UNIVERSITÀ
DEGLI STUDI
DI PADOVA

Dipartimento di *AGRONOMIA ANIMALI ALIMENTI RISORSE NATURALI E AMBIENTE (DAFNAE)*……

_____

SCUOLA DI DOTTORATO DI RICERCA IN SCIENZE DELLE PRODUZIONI VEGETALI

INDIRIZZO:… Agronomia Ambientale…….

CICLO: …… XXV……

METHODS FOR GAP FILLING IN LONG TERM METEOROLOGICAL SERIES AND CORRELATION ANALYSIS OF METEOROLOGICAL NETWORKS

Direttore della Scuola : Ch.mo Prof. Angelo Ramina

Coordinatore d'indirizzo: Ch.mo Prof. Antonio Berti

Supervisore :Ch.mo Prof. Antonio Berti

Dottorando : Gianmarco Tardivo

Declaration

I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which to a substantial extent has been accepted for the award of any other degree or diploma of the university or other institute of higher learning, except where due acknowledgment has been made in the text.

A copy of the thesis will be available at http://paduaresearch.cab.unipd.it/

Dichiarazione

Con la presente affermo che questa tesi è frutto del mio lavoro e che, per quanto io ne sia a conoscenza, non contiene materiale precedentemente pubblicato o scritto da un'altra persona né materiale che è stato utilizzato per l'ottenimento di qualunque altro titolo o diploma dell'università o altro istituto di apprendimento, a eccezione del caso in cui ciò venga riconosciuto nel testo.

Una copia della tesi sarà disponibile presso http://paduaresearch.cab.unipd.it/

To the Mother of God
and Her Holy Son

Table of contents

Riassunto

I dati climatologici sono molto utili in molti campi della ricerca scientifica. Oggigiorno, molte volte questi dati sono disponibili sottoforma di enormi data-base che sono spesso prodotti da stazioni meteorologiche automatiche.

Affinché analisi di ricerca e lavori di modellistica siano possibili su questi data-base, essi devono subire un'opera di omogeneizzazione, validazione e ricostruzione dei dati mancanti. Le operazioni di validazione ed omogeneizzazione sono già per lo più condotte dalle organizzazioni che gestiscono questi dati. Il problema principale rimane quello della ricostruzione dei dati mancanti.

Questa tesi si occupa principalmente di due argomenti: (a) la ricostruzione di valori mancanti di insiemi di dati di precipitazione e temperatura giornalieri; (b) un'analisi fondamentale sulla correlazione spazio-temporale tra le stazioni di una rete meteorologica.

(a)

Per prima cosa, si presenta un nuovo modello adattivo per ricostruire i dati di temperatura. Questo modello viene confrontato con uno non adattivo. Poi si presenterà un'analisi dettagliata sulla scelta ed il numero di predittori per metodi di ricostruzione di tipo multi-regressivo.

Precipitazioni e temperatura sono le più importanti variabili climatologiche, così, viene scelto un metodo per ricostruire anche i dati giornalieri di pioggia, questa scelta viene fatta attraverso un confronto fra 4 tecniche.

(b)

Questi due metodi (ricostruzione di pioggia e temperature) permettono di ricostruire i data-base che vengono usati per il prossimo ed ultimo lavoro: l'analisi di correlazione, attraverso le coordinate spaziale e temporale della rete.

Summary

Climate data are very useful in many fields of the scientific research. Nowadays, in many cases these data are available through giant data-base that are often yielded by automatic meteorological networks.

In order to make possible research analysis and the running of computational models, these data base need to be validated, homogenized, and to be without missing values. Validation and homogenization are common operations, nowadays: the organizations that manage these data-base provide these services. The main problem remain the reconstruction of the missing data.

This dissertation deal with two main topics: (a) the reconstruction of missing values of daily precipitation and temperature datasets; (b) a base analysis on the time and space correlation between stations of a meteorological network.

(a)

At first, a new adaptive method to reconstruct temperature data is described. This method is compare with a non-adaptive one. A detailed analysis of the effects of the number of predictors for a regression-based approach (to reconstruct daily temperature data) and their search strategy is then presented.

Precipitation and temperature are the most important climatological variables, so, a method to reconstruct daily precipitation data is chosen through a comparison of four technique.

(b)

The methods selected in phase (a) make it possible to reconstruct the two data-base (precipitation and temperature) that will be used for the next and last work: the correlation analysis, through time and space of network data.

Chapter 1

General Introduction

Automatic networks and climate data

Agronomy, engineering (civil, naval, aeronautical, etc.), geology, environmental sciences in general, architecture, etc., are among the fields of the science in which climate data play an important role.

Nowadays, in many locations on the surface of the earth, automatic networks of meteorological instruments (thermometers, pluviometers, hygrometers, barometers, etc.) are available to detect the state of the climate.

These instruments transmit their atmospheric measurements to a system of hardware/software devices, via radiowaves and at regular intervals, which in turn systematically record this data in the form of files (electronic structured data).

This measurements are collected in giant database providing a source for research and the analysis of the climatic variables being in the running.

Precisely, this radio-software organization makes it possible to gather these important database affected mainly by **systematic errors**.

This quality of the data were not possible when, formerly, data were collected with the help of operators that directly read the measurements from the instruments; these readings were highly prone to human errors, and the density of the instruments was very low. The data base collected in this way were affected by unsystematic errors that provides difficulties for both statistical and climatic comprehension. Some datasets could be easily lost (for example, due to wartime events). Even with automated systems, however, unsystematic errors can be present e.g. for failures of the system.

In these last years, a large number of research works are obtained from the study and the analysis of this “automatic” database. They led to a rapid increase of deeper analysis, especially in the field of statistics and the applied sciences, through the creation of new mathematical models. In the meteorological field, they make it possible to reach an improved ability to forecast, and an easier comprehension of phenomenologies governing the physics of the atmosphere. Obviously, all this, with the help of computers: permitting a giant number of calculations in a few bits of time.

Statistical errors and non-response

The statistical error can be subdivided into two kinds of errors: sampling errors and non-sampling errors (*Groves R.M., 1989)*.
Among non-sampling errors is non-response, that can be found regardless of the kind of data-base, both manual or automatic; though, in the case of automatic networks, these errors are easier to analyse.
Among non-response errors can be cited: missing values (not observed), outliers (observations that is numerically distant from the rest of the data), out-range (values outside of a reasonable range) and inconsistent values (comparing with other variables).
Often, nowadays, missing data is the only problem that must be solved; in fact, most part of the organizations, managing this networks, carry out detailed analysis to validate (validation: the checking of data for correctness, or the determination of compliance with applicable standards, rules, and conventions) the measurements of the instruments, almost immediately after the data were recorded.

Main topic of the research

The main topic of this research, is the **imputation of the missing values.**
Precipitation and temperature are among the main climatological variables and the two ones we will deal with; although, we retain that some methods that are here described can be applied to other weather variables (sometimes, this idea will be mentioned).
In particular we deal with daily precipitation data (accumulated) and daily temperature data: maximum, mean and minimum temperature.
These data are provided by the Veneto Meteorological Centre of Teolo (Padova, Italy); this Centre manages one of the best weather network operating in Italy.
Not more than 5 percent of the values of these database are missing. The reconstruction of these data may be due to many reasons: the needing to obtain a value of daily precipitation or temperature in a day when unfortunately this datum is not available; the use of algorithms for which, a priori, there must be no gaps in the data to run; etc.

The most important methods to reconstruct missing precipitation and temperature data, of the current literature, are described in the next chapters, through the introducting sections.

Some considerations dealing with reconstructing methods

In order to identify the most performing method to fill the gaps of a database we should understand that the type of model is strictly dependent on *:*

* The **kind of variable** (temperature, precipitation, radiation, humidity, etc.) that we are dealing with;
* The total **number of available stations** and their **density** in the studied area;
* Frequently, a **morphological dependence,** when the territory includes significant gradients in altitude, must be taken into account;
* In some cases the **different climatic zones** involved should be considered;
* When reconstructions are made per gap (calling gap a sequence of contiguous missing values), their **size** has to be considered, in fact, they could still affect the selection of the method.

It is worth noting that sometime the behaviour on space and time of different weather variables can be very different. A typical comparison is between temperature and precipitation. **Temperature** can be considered roughly **continuous.** Instead, **precipitation** generally shows **high gradients** onto both space and time directions; wide area and long periods without phenomena can be found.

Anyway, regardless of the method, it is important to yield reconstructions complying with the climatology that can be deduced from observed data; moreover, the final purposes of the reconstruction have to be considered.

Finally, it is important to cite the **adaptive methods**, they are methods that carry out reconstructions in a localized way, taking into account the particular time and place of the missing value that has to be filled.

Some brief features of the climatology of the Veneto Region

There are two principal climactic zones: the alpine region, characterised by cool summers and cold winters with frequent snowfalls, and the hill and plain areas where the climate is moderately continental. The two coastal areas, along the Adriatic and the Garda lake, give a warmer climate. The lowlands are often covered by thick fog.
From the point of view of precipitations, **convection** is one of the most important meteorological phenomena during the warm season in northern Italy (Calza et al., 2008). The convective activity in Veneto Region was documented for the warm seasons 2005, 2006 and 2007 by Calza et al. (2008), where the overall density map of the warm seasons highlights the province of Vicenza, western part of the Region, as the area with the highest frequency of convective activity.
Verifications in the Calza's paper (2008) revealed that the quantitative precipitation forecast (for the warm seasons - May-Sep - 2005, 2006 and 2007) maxima are located prevalently in the Alpine and Prealpine areas of the Veneto Region, whereas the minima are observed on the southern plains.
Months with a predominance of thermal convection which affects mainly the mountainous parts of Veneto (e.g. Jul 2006) are distinctly different from months where a synoptic influence prevails (e.g. Sep 2006). August appears to be the month with highest convective activity.

Second topic of the research

The other topic of this study, is the **analysis of the correlations** between the stations, taking into account the varying of both time and space through the database.
This topic is directly related to the first, particularly when regressive methods are used to fill the gaps; in fact, often in these cases for a regressive-formula, predictors are selected considering their correlation coefficients with the target station (target station: station where a gap has to be filled or station on which we focus for calculations), so, it is natural

to ask what is the inner structure and the relationship of the whole system of distances (from the target station) and correlations with target stations, considering all the stations of the network. This analysis will be led from both points of view, of space and time.
It is important to note that calculations carried out via computer for this topic, would not be possible if the gaps of the data were not filled in advance. Now it is clear that the two arguments are a vicious circle, but thinking that the total number of missing values to fill, not exceed the 5 percent of the whole database this hindrance is easily overcome:the possible errors induced is minimal.

Description of the sections of the dissertation

First of all, a new dynamic (time-dependent) method to reconstruct temperature data is described. This method works with a set of parameters and most of them are set from experience. Consequently, the second part of this work presents a deeper analysis carried out on the dynamic method to understand how to use these parameters with more scientific awareness. Through this second part, important deductions are made on the relationship between the multiply correlation and the distance of the stations that are used as predictors in the multilinear regression formula.
The third part deals with finding a method to fill the gaps of the precipitation-database; this method must be as simple and performing as possible.
Through these two methods the data of temperature and precipitation are completed, so (fourth part) a deep analysis on the Pearson's correlation coefficients is conducted to obtain important relationships characterizing the correlation system between the stations, through the time and the spatial coordinates of the network.
The second and the fourth parts are closely related: important differences can be noted between the working in multilinear and linear environment, on the behaviour of the stations.
Here is a more detailed description of the chapters 2,3,4,5. Each chapter corresponds to an article. The first chapter is an international published paper, the other three are submitted to international journals.

