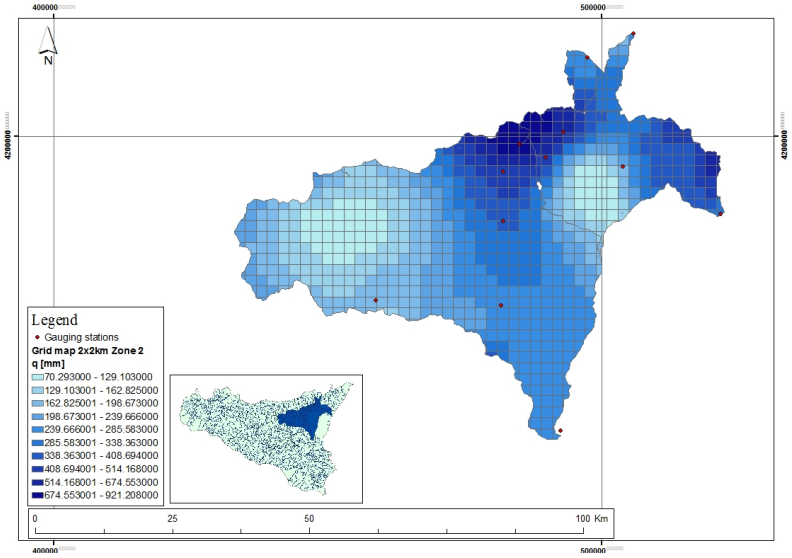

Figure 6.15: Gridded maps of average annual runoff with 8 x 8 km resolution. (Zone 2)

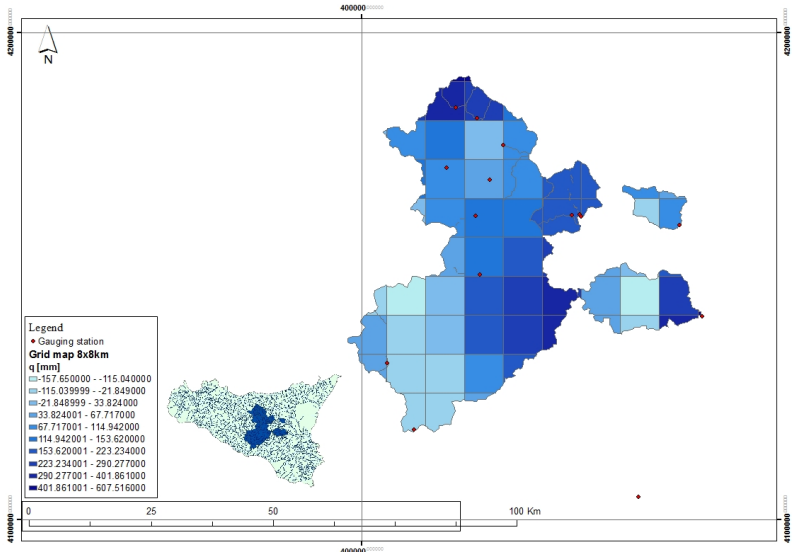

Figure 6.16: Gridded maps of average annual runoff with 8 x 8 km resolution. (Zone 3)

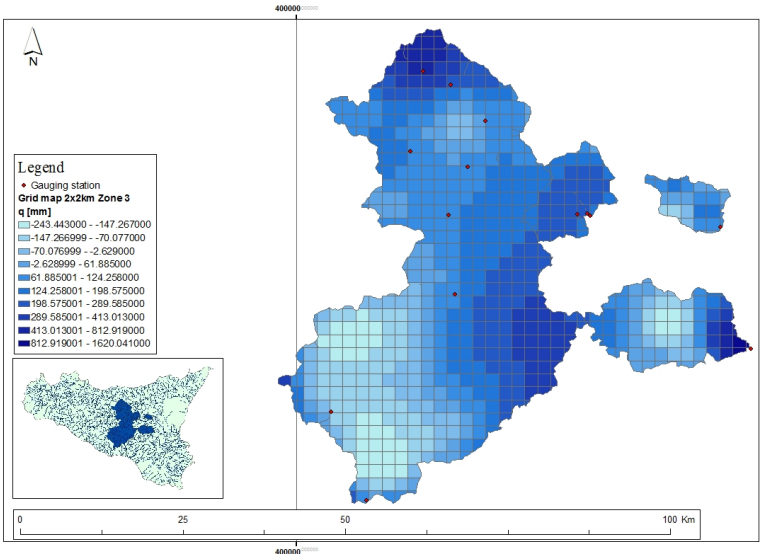

Figure 6.17: Gridded maps of average annual runoff with 8 x 8 km resolution. (Zone 3)

6.2.1 Validation

A cross validation is performed for this mapping technique, in order to validate the methodology.This analysis consists, for each area, of excluding one gauging station in turn from the network and then estimating runoff at this site by applying the interpolation procedure to the remaining stations with the covariance model 6.3 fitted to the full data set. The cross-validation analysis gives valuable insight in the real influence of the global constraint on runoff assessment. For each removed station, the observed runoff is compared to the estimated runoff produced by the watershed upstream this point. This analysis is applied to the first level of hierarchization.

The performance of the method is studied using regression analysis between the observed and the estimated runoff. Figures 6.18, 6.19 and 6.20 show the results of the validation analysis for the different zones taken into account.

Indeed, no objective validation can be proposed for the next level of hierarchization, because of the lack of reliable measurement. The maps can be analysed on visual agreement with observed runoff patterns.

6.3 Results

The performances of the interpolation method have been assessed using two indexes, taking into account the estimated values of runoff obtained with the validation procedure. In particular, the relative deviation (RD) :

$$RD = \frac{(q_{obs} - q_{est})}{q_{obs}} * 100 \quad (\%) \tag{6.5}$$

used to test what percentage of the estimated value differs from that observed,and the linear correlation coefficientCC:

$$CC = \frac{\sum_{i=1}^{N_v} [q_{obs}(\boldsymbol{x_i}) - q_{obs,m}][q_{est}(\boldsymbol{x_i}) - q_{est,m}]}{\sigma_{est}\sigma_{obs}N_v} \tag{6.6}$$

are used. In equation 6.6, $q_{obs,m}$ and $q_{est,m}$ are respectively the mean of measured and estimated runoff values and σ_{est} and σ_{obs} are respectively the standard deviation of measured and estimated runoff values.