2nd Chapter (Abstract: Tardivo G and Berti A, 2012)

A regression-based approach for temperature data reconstruction has been used to fill the gaps in the series of automatic temperature records obtained from the meteorological network of the Veneto Region (North-eastern Italy). The method presented is characterised by a dynamic selection of the reconstructing stations and of the coupling period which can precede or follow the missing data. Each gap is considered as a specific case, identifying the best set of stations and the period that minimizes the estimated reconstruction error for the gap, thus permitting a potentially better adaptation to time-dependent factors affecting the relationships between stations.

The best sampling size is determined through an inference procedure, permitting a highly specific selection of the parameters used to fill each gap in the time series.

With a proper selection of the parameters, the average errors of reconstruction are close to 0 and those corresponding to the 95th percentile are typically around 0.1 °C.

In comparison with similar regression-based approaches, the errors are lower, particularly for minimum temperatures, and the method limits inversions between minimum, mean and maximum temperatures.

3rd Chapter (Abstract: Tardivo G and Berti A, submitted-1)

A suitable search strategy for identifying the best reconstructing stations is a basic requisite for the proper implementation of gap-filling methods.

A detailed analysis of the effects of the number of predictors for a regression-based approach and their search strategy is presented. These information can be used for the reconstruction of missing data in a daily temperature dataset.

Data are recorded by the weather stations of the meteorological network of the Veneto Region (North-eastern Italy). The correlation between stations was studied, checking performances with a recently published regression model. For the network considered, a better performance was achieved by the system when the maximum radius within which to start searching for predictors was set at equal to or greater than 40 km. As a consequence it can be deduced that stations used to reconstruct gaps don' t strictly need to be close to the target station. Setting the maximum number of predictors at 4 and the maximum radius at

exactly 40 km significantly reduces the number of inversions and their attached errors.

4th Chapter (Abstract: Tardivo G and Berti A, submitted-2)

Daily precipitation data are often useful to perform climatological models; nowadays these models make frequent use of computational and algorithmic approaches requiring no missing values to run.
Four straightforward methods to reconstruct gaps of precipitation data-base are here considered and compared through a series of statistical indexes and applications to some practical issues.
The purpose of this paper is to repair the daily precipitation data-base of the Veneto Region (Italy) obtaining values as consistent as possible with the information arising from the observed data.
The methods are compared from many points of view: estimating extreme errors; pairing observed rainfall values and respective errors of each method; ability to predict monthly and annual accumulations, and monthly and annual rainy days; varying the density of the network.
In the two last cases, Linear Regression seem to be the most performing; in the first case, a modified Normal Ratio method seem to have the best behaviour; in the second case, the modified Normal Ratio shares the results with Linear Regression and Inverse Distance Weighting method.

5th Chapter (Abstract: Tardivo G, submitted-3)

A basic issue concerning weather networks, when matched data bases are analysed and studied, is the correlation system that characterizes the set of weather stations.
Some statistical models that reproduce temperature and precipitation data or are used to reconstruct missing data often make use of the Pearson's correlation coefficient, whereby a selection of predictors is carried out.
In this paper a specific analysis has been made to understand the relationships between distance and correlation structures (of the network), and the changing of the order of the stations, passing the time since the birth of the network on, when they are ranked through

the values of correlation coefficient with a target station.
This study has preliminarily been carried out over the entire area (Veneto Region, Italy) and then, subdividing that area into three main climatic zones: mountain, plain and coast. The variables that are involved in this study are : daily precipitation and daily maximum, mean and minimum temperature.
The results of this work that are worth highlighting are: the correlation coefficients of the database of precipitation are, on average, inversely proportional to the mean distances from the target station; the correlation coefficients are higher for the closest stations to the target station; these last sentences are not true for temperature; from 5.5 years from the birth of the network, the temperature variable is characterized by a high stability of the correlation order, up to an wide radius from the target station; this sentence is less evident for rainfall data.

*Brief note to read figures and tables**:***

Fig.02-04 (or **Tab.02-04**) means: the fourth figure (or table) of the second chapter.

References

Groves RM. *Survey errors and surveys costs.* Wiley and Sons, New York, 1989

Calza M, DallaFontana A, Domenichini F, Monai M, Rossa AM. *A radar-based climatology of convective activity in the Veneto region.* Regional Agency for Environmental Protection of Veneto, Meteorological Center of Teolo, Italy. University of Trento. Department of Civil and Environmental Engineering, March 2008.

Tardivo G, Berti A (2012). A dynamic method for gap filling in daily temperature datasets. *J. Appl. Meteor. Climatol.*, **51:** 1079–1086. doi: 10.1175/JAMC-D-11-0117.1.

Tardivo G., and A. Berti (Submitted-1). The selection of predictors in a regression-based method for gap filling in daily temperature datasets. International Journal of Climatology.

Tardivo G., and A. Berti (Submitted-2). Comparison of four methods to fill the gaps in daily precipitation data collected by a dense weather network. Journal of Applied Meteorology and Climatology.

Tardivo G. (Submitted-3). Spatial and time correlation of thermometers and pluviometers in a weather network data-base. Theoretical and Applied Climatology.

Chapter 2

A dynamic method for gap filling in daily temperature datasets

Introduction

Both climatic data analysis and environmental modelling require a complete set of meteorological inputs, but frequently weather stations are subject to some malfunctioning in the monitoring period. The estimation of missing meteorological data is typically done through within-station, between-station or regression-based methods (Allen and De Gaetano 2001). With the first, data for the same station on days when data are available are used, while the other two groups of methods are based on data from neighbouring stations. Short gaps (typically of one or a few days) can be easily filled by simple within-station methods, such as interpolation between available data or moving averages (Kemp et al. 1983), or using data from several days before and several days after the date of missing data in a non-linear regression to fill the data gap (Acock and Pachepsky 2000). When the length of the missing period increases, between-station approaches, considering the specific variability of climate in the period to be reconstructed, tend to give better results; furthermore, Kemp et al. (1983) have shown that regression-based approaches give more accurate estimates than within-station and between-station methods.

For between-station analyses different approaches have been used, ranging from spatial averages or even the use of data from the closest station (Xia et al. 1999), to inverse distance weighting methods (Teegavarapua and Chandramouli 2005).

Also geostatistical interpolation approaches such as kriging (Jeffrey et al., 2001) are considered as reconstructing methods: the main problem with kriging is the computational intensity for large datasets, the complexity of estimating a variogram, and the critical assumptions that must be made about the statistical nature of the variation (WMO 2007; Xia 2001).

Various types of regression approaches (Eischeid et al. 1985, De Gaetano 2000, Allen and De Gaetano 2001) have been also implemented, and they can give good results (Nalder and Wein 1998), even better than those given by kriging approaches (Ian and Ross 1998) and can be easily used for automatic data reconstruction. Further increases in reconstruction performances have been achieved with thin-plate smoothing splines (ANUSPLIN) (Price et al., 2000), which gave positive results in comparison with Gradient plus Inverse-Distance-Squared (GIDS) (Nalder and Wein, 1998).

The paper presents a regression-based approach for daily temperature data reconstruction that involves a dynamic selection of both the reconstructing stations and the length of the coupling period used. Each gap is considered as a specific case, identifying the best set of stations and the period that minimizes the estimated reconstruction error for the gap. The main advantage of this approach is a potentially better adaptation to time-dependent factors affecting the relationships between stations. Examples of these factors can be climatic changes due to both long- and medium- term evolutions that have variable effects on different areas within the region considered, specific local effects such as instrument changes or modification of the conditions surrounding the stations and presence of strong seasonal effects (i.e. mountain areas).

The method has been applied to fill the gaps in the Regional network of meteorological stations of the Veneto Region (North-eastern Italy), composed of 114 stations over an area of 18364 km2.

Materials and methods

The data used in this study have been collected by the Meteorological Centre of the Veneto Region Environmental Protection Agency (Centro Meteorologico di Teolo – ARPA Veneto) on a network of 114 automatic meteorological stations. The stations are distributed quite homogeneously across the Veneto Region (North-eastern Italy) (**Fig.02-01**).

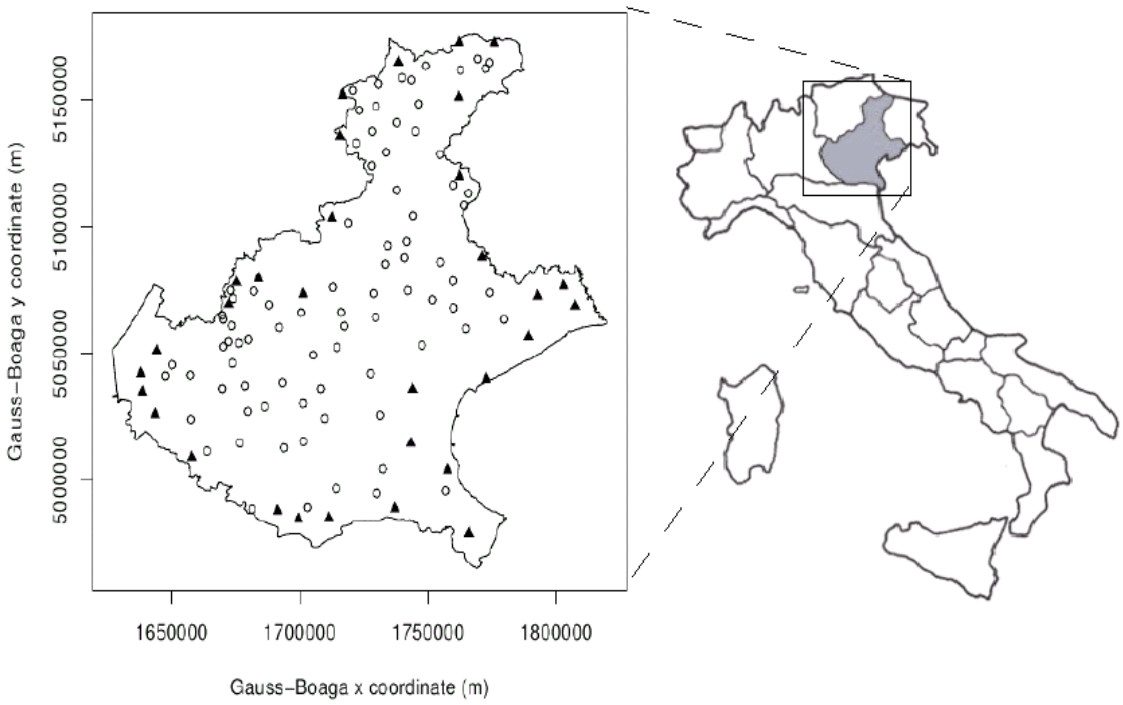

Fig.02-01_Distribution of meteorological stations across Veneto Region. ○ = stations used for reconstructions; ▲ = stations used both for reconstructions and for evaluating the presence of edge effects in the reconstructions

The stations close to the Region boundaries have been used both for the reconstruction of missing data and, in a specific analysis, to evaluate the possible presence of edge effects in the reconstructions. The main characteristics of the stations are reported in Tridello et al. (2009). The time series considered have already been checked for consistency and validated (Tridello et al., 2009).

The time span considered is from January 1, 1993 to December 31, 2008, for a total of 5844 days. Taking the whole set of 114 stations and the three parameters observed (daily

minimum, maximum and mean temperature) gives a total of 1998648 data values. There are 4440 missing data intervals (1480 missing intervals for minimum, maximum, and mean temperature respectively) in the time series, corresponding to a total of 18063 missing data values.