Let's consider the performance of the method for the Zone 1 observing the Figure obtained by the validation procedure (Figure 6.18) . The CC is equal to 0.99. The results are quite sound; the regression line indicates an slight overestimation for the low values and a good estimation for the highest ones. So, the algorithm gives not very good when the lowest values of observed runoff are removed. In this case this overestimation of the observed runoff values can be seen when Baiata at Sapone and Birgi at Chinisia are removed from the data set. These sub-basin are characterized by the runoff values that are lower than the values of other sub-basins of the western area. In fact, the mean value of runoff of the above mentioned sub-basins is equal to 64 mm/year while the mean value of runoff of all sub-basin in the Zone 1 is equal to 170 mm/year. The highest absolute RD is for the Birgi at Chinisia sub-basin and is equal to -21%. Another high values of RD is obtained for the estimate value of runoff when the Fastaia at La Chinea are removed from the data set. In particular RD, in this case, is equal to 18%, suggesting an underestimation of runoff value.

Let's consider, now, the performance of the method for Zone 2 (Figure 6.19). The correlation coefficient is equal to 0.97. In agreement with results of the previous work (Sauquet et al., 2000), the regression line indicates an underestimation for high values, in paricular for the headwater catchments. In this case, the values of runoff of Chiusitta and San Giacomo was underestimated. The highest absolute RD of estimate is for the Chiusitta sub-basin and is equal to 23%, while the RD is equal to 13% when San Giacomo is removed from the data set. In particular the observed mean annual

runoff value is the highest values of the group of basins belonging to the Zone to and it is equal to 845 mm/year.

With reference to the performance of method for the Zone 3 (Figure 6.20), it is possible to observe that the results are quite worse than the results of the first case.The CC is equal to 0.96. In this case, an underestimate of the observed runoff values is noted when Imera meridionale a Petralia sub-basin is removed from the data set. It is an headwater catchments and its observed runoff values are equal to 565.04 mm/year while deviation of estimate value is equal to 23%. Another high values of RD is obtained for the estimate value of runoff when the Capodarso is removed from the data set. In particular RD, in this case, is equal to 24%, suggesting an underestimation of runoff value. Moreover, a very high overestimation for the low value of observed runoff is obtained, in contrast to the previous case.The latter situation happens when Gangi at Regiovanni and Gibbesi at Donnapaola sub-basins are removed from the data set. In this case the highest values of RD has been achieved: RD equal to 36% when Gangi at Regiovanni is removed from the data set and RD equal to 60% when Gibbesi at Donnapaola is removed from the data set.

As previously said, the hierarchical principle allows the calculation of gridded maps for finer and finer resolution satisfying the water balance. A first map was derived from the disaggregation of the mean annual discharge generated by the main basin (Figures 6.12, 6.14 and 6.16) and a second one was the result of the disaggregation of the runoff estimates yielded by each element of the first map (Figures 6.13, 6.17 and 6.15). The accuracy of the method observed on sub-basins partition is only expected on any partition that does not strictly respect the catchment delineation. Indeed, no objective validation can be proposed because of the lack of reliable measurement and the maps are analysed on visual agreement with observed runoff patterns.

Let's consider, the map 8x8 km obtained for the Zone 1. Comparing the Figure 6.9 with the Figure 6.12, one can observed one can observed a good agreement between the mean annual runoff values interpolated by a grid 8x8 km and the observed runoff patterns. In particular, for the Zone 1, both in gridded map of average annual runoff with 8 x 8 km resolution and in gridded map of average annual runoff with 2 x 2 km resolution, the runoff distribution was well estimated. A good agreement between the mean annual runoff values interpolated by a grid 8x8 km and the observed runoff patterns (Figure 6.10 and Figure 6.14) can be observed for the Zone 2. For the Zone 3, instead, in the gridded map of average annual runoff with 8 x 8 km resolution and 2 x 2 km resolution (Figures 6.16 and .....), some negative values can be observed. These results can be attributed to an unreliable original data set of recorded runoff.

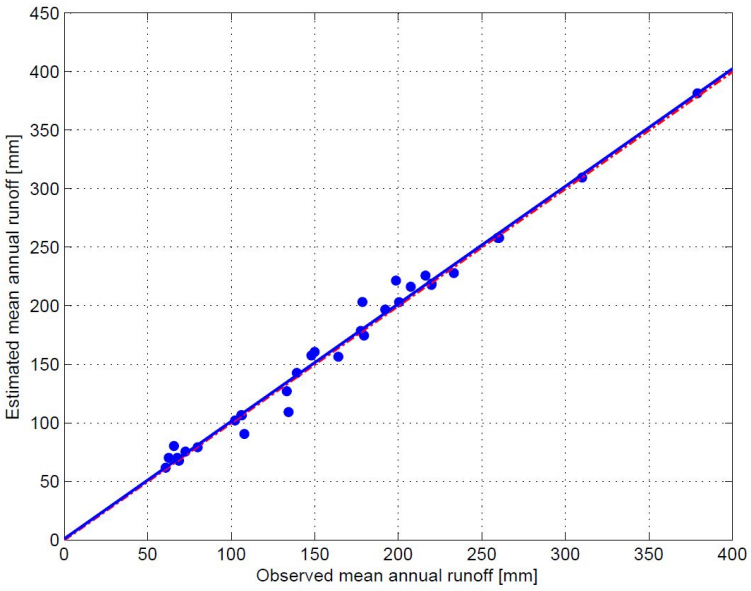

Figure 6.18: Cross-validation of the estimates mean annual runoff with observations for Zone 1 (the regression line (dashed) and the one-to-one line (black) are both represented).