Most missing data intervals are short: 86% are less than 5 days and 98% are less than 20 days, with a single maximum of 213 days of contiguous missing data (**Fig.02-02**).

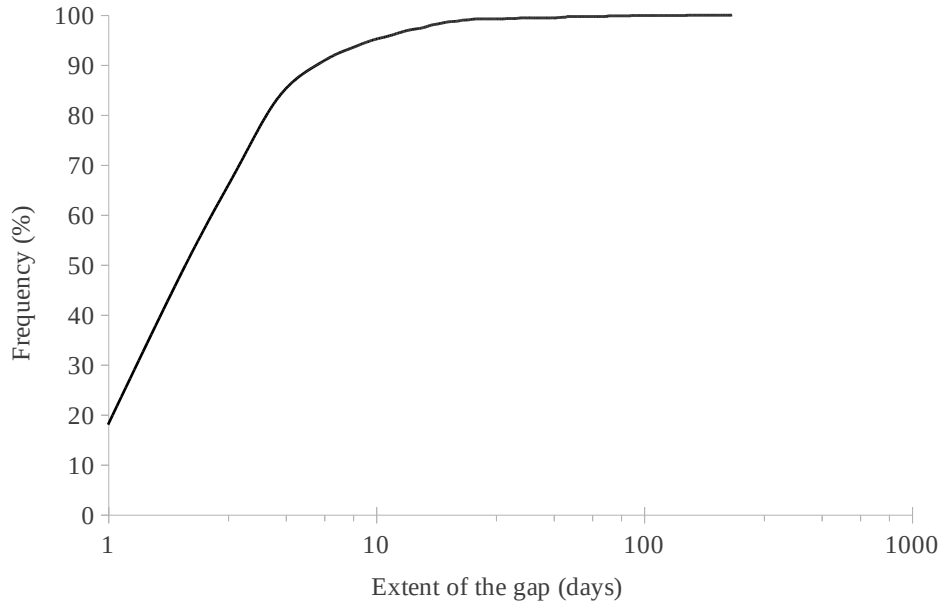

Fig.02-02_Cumulative distribution of the extent of missing data intervals (log scale). For the 114 stations. The time span considered is from January 1, 1993 to December 31, 2008. The total number of intervals of missing days is 1480

The approach used:

The method is based on multiple linear regressions (LS), using a set of surrounding stations as regressors. The approach used for the selection of the stations and identification of the best period of coupling of reconstructing and target stations can be summarised as follows:

1) analysis of the target station to identify a period without gaps of sufficient length contiguous to the gap to be filled preceding and/or following the gap;

2) identification of two groups of stations (maximum 4 stations per group) that can be

potentially used for data reconstruction in the neighbourhood of the target station, one considering data preceding the gap to be filled, the other group following the gap;
3) selection of the period to be considered (before or after the gap);
4) identification of the subset of stations, in the period previously identified, giving the best correlation with the target station for the specific gap to be filled. The search is done considering all the possible subsets within the selected group;
5) identification of the best sampling size (length of the period used for data coupling) that minimises the reconstruction error;
6) reconstruction of the gap.

The procedure is then repeated for the subsequent gaps and, after reaching the end of the time series of the considered target station, for the gaps previously omitted for lack of a period without gaps of sufficient length, using the reconstructed values previously calculated.

Details of the method:

As a first step, the algorithm verifies that in the target station there is a continuous period of at least D days before and/or after and contiguous to the gap (**Fig.02-03**), in order to perform all the required calculations.

The length of the required period without gaps is defined as follows:

D = (U-1) + T + I

where U = number of CV trials (cross-validation trials), user defined;

T = length of the gap (days)

I = maximum sampling size allowed.

The method will then search an optimal sampling size considering sampling sizes (i) ranging from number of reconstructing stations+2 to I.

To evaluate the performances of the proposed approach, the reconstruction process was tested considering a range of values of I from 25 to 500 days and of U from 100 to 500.

The stations that can be used for the reconstruction are those without gaps in the days corresponding to the gap of the target station and in the contiguous period of D days

before and/or after the gap.

The stations within a radius of 40 km satisfying this condition are selected. If there is not at least 1 station with available data within 40 km of the target station, the search radius is increased by 10 km steps until the minimum of 1 suitable station is reached. If more than 4 stations are found, the 4 with the best correlations over the period of D days are selected. The evaluation is done through the coefficient of determination (R^2). The maximum number of stations was limited to 4 following Eischeid et al. (1995), who stated that inclusion of more than 4 stations does not significantly improve the interpolation and may in fact degrade the estimate.

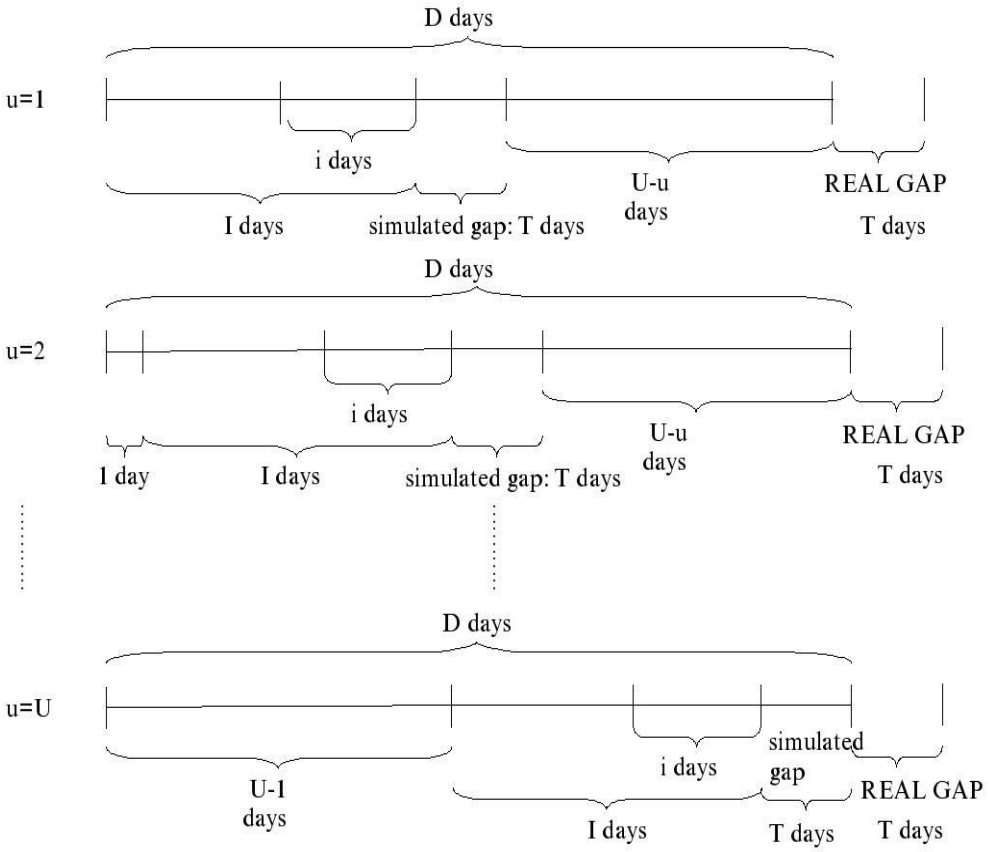

Fig.02-03_ Progress of CV trials for every sampling size in the case of the selection of the period before the real gap. $D = (U\text{-}1) + T + I$; I = maximum sampling size allowed; i = tested sampling size ; U = number of CV trials ; T = length of the gap (days). Each u represents a CV trial.

If the target station has a continuous period of D days both before and after the gap, two groups of potential stations could be identified, one for the period preceding the gap in the target station, the other following the gap. The group containing the station with the best R^2 is then selected.

The final set of stations that will be used for reconstruction is then identified considering all the possible subsets of stations in the previously identified group, and computing a multiple linear regression over the D days considered between the reconstructing stations (x_i) and the target station (y). The final set of reconstructing stations is then selected choosing the one with the lowest mean absolute error (MAE) of the multiple linear regression, exploring all the possible combinations of reconstructing stations.

The next step of the reconstruction process is to identify the best sampling size for the final reconstruction (i.e. number of days contiguous to the gap to be used for determining the multiple regression between reconstructing stations and the target station). As stated before, the length of this period (i) can vary between number of reconstructing stations+2 to a maximum of I days. For each sampling size i, a number of CV trials equal to U is done, starting from the farthest position from the gap and moving the period of T+i days towards the real gap with a 1 day step within the interval of D days, as indicated in **Fig.02-03**.

For each CV trial, a gap of T days is simulated and then reconstructed with the multiple regression obtained for the set of considered reconstructing stations in the period of i days contiguous to the simulated gap; these simulated reconstructions (CV trials) can be used for assessing the performances of the reconstruction method. For each simulated gap, a MAE is computed and, over the U CV trials, a mean MAE (MAE_i) can be computed for each i. The final sampling size (n) that will be used for the reconstruction is equal to the i value giving the minimum MAEi.

The period of n days contiguous to the real gap is then used for calculating the coefficients of the multiple regression between the selected reconstructing stations and the target. These coefficients are specific to the reconstruction of the considered gap.