In fact, the data coming from this group of basins, with the application of denesting method, have presented more negative numbers than those obtained in other zones.

In order to obtain estimated values of runoff also in a zone out of the group of the considerated basins and without any gauging station, the interpolation scheme is applied to the total area of the considered zone. The same procedure, above described, is applied to the basins belonging to the considered zone, whereas the area out of basins is treated differently concerning to constraint. In fact, for this area no constrain must be respected. Also in this case, a target partition defined by the superimposition of a regular 8 x 8 km grid (64 km^2) and a regular 2 x 2 km grid (4 km^2) over is used.

Observing the Figures 6.21 and 6.22 and considering the observed runoff patterns and the rainfall distribution in this zone, a good agreement is obtained between the distribution of the runoff in the map determined by the interpolation process and the actual runoff distribution. In particular, the map represents in a satisfactory way the runoff distribution in the northern and south -eastern zones, while it is not so close to the real distribution in the south-western part.

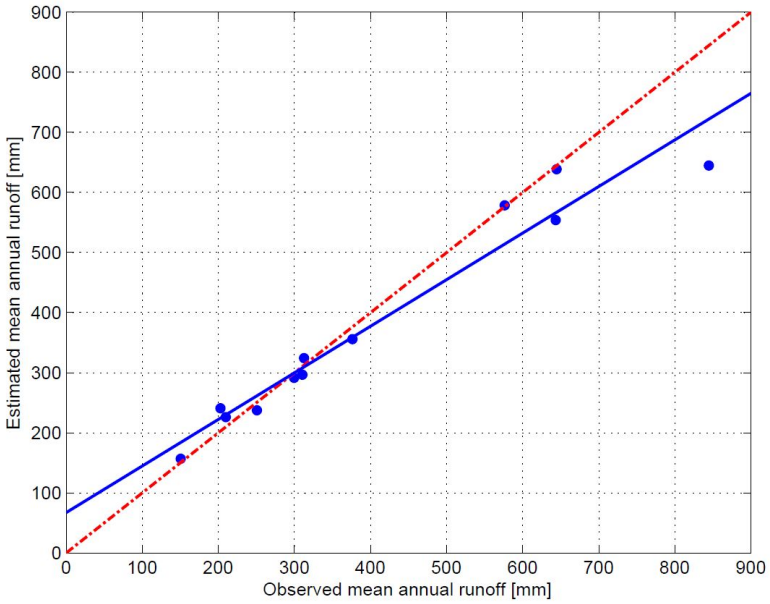

Figure 6.19: Cross-validation of the estimates mean annual runoff with observations for Zone 2 (the regression line (dashed) and the one-to-one line (black) are both represented).

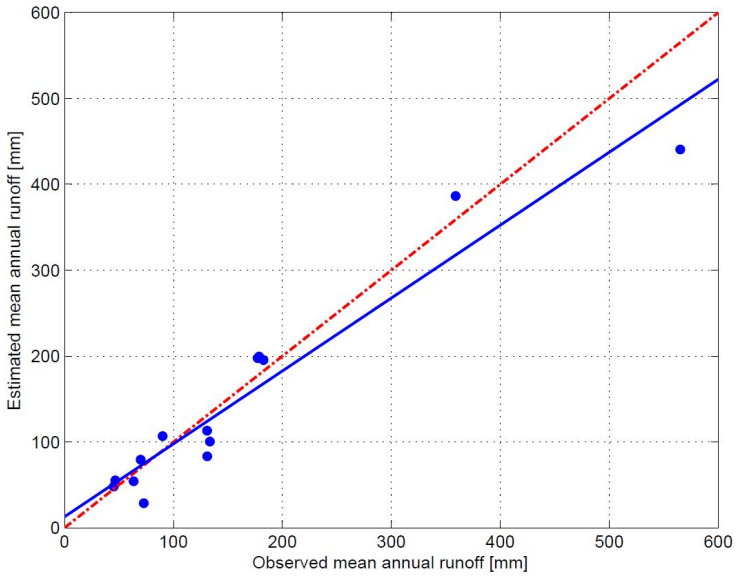

Figure 6.20: Cross-validation of the estimates mean annual runoff with observations for Zone 3 (the regression line (dashed) and the one-to-one line (black) are both represented).

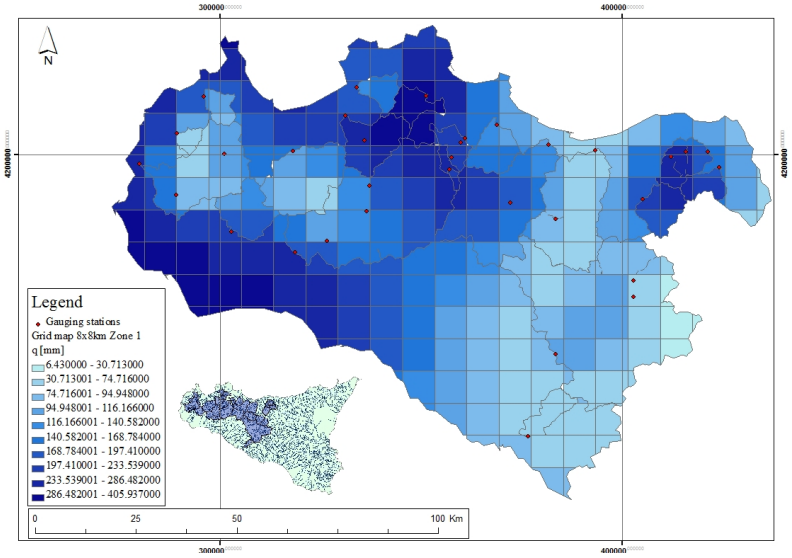

Figure 6.21: Gridded maps of average annual runoff in Zone 1 with 8 x 8 km resolution

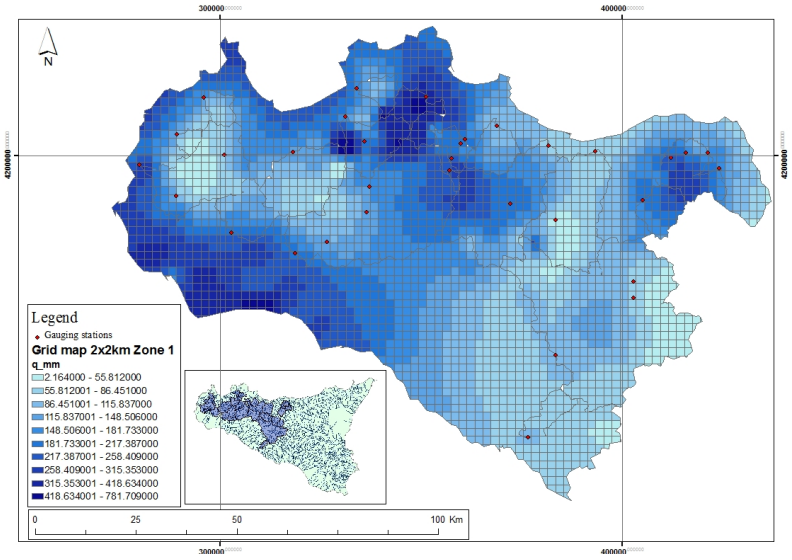

Figure 6.22: Gridded maps of average annual runoff in Zone 1 with 2 x 2 km resolution

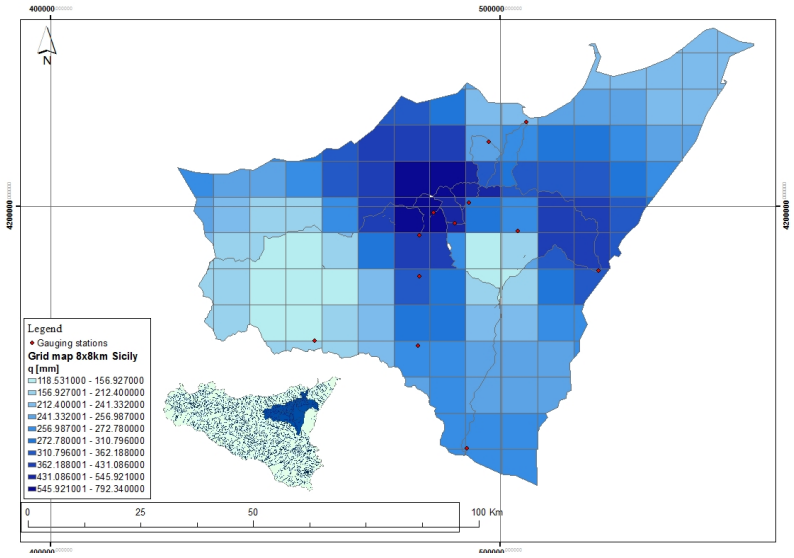

Figure 6.23: Gridded maps of average annual runoff in Zone 1 with 2 x 2 km resolution

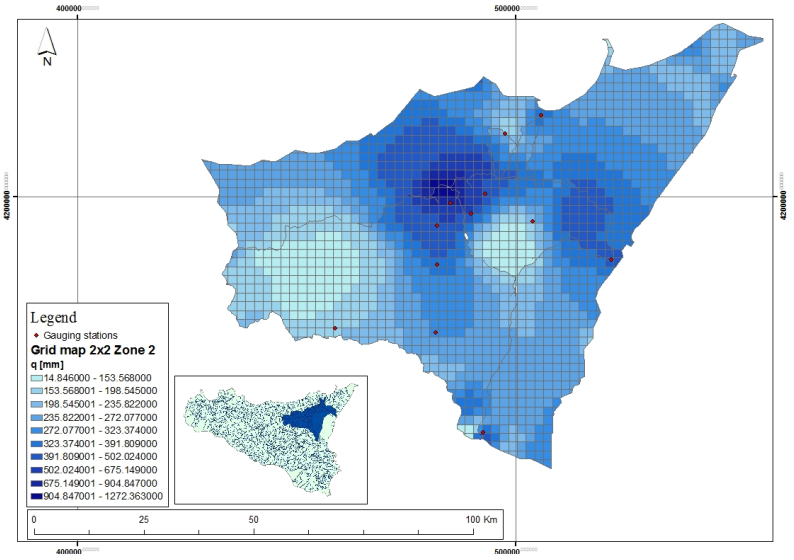

Figure 6.24: Gridded maps of average annual runoff in Zone 2 with 2 x 2 km resolution

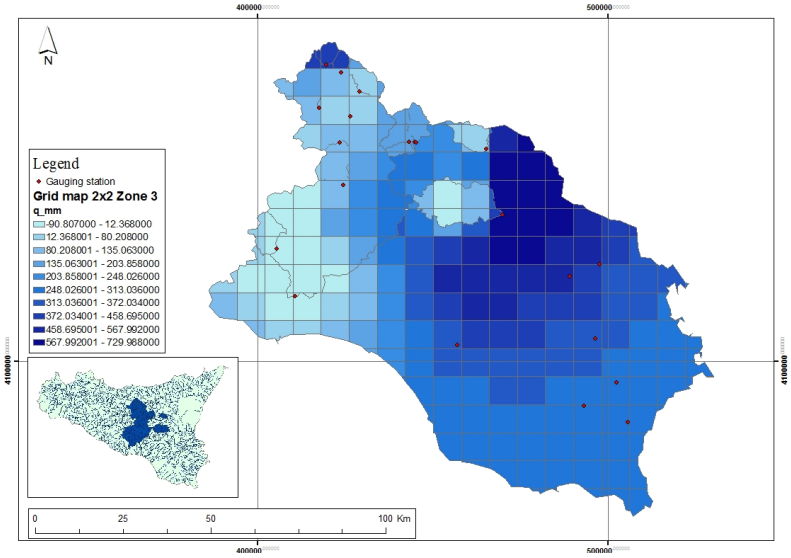

Figure 6.25: Gridded maps of average annual runoff in Zone 3 with 8 x 8 km resolution