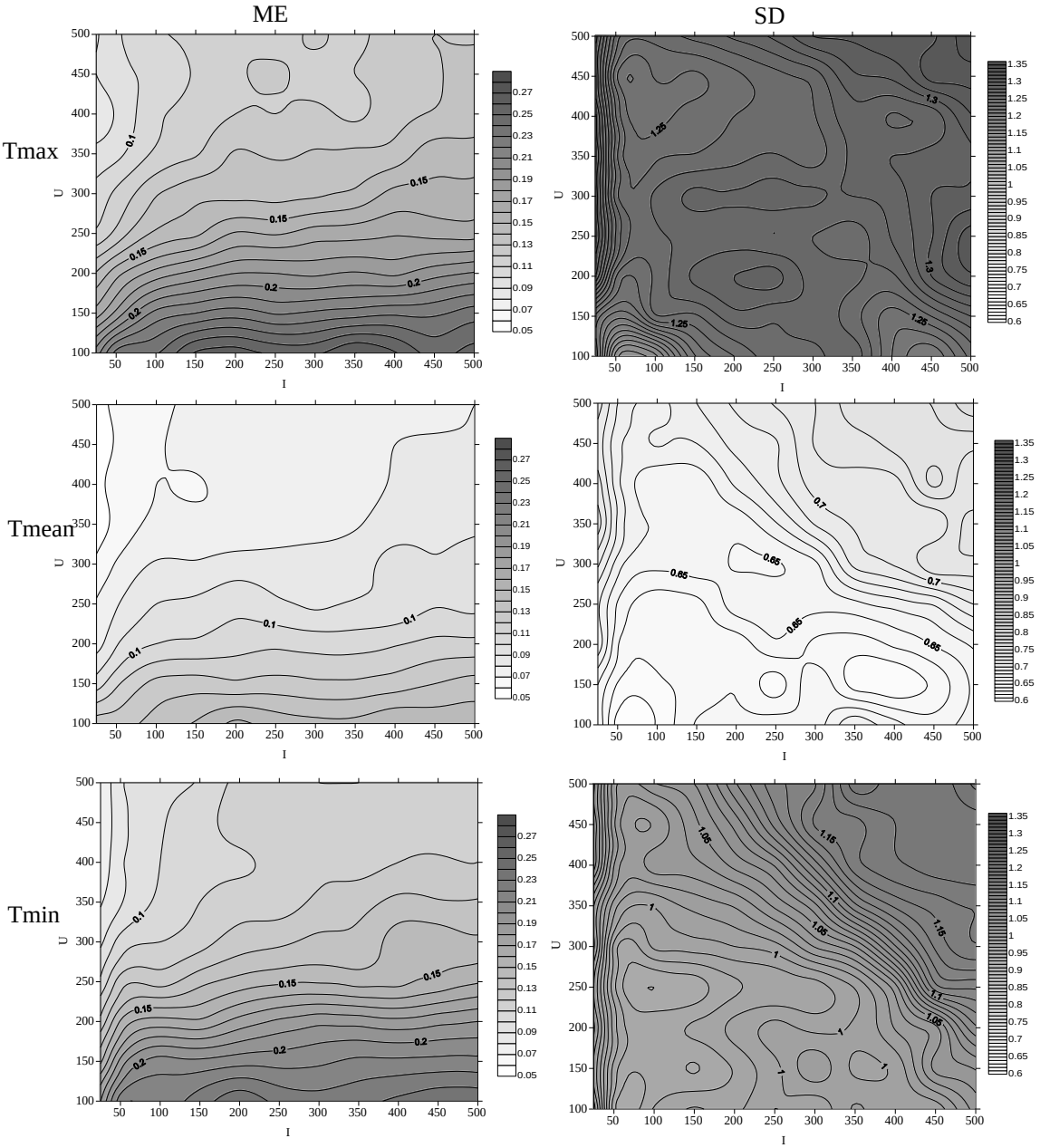

Fig.02-04_95th percentile of the absolute values of mean error (ME) (left) and of the standard deviation (SD) (right) for every *U* and *I* value; *U* = number of CV trials and *I* = maximum sampling size allowed. Tmax = maximum temperature; Tmean = mean temperature; Tmin = minimum temperature.

Results and discussion

As stated before, the method considers a maximum of 4 reconstructing stations for each gap, following the suggestion of Eischeid et al. (1995). A preliminary study was done comparing the reconstruction done with a maximum of 4 or 10 reconstructing stations (data not shown). The average reconstruction error is similar in the two cases, but the number of inversions, i.e. cases where a reconstructed minimum temperature (Tmin) is higher than the mean temperature (Tmean), or maximum temperature (Tmax) is lower than Tmean, was markedly reduced when setting the maximum number of reconstructing stations at 4. This limit was therefore retained for subsequent analyses.

A first step in the evaluation of the method performance is to assess the optimal values of the number of CV trials (U) and maximum sampling size (I) and to evaluate its sensitivity to the change of these parameters

The U values considered ranged from 100 to 500 with 50 units steps, while those considered for I were 25, 50, 100, 150, 200, 250, 300, 350, 400, 450, 500 days, for a total of 99 experiments (i.e. sets of data reconstruction with given I and U values).

The model used requires a contiguous period in the target station of at least D days without gaps for the first data reconstruction. The length of the longest period without gaps is then an important factor limiting the values of both I and U. In our dataset, this period varies within the 114 stations from 464 to a maximum of 5439 days. In order to permit data reconstruction on the whole set of stations, D should therefore not exceed 464. In our case we also considered values of I and U leading to D values above this threshold to evaluate their effect on a wider range of values, accepting the risk of not being able to fill all the gaps of the dataset.

It is worth noting that with our dataset, only in 7 cases over 114 meteorological stations, the size of the maximum contiguous period without gaps did not permit the reconstruction of the missing data with the higher I and U values considered. Even in these cases, however, a proper selection of these two parameters allowed all the stations to be reconstructed.

For each I and U, a number of CV trials equal to U is done for determine the optimal sampling size n for a specific gap; each of them allows the accuracy of the method to be

evaluated. A ME have then be computed for each data value and the average ME have been used to evaluate the reconstruction performances. Considering the whole set of data, the absolute value of the mean of the ME of the simulated reconstructions varied from 0.000 and 0.009 °C, this shows the method is essentially unbiased.

The 95th percentile of the absolute values of ME (**Fig.02-04**) confirms the great reliability of the method. However, both I and U can affect the accuracy of the method, with values generally inversely proportional to U , provided I is larger than around 100. Taking into account the 95th percentile of SD, there is a wider dispersion for the smaller sampling sizes and a tendency to reduce the variability of estimates for I values ranging from 50 to 150 days (**Fig.02-04**).

The final sampling size selected by the method is in most cases close to I when this parameter is less than 100 days. For higher I values, the final sampling size is distributed over the whole range of possible values, even if the median values don't exceed 150 days (**Fig.02-05**).

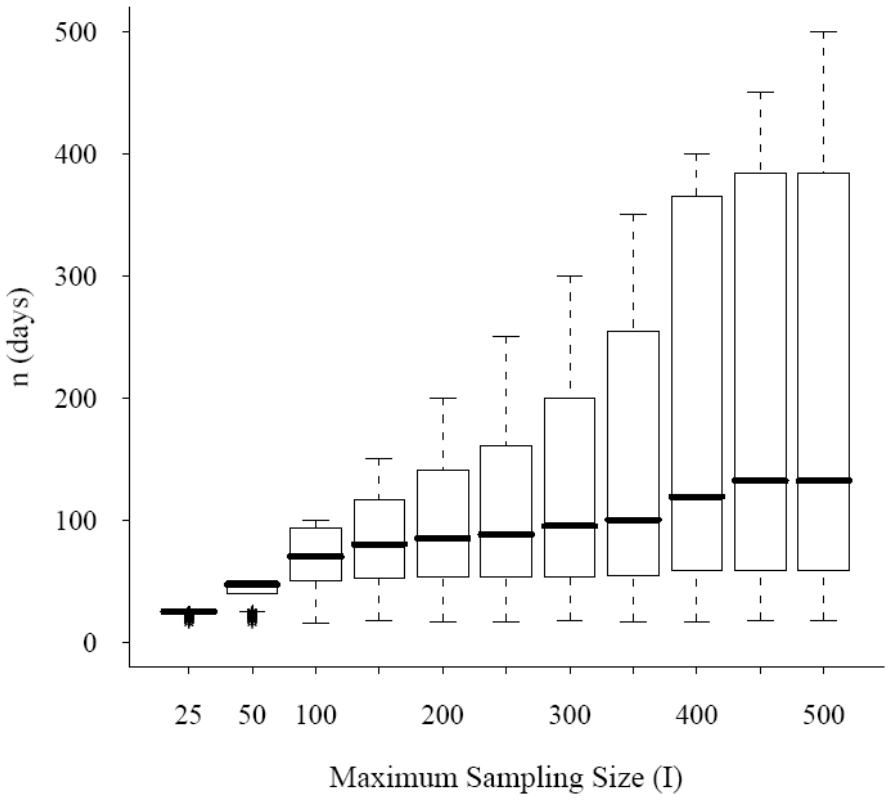

Fig.02-05_Box-plots for the distribution of the final sampling size (*n*) with different *I* thresholds. *U*=450. *U* = number of CV trials and *I* = maximum sampling size allowed.

The number of regressors used for the reconstruction is limited to 4 stations; however the procedure can select fewer stations, depending on the number of reconstructing stations found in the neighbourhood of the target station and their correlation. In most cases the final number of regressors is equal to the maximum number allowed (**Fig.02-06**), mainly because the network of stations is relatively dense, so there are a large number of stations within 40 km of the target station.

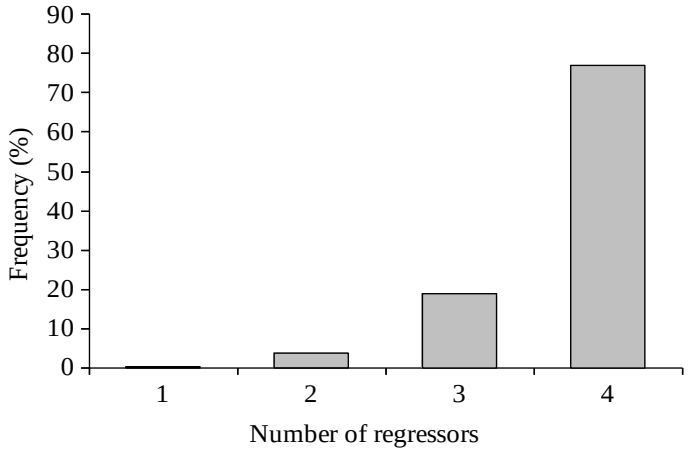

Fig.02-06_Number of regressors (reconstructing stations) selected. Maximum, minimum or mean temperature share the same graph. *U*=450, *I*=150. *U* = number of CV trials and *I* = maximum sampling size allowed.

The average distance of the stations selected as regressors slightly increases with both I and U even if the mean distance is within a 20 km range for all the combinations of I and U (data not shown). With small maximum sampling sizes (small I) the required length of the period without gaps before and after the gap to be filled is short, so a large number of possible stations are available around the target one and the search can be concluded within a relatively short radius. Increasing both I and U, the required length of the period without gaps increases, thus reducing the number of suitable stations and forcing the search over a wider radius.

To evaluate if there is some edge effect on the reconstruction of data in stations laying on the boundary of the region, the 31 stations highlighted in (**Fig.02-01**) were considered.

The 95th percentile of the reconstruction errors and of their standard deviations were very close to those of the whole set of reconstructions and the differences never exceeded ± 0.006 °C for 95th percentile and ± 0.08 °C for its standard deviation, thus indicating that the method can give reliable results even in these conditions.

A further aspect to be considered when evaluating the reconstruction performance is the number of inversions (i.e. cases where a reconstructed Tmin is higher than Tmean, or Tmax is lower than Tmean) (Xia, 2001). This type of error is quite rare, ranging from 1 to 37 inversions in the whole set of ca. 6000 missing days (**Tab.02-01**). An effect of both I and U is anyway evident, with the lesser number of inversions for I ranging from 100 to 250 days and U more than 150.

Tab.02-01_Number of inversions observed as a function of I and U parameters.

U	100	150	200	250	300	350	400	450	500
I	number of inversion								
25	33	31	30	27	18	18	29	31	37
50	26	12	18	14	11	17	18	20	19
100	6	17	11	9	9	9	9	10	10
150	11	4	5	5	3	7	7	1	4
200	3	4	9	7	5	9	3	5	7
250	5	4	7	6	7	3	7	8	11
300	5	4	6	9	6	6	10	11	15
350	5	2	9	6	9	10	11	13	11
400	1	8	7	7	8	10	10	11	10
450	2	5	6	10	8	11	11	11	10
500	6	7	7	9	11	9	11	11	12

Considering both the ME and its standard deviation as well as the number of inversions, the parameter values I=150 and U=450 seem to be appropriate in most cases, leading to 95th percentile errors of 0.112, 0.072 and 0.107 °C for minimum, mean and maximum temperatures respectively (**Fig.02-04**). These values were selected for subsequent analyses.

To compare the results obtained with the method presented here with other approaches, the procedure proposed by Eischeid et al. (1995) has been applied to daily data, as proposed by Allen and De Gaetano (2001).

The approach of Eischeid et al. (1995), as our method, is based on multiple regressions, using the absolute deviations from the regression (LAD) as minimising function.

The aim of this comparison was mainly to check if the use of a dynamic and gap-specific selection of both the reconstructing stations and the sampling size should improve the reconstructions in comparison with a static selection of these two parameters.

The meteorological stations used for data reconstruction were selected over an initial radius of 40 km around the target station, considering a minimum of 1 station and a maximum of 4.

The LAD method was applied for all the cross-validations (simulated gaps) already done for the LS method.

The mean ME are similar for both methods and, in absolute terms, very small (**Tab.02-02**). The method presented here permits a reduction of 95th percentile SD, particularly in extreme cases, with a consistent reduction of 95th percentile ME, and greatly reduces the number of inversions, which pass from 41 with a maximum error of 2.002 °C for the method of Eischeid et al. (1995) to 1 inversion with a maximum error of 0.025 °C for the method presented here.

It is worth noting that, in the estimation procedure, Eischeid et al. (1995) used only stations presenting a correlation coefficient higher than 0.35. In our case this criterion was not used due to the very high correlation coefficients of the stations used for data reconstruction, typically over 0.95.

Tab.02-02_Comparison of LAD and LS methods for data reconstruction. 95th percentile of the absolute values of ME, mean and median values and 95th percentile of the standard deviations are presented. Parameters for LS method: I=150, U=450; for LAD I is not considered while U is the same as the LS method. U = number of CV trials and I = maximum sampling size allowed.

	LAD method			
	95th percentile	mean	median	SD 95th percentile
		°C		
Tmax	0.347	-0.011	-0.002	1.278
Tmean	0.277	0.001	-0.003	0.917
Tmin	0.339	0.001	0.003	1.414
	LS method			
Tmax	0.112	-0.001	0.002	1.247
Tmean	0.072	-0.003	-0.003	0.670
Tmin	0.107	-0.004	-0.004	1.048

Conclusions

The proposed method proved to be a reliable procedure for reconstructing missing data in long-term daily temperature series. With a proper selection of the I and U parameters, the average errors are close to 0 and those corresponding to the 95^{th} percentile are always very low, typically around 0.1 °C. This aspect is of great importance because these errors are often related to extreme observations, thus indicating a potential ability of the method also in reconstructing extreme events. The selection of the number of cross-validations (U) and maximum sampling size allowed (I) is crucial for the reliability of the method, both for limiting the reconstruction error and the number of inversions. On the other hand, augmenting U and I, the number of calculations and the time required for the reconstruction increases. The values of I=150 and U=450 appears to be a good compromise, allowing a good reconstruction of all the missing data while maintaining a feasible computing time. Further studies should be necessary to assess if these parameters can be considered valid also in other situations or are highly dependent on the specific characteristics of the network considered.

In comparison with similar regression-based approaches (Eischeid et al. 1995), the errors are lower, and the method proposed here limits to a minimum the case of inversions between minimum, mean and maximum temperatures.

References

Acock, M. C., and YA. A. Pachepsky, 2000. Estimating Missing Weather Data for Agricultural Simulations Using Group Method of Data handling. J. Appl. Meteorol., 39: 1176-1184.

Allen, R. J., and A. T. De Gaetano, 2001: Estimating missing daily temperature extremes using an optimized regression approach. Int. J. Climatol., 21: 1305-1319. doi:10.1002/joc.679. SIAM J. Num. Anal., 10, 839-848.

DeGaetano, A. T., 2000. A serially complete observation time metadata file for US. Daily Historical Climatology Network Stations. Bulletin of the American Meteorological Society 81: 49–67.

Eischeid, J. K., C. B. Baker, T. R. Karl and H. F. Diaz, 1995: The quality control of long-term climatological data using objective data analysis. J. Appl. Meteorol., 34: 2787-2795.

Ian, A. N., and W. W. Ross, 1998: Spatial interpolation of climatic Normals: test of a new method in the Canadian boreal forest. Agricultural and Forest Meteorology, 92: 211-225.

Jeffrey, S.J., J.O. Carter, K. B. Moodie, A. R. Beswick, 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environmental Modelling & Software 16:309–330.

Kemp, W.P.D., D. G. Burnell, D. O. Everson, A. J. Thomson, 1983. Estimating missing daily maximum and minimum temperatures. Journal of Climate and Applied Meteorology, 22: 1587–1593.

Kotsiantis, S., A. Kostoulas, S. Lycoudis, A. Argiriou and K. Menagias, 2006: Filling missing values in weather data banks. 2nd IEE International Conference on Intelligent Environments, 5-6 July, 2006, Athens, Greece. 1: 327-334.

Nalder, I. A., and R. W. Wein, 1998. Spatial interpolation of climatic Normals: test of a new method in the Canadian boreal forest. Agric. For. Meteorol. 92: 211–225.

Price, D. T., D. W. McKennedy, I. A. Nalder, M. F. Hutchinson and J. L. Kesteven: A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data. Agricultural and Forest Meteorology. 101:81-94.

Teegavarapu, R.S.V., V. Chandramouli, 2005. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. Journal of Hydrology, 312: 191-206.

Tridello, G., A. Chiaudani, F. Rech, G. Tardivo, P. Meneghin, F. Checchetto, I. Delillo, S. Orlandini, V. DiStefano, G. Bartolini, M. Mariani, G. Cola, M. Borin, A. Berti and A. Bonamano: Italian Journal of Agrometeorology (2) 2009. Quaderno degli Abstract. 12° Convegno Nazionale di Agrometeorologia. Sassari, 15-17 giugno 2009.

WMO, 1983. Guide to climatological practices. 2nd edition.

WMO, 2007. Guide to climatological practices. 3rd edition (draft).

Xia, Y., P. Fabian, M. Winterhalter and M. Zhao, 2001: Forest climatology: estimation and use of daily climatological data for Bavaria, Germany. Agric. For. Meteorol. 106: 87–103.

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

Chapter 3

The selection of predictors in a regression-based method for gap filling in daily temperature datasets

Introduction

Long-term time series often contain gaps due to failures of the measuring instruments or radio-software systems acquiring data from them. This issue is particularly acute in meteorological and climatological fields where monitoring station networks are frequently used to detect key variables such as temperature, precipitation, pressure, humidity, radiation, etc. Studies began some decades ago on models, to provide as accurate as possible climatological data reconstructions. The arrival of computers made it possible to significantly increase their performance; they enable models to making use of more sophisticated mathematical and statistical methods supported by important algorithmic structures.

Many climatological models require a number of surrounding stations to reconstruct missing values of a given station; for example: between-station methods (Kemp et al., 1983), kriging approaches (Jeffrey et al., 2001), thin-plate smoothing splines (Price et al., 2000), artificial neural networks (Kim and Pachepschy, 2010) etc. The number and closeness to the target station of these surrounding stations (predictors) is strictly dependent on the type of model and on the total number of available stations and their density in the study area; there is also frequently a morphological dependence when the territory includes significant altitudinal gradients. Depending on the characteristics of studied variables, the different climatic zones involved should be taken into account.

It must be stressed that a sufficient number of stations are required in the network; were it not for this, a more thorough numerical analysis would make no sense.

In this context,there are essentially three main ways to rank and select predictors: by climatic zone, by distance and by correlation indices.

Some methods, such as Steurer' s (1985), rely on selection of stations using broad and somewhat arbitrary climate or political boundaries; DeGaetano et al. (1995) have shown that using a distance criterion versus climate boundaries significantly reduces the overall range of errors. But in other methods such as the Normal Ratio (Young 1992), the choice was made by the value of the correlation coefficient (using the best three stations).

Temperature is one of the least problematic variables and reconstructing methods are often based on multilinear regression. There are many variants in the literature, more or less

complicated, each one tailored to a specific situation.

These methods include one by Eischeid (1995), in which the problem of number of predictors was solved by requiring a preliminary choice of 10 stations closest to the target station; this author also stated that a minimum of one station is needed as predictor and a maximum of 4, while using more than 4 stations could degrade the estimate.

Another adaptive and regression-based method was recently published (Tardivo and Berti, 2012), consisting of a statistical-computational approach that tackles each gap separately.

The method was compared with that of Eischeid; setting the maximum number of reconstructing stations at 4 or 10, it was observed that the best behaviour was achieved with 4 stations, especially from the point of view of inversions (cases when reconstructed data present Tmax values less than Tmean or Tmean less than Tmin), roughly agreeing with Eischeid' s assertion.

This chapter describes a more detailed analysis for better understanding the variation in performance resulting from the changing of these numbers, and to explore the more appropriate maximum distance from the target station within which to start searching for predictors. In some cases this problem was solved by using stations as close as possible to the target station (Eischeid, 1995).

In Chapter 2, the initial radius was set at 40 km; this was obtained through a rudimentary system of tests and no specific analysis was conducted.

It should be mentioned that this work uses regression dynamic methods: for each gap and target station a preliminary selection of predictors is needed that adapts locally to the period of the gap. The analysis of the behaviour of such a system in relation to the search parameters is of interest even for non-dynamic but regressive systems, which anyway look for predictors that have no missing days corresponding to the gap period.

Finally, it should be pointed out that spatial predictions are dealt with here, but a similar analysis was conducted in the weather forecasting field (Carr, 1988), where important suggestions were made regarding the number of predictors, referring in particular to the Model Output Statistics (MOS) forecasting technique.

Materials and methods

This work was carried out over the stations already used in chapter 2.

Selection of predictors

Reconstruction of missing data requires a coupling period to study the correlation between the target and the reconstructing station(s). In the present paper a continuous period contiguous to the gap of 600 days (plus the extent of the gap minus one) is required (see Chapter 2).

Reconstructing stations can have an available period either prior to or after the gap period. If a target gap has sets of stations available on both sides, a choice is made ranking the two sets by R^2 and picking the set having the station with the best R^2. From this last set the subgroup of stations reporting the best MAE (Mean Absolute Error) with the target station (on the coupling period of 600 days) is selected.

In this work this selection procedure is performed searching for a number of stations ranging from a minimum (mn) to a maximum (Mx), varying mn and Mx. If more than Mx stations are found, the Mx stations with the best R2 are selected. The system begins to search for perdictors within a radius of Sr km (varying Sr). When mn stations are not found, this value (Sr) is increased by 10 km until the mn number of stations is found.

This method of searching for predictors was described in chapter 2. In this chapter these three parameters are varied, ranging mn and Mx from 1 station to 12, with mn less than or equal to Mx and Sr from 10 to 60 km with a 10 km step.

Setting Sr, mn and Mx for each target station and each of their gaps, a subgroup of predictors can be found; a set of subgroups (hereafter referred as "subgroups-set") is obtained (matched with the previously set Sr, mn, Mx) scanning the whole network of target stations with their gaps. These subgroups-sets were analyzed varying Sr, mn and Mx parameters as required by the method described and according to the purposes of this particular work.

This chapter deals with the effect of the size of predictor-subgroups and their distance from

the target station. More specifically, for each subgroups-set, the mean values of sizes of all its subgroups and their mean distance (calculated over all the average distances of subgroups, of the subgroups-set, from their respective target station) were considered.

Evaluation of reconstruction performances

After choosing the predictors of a gap, the second phase consists of an inference procedure: the identification of the best sampling size (length of period used for data coupling when the real reconstruction has to be done) that minimises the reconstruction errors of a series of cross validation-trials carried out over the period of 600 days close to the real gap. This method of obtaining performance values is used to evaluate the different set mn, Mx, and Sr values over all of the gaps and, separately, for each temperature (Tmax, Tmean and Tmin).

For these tests and this network of stations the best performing values of I and U already identified in the preceding chapter were maintained, setting I=150 and U=450 days (for the sum of 600 days); the same was done for the value of maximum searching distance for predictors, which was set at 100 km.