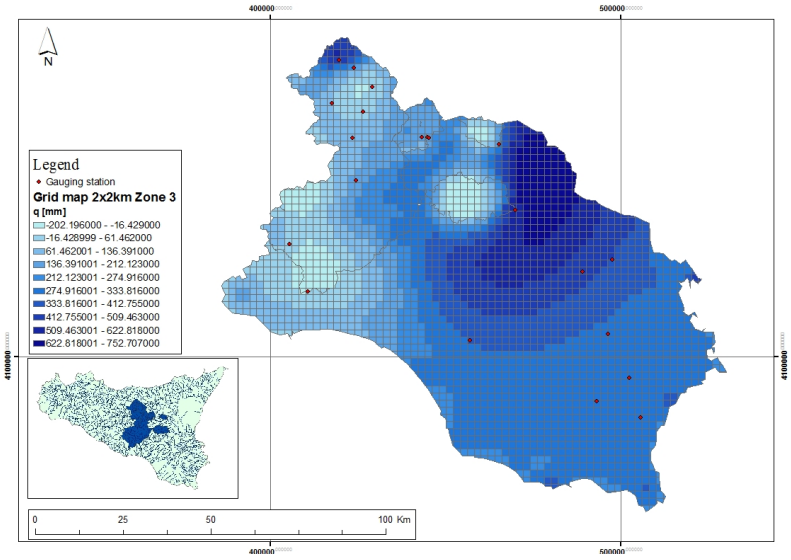

Figure 6.26: Gridded maps of average annual runoff in Zone 3 with 2 x 2 km resolution

6.4 Conclusions

For the runoff variable, estimated by a mapping technique, a cross-validation is performed to test the obtained results. It is important to highlight that in this case only an annual scale analysis has been performed to obtain the gridded maps of estimated average annual runoff.

The cross-validation analysis gives valuable insight in the real influence of the global constraint (one for each major drainage basins assessment) on runoff in the application of the method. On the contrary, the accuracy of the method on sub-basins partition (Gridded maps of average annual runoff with 8 x 8 km and 2 x2 km resolution) can be only assessed verifying if the partition strictly respects or not the hyerarchical structure of the catchment in comparison with the observed runoff pattern. Indeed, no objective validation can be proposed because of the lack of reliable measurement and the maps are analysed on visual agreement with observed runoff patterns. So, taking into account the informations obtained by the cross validation methods, for all groups of basins in the considered zones (Zone 1 - western area of Sicily), Zone 2 - eastern of Sicily and in Zone 3 - south-eastern of Sicily), the results are quite sound, in terms of real influence of the global constraints.But the regression line indicates an underestimation for high values (i.e. the headwater catchments) and an overestimation for the low values. So, the algorithm gives poor results when the observation removed from the dataset is one of the extreme values (highest or lowest); this situation agrees with the results of the previous work of Sauquet et al., 2000.

From the visualization of the maps, it is possible to observe that in Zone 1 and Zone 2 a good agreement between the observed and estimated runoff patterns. Furthermore, almost all the partitions strictly respect the hierarchical structure of the catchment. A different situation is encountered for Zone 3, where the presence of negative runoff estimated values demonstrates that the methods doen not succeed in reproducing correctly the runoff pattern and most of the partitions do not strictly respect the hierarchical structure of the catchment. This could be due to a low quality of the input data. Finally, the application of this methods gives the annual runoff estimated data for the stations that have been out of work in the chosen time window and that are characterised by a dataset affected by missing data. Moreover, since the hierarchical principle allows the calculation of gridded maps for finer and finer resolution annual runoff estimated values can be obtained also for the areas of the basin not provided with gauge stations.

Conclusions

The availability of complete and reliable datasets of hydrological variables is of paramount importance when dealing with both hydrological modelling and planning/ management issues. In order to cope with this problem, the study and analysis of different estimation methods, for filling missing data in time series of hydrological datasets, is carried out in this thesis. The following hydrological variables have been investigated: *precipitation*, *temperature*and *runoff*.

The reconstruction of dataset affected by missing data of precipitation, temperature and runoff, takes into account only the spatial structural dependence of these considered variables neglecting the spatial-temporal dependence.

A preliminary classification of the considered variables has been done on the basis of their possible representation by point or point/areal processes. Then, on the basis of the variables specifitied, different estimation methods have been considered, described and applied to solve the problem of missing data. In particular, for the variables as precipitation and temperature that are point processes, the following algorithms, used for the spatial interpolation, are applied: inverse distance weighting, radial basis function with thin plate spline, simple linear regression, multiple regression, geographically weighted regression, artificial neural network, ordinary kriging, residual ordinary kriging.

On the other hand, for runoff, the work stemmed from the consideration that it can be described as an areal process. With this assumption, a more accurate estimation of the considered variables can be obtained. This approach has very few examples in scientific literature but appears to be very promising in the considered field. For this reason, the estimation method chosen for the runoff is a stochastic method to construct maps with a geostatistical approach. It is, in particular, a stocastic interpolation system that can be assimilated to kriging system with the explicit consideration of the runoff variable as an areal process.

With regard to the precipitation and temperature, from the comparison of the methods, it is possible to highlight the following results:

- for the estimation of variables characterized by high spatial dependence, the methods that provide the best results are the geostatistical methods. In fact, among the univariate methods, the best performance has been obtained with a geostatistical method: ordinary kriging, for the precipitation, and with a deterministic method: RBF, for temperature. The geostatistical method provide the best results, because unlike the simpler methods, takes into account most of the spatial pattern which can be observed for rainfall data. From a comparison among the three univariate methods, in the case of precipitation it can be seen that with the application of RBF a very low value of RMSE is obtained but this in the same time this method provides the highest value of MBE, and so a sistematic underestimation of rainfall value. The worst results, for the precipitation, are obtained with the application of IDW.On the contrary, for the temperature, a deterministic methods provides the best results in terms of accuracy and unbiasedness. The spatial distribution of temperature in a region is characterized by greater uniformity than that of rainfall, so it is possible obtaining good results also with the application of a methods that ignore the pattern of spatial dependences of temperature.