Results and discussion

Station Selection

Stations that can be used as predictors in gap reconstruction must have a sufficient number of days available. If a dynamic selection method is used (Chapter 2), a further condition is that the required number of days must be continuous and contiguous to the missing period in the target station. This period permits statistical (and therefore climatological) relationships with the target station to be analyzed in a localized way.
In Chapter 2 this period (D) was the sum of the number of days: number of cross validation-trials (U), maximum sampling size allowed (I) and gap-size (T) minus one. In the present chapter a period of 600 days has been considered. This length of period is assumed as sufficiently wide to allow a detailed study on the relationships between weather stations.
Generally a dynamic and/or regression-based model deals with station-selection by r (Pearson correlation coefficient), R^2 (coefficient of determination) and MAE (Mean Absolut Error) as statistical indexes. R^2 is used to rank stations when the number of searched station is equal to one, while when more than one station is searched, the possible subgroups of stations are selected via combinatorial calculus, ranking them via MAE.
The selection of station appeared to be strongly affected by Mx, mn and Sr parameters, with effects involving both the number of stations selected and their distance from the target one.
A preliminary analysis can be discussed observing variations of both means of the sizes of subgroups and their distances from target station, when mn=1. Setting mn=1 permits the system to be independent from mn values and to study just the relationship of Mx and Sr.
The mean of the number of predictors is reported in **Fig.03-01** for each subgroups-set with mn=1 (this graph refers to Tmax data, the same behaviour was found for Tmean and Tmin) and **Tab.03-01** representing mean distance of the closest stations, from all stations, having an available period of D days contiguous to the gaps.
Observing **Tab.03-01** it can be noted that, if Sr≤30 km, the system can find 7 suitable

stations at most on average; instead, if Sr>30 km, the system can find more than 12 suitable stations on average.

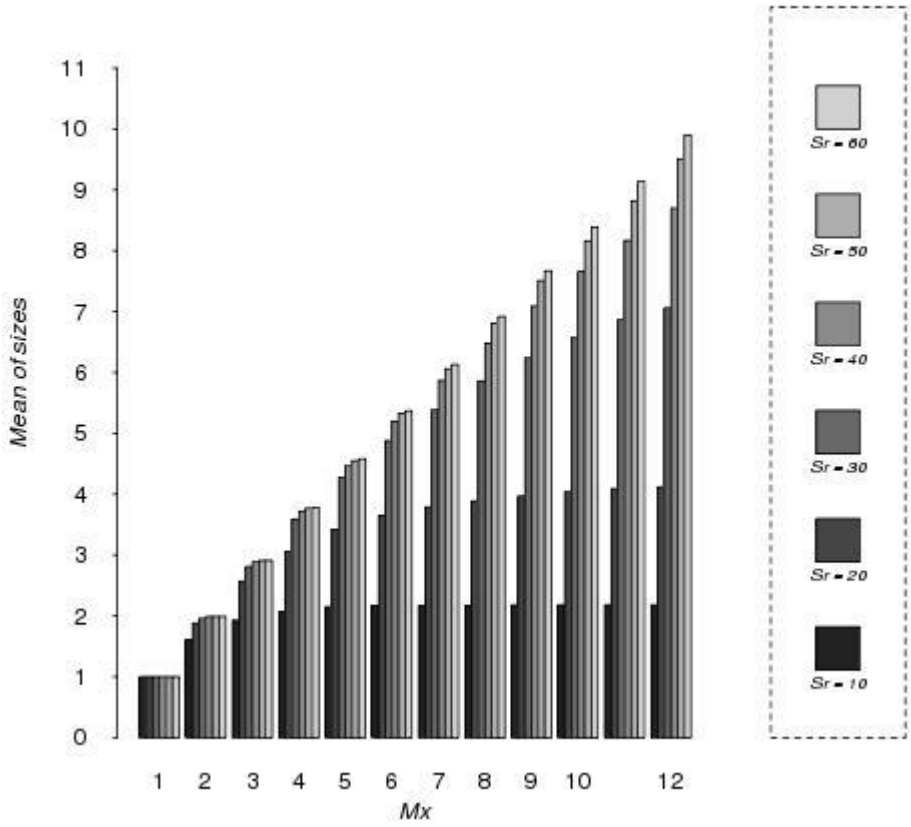

Fig.03-01_Mode-values of histograms of the whole set of sizes of selected subgroups for each collection, with mn=1 (this graph refers to Tmax data, the same behaviour was found for Tmean and Tmin).

Analysing **Fig.03-01** the system increases the mean of the sizes (of subgroups) with Mx, when Sr>30 km; instead, when Sr≤30 km, this mean remains constant when the corresponding number of stations (see **Tab.03-01**: station-order column) is approximately reached; indeed, looking at the graph with Sr=30 km, it can be seen that at Mx=9, the subgroups-sets reach a mean size of 7 stations, and this is maintained up to Mx=12; this is in accordance with **Tab.03-01**, where, for Sr=30 km the system can find 7 stations (on average).

When Sr is greater than 30 km, **Tab.03-01** shows that the system finds more than 12

stations, whereas **Fig.03-01** shows an increase of the subgroups-set size with Mx.
This suggests that the distance between the target station and the reconstructing one is not the main criterion of selection and that it is possible to find stations (and subgroups) with a high correlation even at very long distances from the target.

Tab.03-01_ Mean distances of the stations with suitable and available D-period. (From the 1st to the 12th nearest station).

station order	suitable station
1	10183.2
2	13784.5
3	16955.2
4	20124.4
5	23157.5
6	25868.5
7	28282.5
8	30373.6
9	32449.6
10	34512.0
11	36619.9
12	38499.3

Furthermore, it was observed that the average distance of the selected stations seems to be independent from Mx and varies according to mn at the smaller Sr values, while the opposite is evident for the larger Sr values. This last dependence was acquired gradually, changing Sr from 10 to 60 km.
The behaviour of the system in terms of distances is summarized in **Fig.03-02**, where the

graphs of average distances are presented varying Sr from 10 to 60 km and setting as constant first the mn and then the Mx parameter. The left hand column contains graphs with

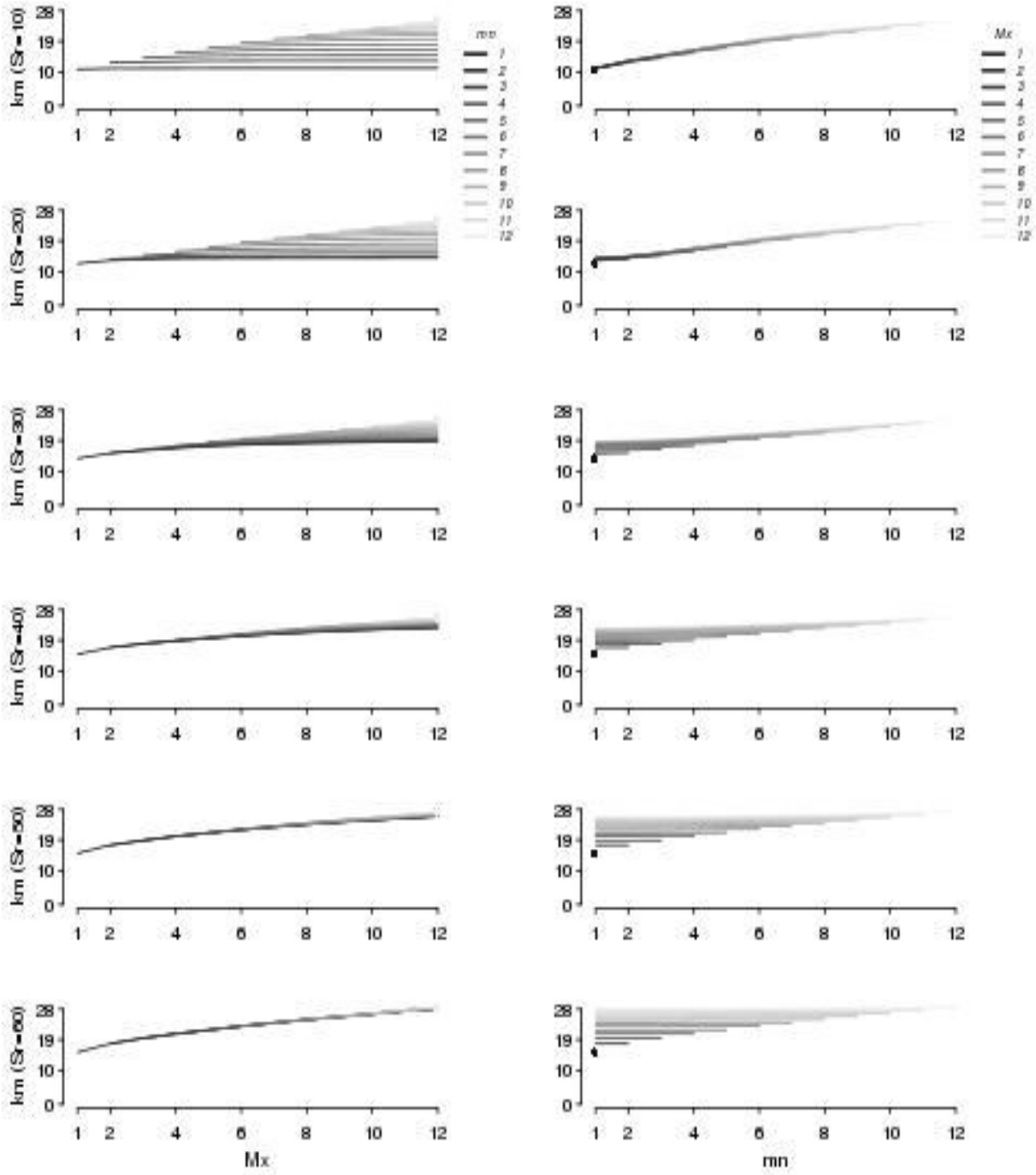

Fig.03-02_Average distances of subgroups varying Sr from 10 to 60 km and setting as constant first the mn parameter (left-side) and then the Mx parameter (right-side).

mn=constant: from Sr=10 to 60 km, it can be seen that there is a tendency to increase the dependence on Mx of the average distance of subgroups selected; at Sr=60 km this distance increases with Mx and this increase becomes independent of the mn value. Instead, looking at the right hand column: from Sr=10 to 60 km, the increasing of independence from mn when Mx is constant can be noted; at Sr=10 km dependence is quite equal for each Mx constant value.
The mean value of distances reported in each graph increase with Sr; referring to Tmax as an example: mean value is 16802 meters for the graph of Sr=10 km, rising to 17523, 19663, 21832, 23709, up to 24961 for Sr=60 km.
When the search is restricted to a relatively reduced radius, the optimum number of predictors is mainly dictated by the parameter mn; when a number of stations equal to the minimum allowed is available, the search stops. With a wider search radius it is possible to identify many possible predictors, generally above the maximum allowed Mx; in these conditions the number of predictors selected is mainly dictated by Mx and, frequently, stations as far away as 84 km but, nevertheless, highly correlated with the target one can be found.

Estimation of performance

To evaluate the performance of this selection procedure the cross validation-trial method of dynamic model (Chapter 2) has been used, where calculations of 95%, SD95% and inversion errors were done for each subgroups-set.
For each Sr value the values of 95th percentile of the mean error (varying mn and Mx parameters) are presented in **Tab.03-02**, together with their standard errors (SD95); the absolute values of these errors are very low and are only marginally affected by Sr.
The SD95 values tend to reach a minimum when mn=Mx in comparison with the cases when mn was less than Mx, so, looking at these values (Mx=mn) (**Fig.03-03**) the standard deviation of Tmax decreases rapidly with both Sr and the number of predictors, becoming roughly stable for Sr≥40 km and for more than 6 predictors. With a proper selection of both Sr and number of predictors it is anyway possible to obtain SD95 values very close to the best values: when Mx=mn=4 stations and Sr≥40 km, the values are only 0.015 °C higher

Tab.03-02_ Differences between maximum and minimum values of distances, 95% and SD95%, for each Sr and each temperature (Tmax, Tmean, Tmin), varying mn and Mx.

Sr	Distance			95			SD95		
	Tmax	Tmean	Tmin	Tmax	Tmean	Tmin	Tmax	Tmean	Tmin
10	14242	14852	14925	0.029	0.062	0.041	0.659	1.110	1.145
20	12674	13171	13219	0.017	0.023	0.010	0.279	0.433	0.542
30	11381	11302	11292	0.023	0.017	0.007	0.246	0.261	0.383
40	10810	10975	10966	0.021	0.013	0.011	0.195	0.183	0.186
50	12206	12607	12521	0.019	0.010	0.011	0.187	0.178	0.161
60	13258	13944	13999	0.019	0.012	0.008	0.184	0.169	0.155

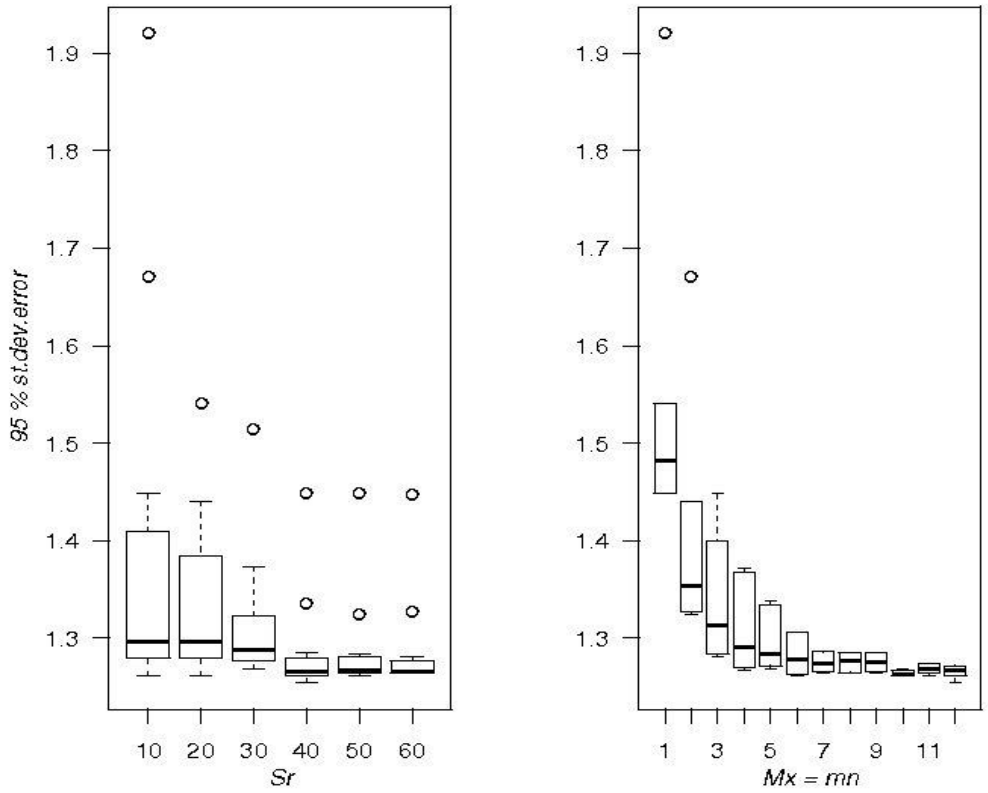

Fig.03-03_Boxplot of 95% st. dev. errors matched to collections varying Sr from 10 to 60 km (left-graph) and varying mn=Mx (right-graph); for Tmax. The same behaviour was found for Tmean and Tmin.

than the best one; the same happens with Mx=mn=10 stations, independently of Sr.
The same behaviour was found for Tmean and Tmin with minimum SD95 values of 1.254, 0.584 and 0.946 °C for Tmax, Tmean and Tmin respectively.
In the case of the number of inversions (cases when reconstructed data present Tmax values less than Tmean or Tmean values less than Tmin) (**Tab.03-03**) it can be observed that the best results were found when Mx=mn was equal to or greater than 4 and Sr≥40 km; while, considering the associated errors of inversions (**Tab.03-04**), the best entries are found when Mx=mn was equal to or greater than 3.

Tab.03-03_Number of inversions found, for each reconstruction procedure. (For each Sr value and each Mx=mn).

Sr	Mx=mn											
	1	2	3	4	5	6	7	8	9	10	11	12
10	91	48	26	21	13	9	11	16	10	10	12	10
20	61	45	24	21	12	9	11	16	10	10	12	10
30	35	45	26	16	9	10	9	14	10	11	12	10
40	25	23	10	2	4	6	4	12	12	9	6	10
50	28	26	14	7	5	9	5	11	7	9	7	9
60	29	28	11	7	6	6	6	11	9	14	4	9

Tab.03-04_The maximum absolute error of inversions found (°C), for each reconstruction procedure. (For each Sr value and each Mx=mn).

Sr	Mx=mn											
	1	2	3	4	5	6	7	8	9	10	11	12
10	4.003	1.982	1.061	1.604	0.718	0.390	1.443	1.033	0.615	0.565	0.950	0.296
20	6.652	1.982	0.877	1.604	0.544	0.390	1.443	1.033	0.615	0.565	0.950	0.296
30	3.906	3.007	0.956	1.659	0.932	0.390	1.443	1.033	0.615	0.565	0.700	0.296
40	3.906	3.007	0.549	0.025	0.210	1.030	1.097	0.715	0.590	0.517	0.327	0.315
50	3.906	3.007	0.549	1.084	0.323	0.393	0.660	1.394	1.827	0.363	0.574	0.447
60	3.906	3.007	0.549	1.084	0.323	0.393	0.660	1.394	1.827	0.646	0.851	0.779

Conclusions

The results presented highlight that it is preferable to search for suitable reconstructing stations over a wide search radius. This may seem counter-intuitive, being the closeness of the predicting station being a widely used and accepted criterion for station selection. In effect in most cases the closer stations are highly correlated with the target one but it is anyway possible to identify other stations that, despite their distance from the target one, present high correlations due to some specific local trait.

A further advantage of a wide search radius is that it is possible to identify a large number of possible reconstructors and the increase in the number of predictors permits the reconstruction error to be limited. In our case, setting $Sr \geq 40$ km it is already possible to obtain a saturated selection system, i.e. a system that almost always finds at least Mx stations independently from mn.

Considering both SD95 and inversion errors, the best results are obtained for Sr=40 km and mn=Mx=4, reaching 1 inversion with an error of 0.025 °C. Extrapolating from the specific situation, it seems to be appropriate to use a search radius allowing the identification of a number of stations roughly three times the required number of predictors. This generally permits an optimal subgroup of predictors to be identified, limiting the reconstruction error to a minimum.

References

Jeffrey SJ, Carter JO, Moodie KB, Beswick AR. 2001. Using spatial interpolation to construct a comprehensive archive of Australian climate data. *Environmental Modelling. Software*, **16**:309–330.

Price, D. T., D. W. McKennedy, I. A. Nalder, M. F. Hutchinson, and J. L. Kesteven, 2000: A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data. Agric. For. Meteor., 101, 81–94.

Kemp W.P., D.G. Burnell, D.O. Everson, and A.J. Thomson. 1983. Estimating missing daily maximum and minimum temperatures. Journal of Climate and Applied Meteorology 22: 1587–1593.

Kim J. W. and Y. A. Pachepsky. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. Journal of Hydrology. DOI : 10.1016/j.jhydrol.2010.09.005.

Steurer P. 1985. Creation of a serially complete data base of high quality daily maximum and minimum temperatures. National Climatic Data Center, NOAA.

DeGaetano A.T., K.L. Eggleston, W.W. Knapp. 1995. A method to estimate daily maximum and minimum temperature observations. Journal of Applied Meteorology 34: 371–380.

Young K. C. : A three-way model for interpolating for monthly precipitation values. Monthly Weather Review. DOI : http://dx.doi.org/10.1175/1520-0493(1992)120

Eischeid J.K., C. B. Baker, T. R. Karl, and H. F. Diaz : The Quality Control of Long-Term Climatological Data Using Objective Data Analysis. Journal of Applied Meteorology. http://dx.doi.org/10.1175/1520-0450(1995)034

Tardivo G., and A. Berti : A dynamic method for gap filling in daily temperature datasets. Journal of Applied Meteorology and Climatology. DOI : 10.1175/JAMC

Megg Brady Carr . Determining the Optimum Number of Predictors for a Linear Prediction Eqaution. 1988. Monthly Weather Review. DOI : http://dx.doi.org/10.1175/1520-0493(1988)116

Chapter 4

Comparison of four methods to fill the gaps in daily precipitation data collected by a dense weather network

Introduction

Precipitation databases are very important in many research fields, including hydrology (e.g. evaluation of basin flows), agronomy (e.g. calculations of evapotranspiration), and climatology and meteorology (e.g. precipitation forecasting).

Nowadays, one of the most important issues concerning the use of a database to obtain relevant information about rainfall, is the reconstruction of the missing values, required by many algorithms used in data analysis.

Two important issues can be highlighted about the filling of a precipitation database: the ability of a reconstructing method to allow accurate computations on simple averages (e.g. on monthly or annual periods); and the ability to reconstruct extreme values.

In general, applying reconstructed data to obtain monthly or annual accumulated rainfall values or rainy days as consistent as possible with the information from the observed data, is a usual way to test the method.

In most cases, filling missing gaps in daily precipitation data is a difficult task. Indeed, this can be clearly seen when comparing precipitation and temperature variables: generally, precipitation is characterized by higher values of both space and time gradients. This may be due to the climatic zone involved (as may be the case for northern Italy), however this feature can be considered as an intrinsic characteristic of this variable.

Nevertheless, it must be considered that summer precipitations in northern Italy are characterized by short-range storm cells. Sometimes these cells are very localized, so that only one pluviometer in the grid can adequately record the event and this interferes with data reconstruction. For example, if this instrument failed to record the event, there would be no way to estimate this datum from surrounding stations.

Many approaches have been used for filling time series, for example: kriging and thin plate smoothing splines; in many cases, such as basin flow evaluation (hydrology), Artificial Neural Network models (ANN) are very reliable (Kim and Pachepsky, 2010), but when more accurate evaluations of extreme values are required, ANN models are less effective in reproducing the events (Tirozzi et al., 2006; p. 162); in comparison many other straightforward methods, such as Normal-Ratio (NR) (Paulhus and Kohler, 1952;

Young, 1992), Multiple Linear Regression (MLR), Multiple Discriminant Analysis (MDA) (Young, 1992), Nearest-Neighbour (NN), Inverse Distance Weighting (IDW) and Linear Regression (LR) (Serrano et al., 2009) seem to show lower values of errors in reproducing extremes, though they are generally less effective, on average, on non-extreme values.

In this work, NN, IDW, LR and a slightly modified Young' s NR method are tested and compared.

The methods are compared from many points of view: estimating extreme errors; pairing observed rainfall values and respective errors of each method; ability to predict monthly and annual accumulations, and monthly and annual rainy days; varying the density of the network.

In the two last cases, LR seems to be the best performer; in the first case, a modified NR method seems to have the best behaviour; in the second case, the modified NR gives the same results as LR and IDW methods.