- the morphology cannot be neglected when an interpolation of a climatic variable is carried out. The introduction of the elevation information improves the performances of the Variables-Aided Interpolation (VAI) methods significally in terms of unbiasedness both for the precipitation and the temperature. The application of VAI (variables-aided interpolation) methods, which take into account the elevation of raingauges, in the case of precipitations, and the geographic and morphologic parameters of the temperature stations, in the case of temperatures, has been investigated starting from the simpler deterministic methods (i.e., linear regression and geographically weighted regression) and proceeding to more sophisticated ones. In particular, the application of ordinary kriging to the residuals, coming from to the deterministic methods, has shown that the introduction of the elevation information improves the performances of the VAI methods. Then it has been pointed out that, for regions characterized from a really complex morphology (Goovaerts, 1999, Diodato and Ceccarelli, 2005) as Sicily, it is really important to take into account the elevation information to carry out a reliable variables estimate.

- On the whole, the best results have been achieved through the application of a combination of an EAI deterministic methods and a geostatistical method: the residual kriging of linear robust regression or, alternatively, artificial neural network. This matter demonstrates that not all the deterministic component is definitely explained by the simple VAI methods and that the obtained residuals have still a spatial correlation.
- While the residual kriging application improves the accuracy of the underlying deterministic methods, the application of the same method increases, unfortunately, the bias of the deterministic ones.

It is important to highlight that this kind of approach, totally focused on spatial interpolation methods, neglects the temporal dependence typical of precipitation and temperature time series, at each gauge sites and this could be the cause of a not perfect reproduction of rainfall and temperature time series statistics.

Finally, the study has shown that, through the spatial interpolation carried out with the best method, it is possible to perform a good infilling process both in the annual rainfall and in the monthly rainfall records in the Hydrologic Annals of Sicily.

To summarize the novelty and original contribution coming from this part of the thesis, it is possible to state that a comprehensive overview and comparison of several estimation methods effectiveness in retrieving the missing data of rainfall and temperature for the Sicilian region is performed. Scientific literature does not exhibit previous contributions giving a so complete analysis of spatial interpolation methods for filling missing data in hydrological datasets and in in the considered region.

With regard to the runoff, it is possible to highlight the following results:

- for all groups of basins in the considered zones (Zone 1 - western area of Sicily, Zone 2 - eastern of Sicily and in Zone 3 - south-eastern of Sicily), the stochastic interpolation method applied provides quite sound results, in terms of real influence of the global constraints.
- From the visualization of the maps, it is possible to observe that in Zone 1 and Zone 2 a good agreement between the observed and estimated runoff patterns. Furthermore, almost all the partitions strictly respect the hierarchical structure of the catchment. A different situation is encountered for Zone 3, where the presence of negative runoff estimated values demonstrates that the methods doen not succeed in reproducing correctly the runoff pattern and most of the partitions do not strictly respect the hierarchical structure of the catchment. This could be due to a low quality of the input data.

Finally, the application of this methods gives the annual runoff estimated data for the stations that have been out of work in the chosen time window and that are characterised by a dataset affected by missing data. Moreover, since the hierarchical principle allows the calculation of gridded maps for finer and finer resolution annual runoff estimated values can be obtained also for the areas of the basin not provided with gauge stations.

To summarize the novelty and original contribution coming from this part of the thesis, it is important to highlight that the previous applications of the presented approach are done in homogeneous climatic contexts with propitious conditions of the flow regime to apply the procedure. On the contrary, here, for the first time, the method is applied in the Sicilian context where both the climatic and morphological profiles are strongly inhomogeneous.

References

References:

Andrews, D. F., 1974. A robust method for multiple linear regression. Technometrics 16 (4), 523–531.

Battiti, R., 1992. 1st-order and 2nd-order methods for learning between steepest descent and newton method. Neural Computation 4, 141–166.

Bishop, C., 1995. Neural networks for pattern recognition. Oxford Press, Oxford, UK.

Bono, E., 2004. Tecniche di interpolazione spaziale finalizzate alla ricostruzione delle serie storiche di dati climatici. bsc thesis. Università di Palermo, Italy.

Bono, E., G. La Loggia, G., Noto, L., 2005. Spatial interpolation methods based on the use of elevation data. Geophysical Research Abstracts, 1607–7962/gra/EGU05–A–08893.

Bruce, J., Clark, R., 1969. Introduction to hydrogeometeorology. Pergamon Press, New York.

Brunsdon, C., McClatchey, J., Unwin, D., 2001. Spatial variations in the average rainfall652 altitude relationship in great britain: an approach using geographically weighted regression. International Journal of Climatology 21, 455–466.

Coulibaly, P., Evora, N., 2007. Comparison of neural network methods for infilling missing daily weather records. Journal of Hydrology 34 (12), 27–41.

de Smith M.J.; Goodchild M.F.; Longley P.A., 2006. Geospatial Analysis - a comprehensive guide. 3rd edition. Troubador Publishing Ltd, Leicester, UK.

Demyanov, V., Kanevsky, M., Chernov, S., Savelieva, E., Timonin, V., 1998. Neural network residual kriging application for climatic data. Journal of Geographic Information and Decision Analysis 2, 215–232.

Dennis, J., Schnabel, R., 1983. Numerical methods for unconstrained optimization and non662 linear equations. Prentice-Hall, NJ, USA.

Diodato, 663 N., Ceccarelli, M., 2005. 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.

Eischeid, J., Baker, C., Karl, T., 1995. The quality-control of long-term climatological data using objective data-analysis. Journal of Applied Meteorology 34, 2787–2795.

Eischeid, J., Pasteris, P., Diaz, H., Plantico, M., Lott, N., 2000. Creating a serially com669 plete, national daily time series of temperature and precipitation for the western united states. Journal of Applied Meteorology 39, 1580–1591.

Goovaerts, P., 1999. Using elevation to aid the geostatistical mapping of rainfall erosivity. Catena, 227–242.

Goovaerts, P., 2000. Geostatistical approaches for incorporating elevation into the spatial in674 terpolation of rainfall. Journal of Hydrology 228, 113–129.

Govindaraju, R. S., Rao, A. R. (Eds.), 2000. Artificial neural networks in hydrology. Kluwer Academic Publishers.

Griffith, D., 1987. Spatial autocorrelation: A Primer. Association of American Geographers, Washington, DC.

Hassibi, B., Stork, D., 1993. Second order derivatives for network pruning: optimal brain surgeon. Vol. 5. Morgan Kaufmann, pp. 164–171.

Isaaks, E., Srivastava, R., 1990. An introduction to applied geostatistics. Oxford University Press, UK.

Jeffrey, S. J., Carter, J. O., Moodie, K. B., Beswick, A. R., 2001. Using spatial interpolation to construct a comprehensive archive of australian climate data. Environmental Modelling and Software 16, 309–330.

Kitadinis, P., 1997. Introduction to Geostatistics: Applications in Hydrogeology. Cambridge University Press, UK.

Larsen, J., L.K., H., 1994. Generalization performance of regularized neural network models. In: Proceedings of the IEEE workshop on neural networks for signal processing IV. pp. 42–51.

Larson, L., Peck, E., 1974. Accuracy of precipitation measurements for hydrologic modeling. Water Resources Research 10, 857–863.

Lin, G., Chen, L., 2004. A spatial interpolation method based on radial basis function networks incorporating a semivariogram model. Journal of Hydrology 288, 288–298.

Martinez-Cob, A., 1996. Multivariate geostatistical analysis of evapotransiration and precipi695 tation in mountainous terrain. Journal of Hydrology 174, 19–35.

Matheron, G., 1965. Les variables régionalisées et leur estimation. Masson, Paris.

Moller, M., 1993. A scaled conjugate gradient algorithm for fast supervised learning. Neural Networks 6, 525–533.

Prudhomme, C., Reed, D., 1999. Mapping extreme rainfall in a mountainous region using geo700 statistical techniques: A case study in scotland. International Journal of Climatology 19 (12), 1337–1356.

Revfeim, K. J. A., 1990. A theoretically derived distribution for annual rainfall totals. Inter703 national Journal of Climatology 10 (6), 647–650.

Stooksbury, D., Idso, C., Hubbard, K., 1999. The effects of data gaps on the

calculated monthly mean maximum and minimum temperatures in the continental united states: A spatial and temporal study. Journal of Climate 12, 1524–1533 Part: 2.

Suhaila, J., Jemain, A., 2007. Fitting daily rainfall amount in malaysia using the normal transform distribution. Journal of Applied Sciences 14 (7), 1880–1886.

Tang, W., Kassim, A., Abubakar, S., 1996. Comparative studies of various missing data treat710 ment methods malaysian experience. Atmospheric Research 42, 247–262.

Teegavarapu, R., Chandramouli, V., 2005. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. Journal of Hy713 drology 312, 191–206.

Tobler, W., 1970. A computer movie simulating urban growth in the detroit region. Economic Geography 46, 234–240.

Vasiliev, I., 1996. Visualization of spatial dependence: an elementary view of spatial autocor717 relation. CRC Press, Boca Raton, USA.

Vieux, B., 2001. Distributed Hydrologic Modeling using GIS. Kluwer Academic Publishers, Dordrecht, The Netherlands.

ANDREASSON, J., LINDSTROM, G., GRAHN, G., JOHANSSON, B., (2004), "Runoff in Sweden – Mapping of Climate Change Impacts on Hydrology," Nordic Hydrological Programme II, 625-632;

ARNELL, N. W., (1995), "Grid Mapping and River Discharge," Journal of Hydrology, 167, 39-56;

BISHOP, G.D., CHURCH, M.R., (1992), "Automated Approaches for Regional Runoff Mapping in the Northeastern United States," Journal of Hydrology, 138, 361-383;

BISHOP, G.D., CHURCH, M.R., (1995), "Mapping Long-Term Regional Runoff in the Eastern United States Using Automated Approaches," Journal of Hydrology, 169, 1, 189-207;

BISHOP, G.D., CHURCH, M.R., ABER, J. D., NEILSON, R. P., OLLINGER, S. V., DALY, C., (1998), "A comparison of Mapped Estimates of Long-Term Runoff in the Northeast United States," Journal of Hydrology, 206, 3-4, 176-190;

BLOSCHL, G., (2005), "Rainfall-Runoff Modelling Ungauged catchments," Encyclopedia of Hydrological Sciences, Artiche 133, 2061-2080, Wiley, Chicester;

BONO, E., (2004), "Tecniche di Interpolazione Spaziale Finalizzate alla Ricostruzione delle Serie Storiche di Dati Climatici," Tesi di Laurea, Università degli Studi di Palermo

BRUNO, G., CANNAROZZO, M., CIRAOLO, G., (2003), "Le Grandi Dighe in Sicilia," Dipartimento di Ingegneria Idraulica e Applicazioni Ambientali, Università degli Studi di Palermo;

CHURCH, M. R., BISHOP, G. D., CASSELL, D. L., (1995), "Maps of Regional Evapotranspiration and Runoff/Precipitation Ratios in the Eastern United States," Journal of Hydrology, 168, 1, pp. 283-298;

FERRARESI, M., FRANCHINI, M., (1988), " Analisi regionale dei Deflussi nei Bacini dell'Ofanto e del Fortore: la Regionalizzazione dei Deflussi Medi," XXI Convegno di Idraulica, L'Aquila;

FERRARESI, M., TODINI, E., FRANCHINI, M., (1988), " Un metodo per la Regionalizzazione dei Deflussi Medi," XXI Convegno di Idraulica, L'Aquila.

FOYSTER, A. M., (1975), "Mapping Runoff by the Grid Square Technique," Nord Hydrology, 6, 207-221;

GANNETT, H., (1912), "Maps of United States Showing Mean Annual Runoff," Surface Water Supply of the United States, U.S. Geological Survey, 301-312;

GEBERT, W. A., GRACZYK, D. J., KRUG, W. R., (1987), "Average Annual Runoff in the United States 1951-80," Hydrological Investigations Atlas HA-710 U. S. Geological Survey, Reston VA;

GIUSTOLISI, O., SIMEONE, V., LAUCELLI, D., DOGLIONI, A., (2003), "Strategie Evolutive per la Modellazione Afflussi Deflussi in Bacini Strumentati," Giornata di studio:Metodi statistici e matematici per le analisi idrologiche,Roma 2003;

GOTTSCHALK, L., (1993a), "Correlation and Covariance of Runoff," Stocastic Hydrology and Hydraulics, 7, 85-101; GOTTSCHALK, L., (1993b), "Interpolation of Runoff Applying Objective Methods," Stocastic Hydrology and Hydraulics, 7, 269-281;

GOTTSCHALK, L., KRASOVSKAIA, I., (1993), "Interpolation of Annual Runoff to Grid Networks Applying Objective Methods," IAHS, 214, 81-89;

GOTTSCHALK, L., KRASOVSKAIA, I., (1998), "Development of Grid-Related Estimates of Hydrological Variables," Report of the WCP-Water Project B.3, WCP/WCA, Geneva, Switzerland;

GOTTSCHALK, L., KRASOVSKAIA, I., LEBLOIS, E., SAUQUET, E., (2006), "Mapping Mean and Variance of Runoff in River Basin," Hydrology and Earth System Sciences Discussion, 3, 299-333;

GRACZYK, D. J., GEBERT, W. A, KRUG, W. R., ALLORD, G. J., (1987), "Maps of Runoff in the Northeastern of Region and the Southern Blue Ridge Province of the Unites States durino periods in 1983-1985," U. S. Geological Survey, Madison,

WI;

GRAHN, G., GYLLANDER, A., JOHANSSON, B., SVENSSON, P., (2002), "Runoff Map of Sweden – A Method for Continuous Production," Nordic Hydrological Programme II, 491-196;

HERSCHY, R. D., (1998), "Mean Annual Runoff: Correlation with Catchment Characteristics," In Encyclopedia of Hydrology and a Water Resources, Cambridge;

HUANG, C., YANG, F. T., (1998), "Streamflow Estimations Using Kriging," Wat. Resour. Res., 34, 1599-1608;

JENSON, S. K., J. O. DOMINGUE, (1988), "Extracting Topographic Structure from Digital Elevation Data for Geographic Information System Analysis," Photogramm. Eng. Remote. Sens., 54, 1593-1600;

KORZUN, V. I., (1978), "World Water Balance and Water Resources of the Earth," UNESCO, Studies and reports in Hydrology, 25, Paris; LANGBEIN, W. B.et al., (1949), "Annual Runoff in the United States," U. S. Geological Survey,Washington, DC;

LIEBSCHER, H., (1972), "A Method for Runoff Mapping from Precipitation and Air Temperature data" World Water Balance, Vol. 1, UNESCO/IASH/WMO, Gentbrugge, Belgium, 115-121;

MELESSE, A.M., GRAHAM, W.D., JORDAN, J.D., (2003), "Spatially Distributed Watershed Mapping and Modelling: Gis-based Storm Runoff Response and Hydroraph Analysis: part 2," Journal of Spatial Hydrology, Vol. 3, No.2, Fall 2003;

NOTO, M.T., LA LOGGIA, G., NOTO, L.V., (2005), "Tecniche GIS per il Calcolo del Bilancio Idrologico delle Acque Superficiali,".....

NOTO, M. T., (2002), "Bilancio Idrologico per la Stima delle Risorse Idriche: Innovazioni Applicative tramite Tecnologie GIS," Tesi di Laurea, Università degli Studi di Palermo;

PREDEEK, A., ISELE, K., (1992), "A Study on the Transformation of Point Measured Runoff Data into Grid based Data," Document presented at the second planning meeting on "Grid estimation of runoff data,"WCP-Water Project B. 3., Warsaw, Poland, 6-9 May;

PRIESTLEY, C. H.B., TAYLOR, R. J., (1972), "On the Assessment of Surface Heat Flux and Evaporation Using Large-Scale Parameters," Monthly Weather Review, 100, No. 2, 81-92, Febrary;

ROTULO, L., (2005), "Utilizzo degli L-moments per l'Analisi Regionale delle Portate al Colmo. Applicazione alla Sicilia," Tesi di Laurea, Università degli Studi di Palermo;

SAUQUET, E., (2004), "Mapping Mean Annual and Monthly River Discharges: Geostatistical Developments for Incorporating River Network Dependencies," Hydrological Regimes and Water Balance, Ohrid, FY Replubic of Macedonia, 25-29 May 2004;

SAUQUET, E., GOTTSCHALK, L., LEBLOIS, E., (2000), "Mapping Average Annual Runoff: a Hierarchical Approach Applying a Stochastic Interpolation Scheme," Hydrological Sciences Journal, 45(6), 799-815;

SAUQUET, E., KRASOVSKAIA, I., LEBLOIS, E., (2000), "Mapping Mean Monthly Runoff Pattern using EOF Analysis," Hydrology and Earth System Sciences, 4(1), 79-93;

SERVIZIO IDROGRAFICO ITALIANO, sezione di Palermo: Annali Idrologici, parte II, Servizio Poligrafico dello Stato, 1923-1997;

SKOIEN, J. O., MERZ, R., BLOSCHL, G., (2006), "Top Kriging – Geostatistics on Stream Networks," Hydrology and Earth System Sciences, 10, 277-287;

VIGLIONE, A., CLAPS, P., LAIO, F., (2006), "Utilizzo di Criteri di Prossimità nell'Analisi Regionale del Deflusso Annuo," atti del XXX° Convegno di Idraulica e Costruzioni Idrauliche – IDRA 2006;