Materials and methods

Data

The data span from 1st January 1993 to 31st December 2007. The network has 109 stations, distributed over 18400 km^2 of the Veneto Region. Each station has a number of missing data that does not exceed 5% of the spanned period. 62% of the data were equal to zero-precipitation.

The stations of the network are automatic, they are radio-connected to a system of software/hardware devices that record the measurements.

The instruments are tipping bucket rain gauges.

Methods

Three of the four methods compared are presented in detail in Serrano et al. (2009).

Some differences from Serrano's application (due to the network and area involved) are described below.

Nearest-neighbour method (NN) uses two criteria for the selection of the predictor: the nearest neighbour has to be within a radius of 40 km of the target station and have a correlation coefficient (with the target station) higher than 0.6 (instead of 15 km and r=0.5 requested in Serrano's paper; the reason is due to the differences in the spatial-densities and distribution of the correlations between this dataset and the Serrano one) and at this distance the correlation (Pearson's r) between the daily precipitation series from both stations is equal to 0.72 on average (instead of 0.62, see Serrano's paper). In this case, setting these thresholds (40 km and 0.6) appeared to be the best compromise. The gaps are filled directly with data from the closest station meeting the criteria.

The low percentage of missing days in the database makes it possible to not consider the problem of common data between the predictor and target stations (see Serrano's paper).

In the **linear regression method** (LR), missing data were obtained by identifying the station more correlated with the target one, forcing its regression line with the target station to pass through the origin. Only the slope coefficient was used to provide reconstructed data (see Serrano et al., 2009).

Inverse distance weighting (IDW), where $(1/d)^2$ is the weighting factor, (d) being the distance between target and neighbouring station. Serrano et al. (2009) used a maximum radius of 15 km for the interpolation, while a radius of 40 km has been considered in this paper, for consistency with the NN method.

The last compared method is a variant of the **Normal-Ratio** (NR), first proposed by Paulhus (1952). A modified version was proposed by Young (Young et al., 1992), using functions of r-Pearson coefficients as weights of neighbouring stations. In this paper, different functions of these coefficients are proposed, obtaining the formula:

$$x_T = (\sum r_i \, x_i) / (\sum r_i^{1.75})$$

Where x_i are the values of surrounding stations, r_i the respective Pearson's coefficients.
This formula is applied to a maximum of three stations with the best correlation coefficient, and within a radius of 40 km.
This variant of Young' s method was considered because of the relatively small value of the highest error presented in reconstructing the whole network (through the cross-validation system); precisely: the exponent 1.75 in the denominator gave the smallest value of the highest error, in comparison with other exponents and other types of functions of r-Pearson.
The choice of a 40 km radius was due to the different structure of the network in comparison with that studied by Serrano's paper; here, a greater distance was needed within which the target stations gather a sufficient number of well-correlated predictors.
The performance of each method was studied by doing a large number of cross-validations, reconstructing one day in each of them and comparing the reconstructed value with the

observed one.

The approaches were then compared using the Maximum Absolute Error (MAE) and Residual Mean Square Error (RMSE) and evaluating the amount and distribution of outliers.

Results and discussion

A total of 515117 cross-validations were done to evaluate and compare the four methods. For all four methods, the upper range of error centiles (from 88 to 100th centiles) are shown in **Fig.04-01**. Up to the 98th centile the four methods present roughly the same behaviour while important differences are evident for extreme errors which are well differentiated between the methods, with NN presenting the highest maximum errors and NR the lowest.

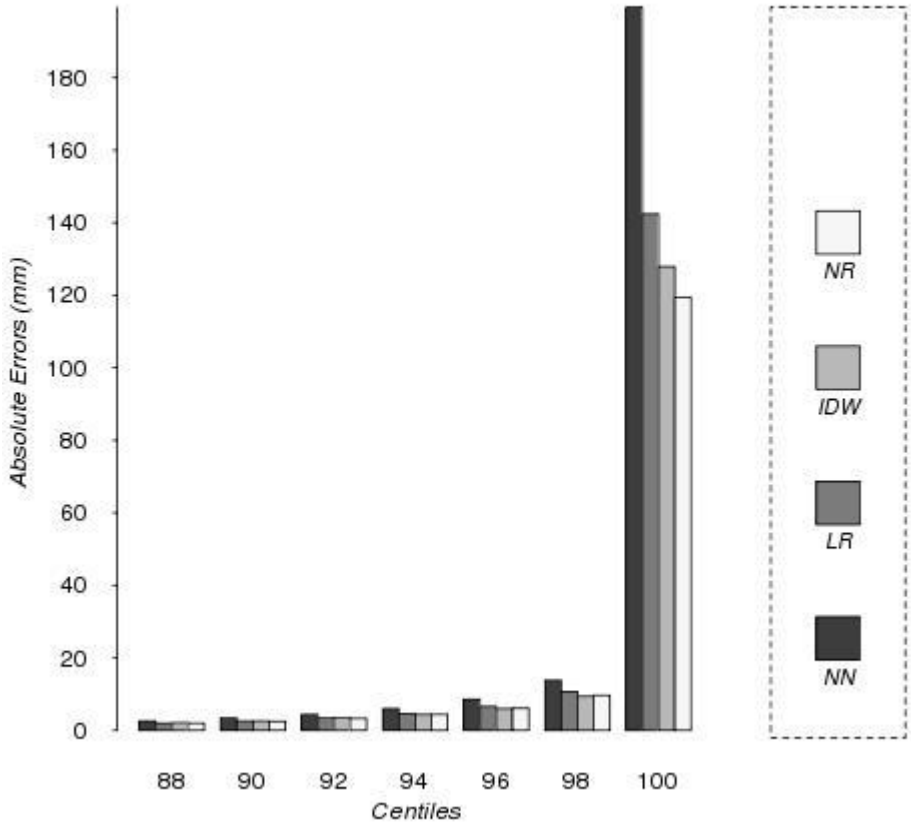

Fig.04-01_Centiles of errors (from the 88th to the 100th centile) through the whole set of cross-validations.

It is worth noting that IDW, while presenting a slightly higher MAE than NR, has the lower RMSE (**Tab.04-01**). To define the level of significance of the indices (maximum error,

MAE, RMSE), a series of 1000 bootstraps over the cross-validations were carried out (**Fig.04-02**). It can be seen that NN and LR methods present significantly higher MAE and RMSE, while NR has a significantly lower MAE than IDW method.

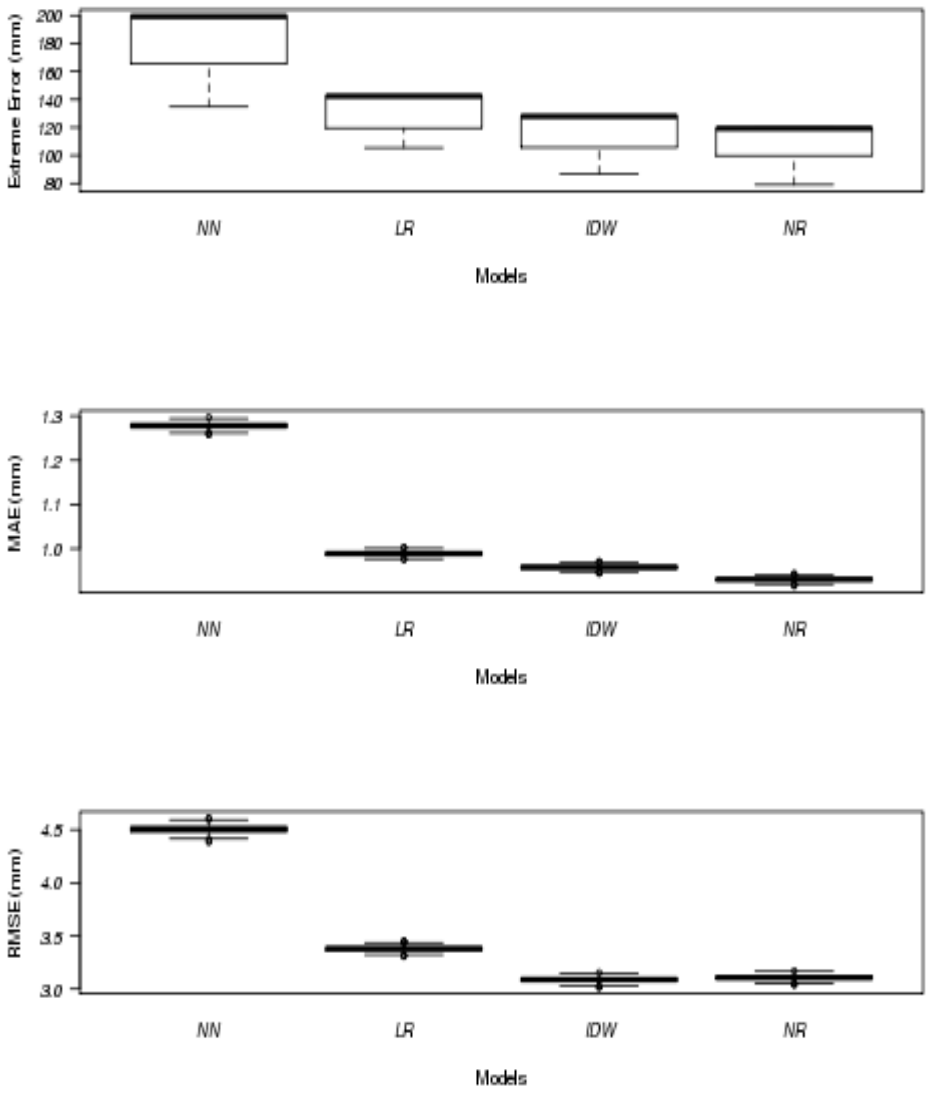

Fig.04-02_Box-plots of the values deduced from the bootstraps carried out to estimate the significance of the performance of NR method.

Tab.04-01_MAE and RMSE of the four methods.
These values are obtained for all the cross-validations.

	MAE	RMSE
NN	1.28	4.50
LR	0.99	3.38
IDW	0.96	3.09
NR	0.93	3.11

To evaluate the performances of the methods in the case of outliers, two analyses were conducted. Firstly, the relationship between the real values of daily rainfall and the associated outliers of errors of reconstruction was studied for each method.

In this paper errors that are over 1.5×IRQ (Inter Quantile Range) +75th centile or below 25th- 1.5×IRQ of the distribution of errors, are considered outliers of errors. **Tab.04-02** presents the numbers of outliers of errors, subdividing the set of observed daily precipitation values (p), of the whole network, into 7 intervals: p=0 mm; $0 < p \leq 2$ mm; $2 < p \leq 20$ mm; $20 < p \leq 40$ mm; $40 < p \leq 70$ mm; $70 < p \leq 88.4$ mm; $p > 88.4$ mm. It can be seen that when p is equal to 0, the fourth method shows a greater number of outliers of errors, but the mean value of these errors is the smallest; in the other cases the number of outliers is similar for all the methods, but the mean of the fourth is always among the best.

The threshold value of 88.4 mm was selected to evaluate the outliers of the real daily rainfall values, and calculated following the method proposed by Eischeid et al. (1995): an outlier was flagged when it was greater than f*IRQ+50th centile, f being a multiplication factor; choosing the multiple (f) of the IRQ where the slope of the function of the number of outliers flagged varying f was sufficiently near zero. Setting f=10 (instead of 4, see Eischeid), all values greater than 88.4 mm were found as outliers in the whole series (see **Fig.04-03**).

These analysis are not sufficient and are too specific in order to determine the goodness of a method. A reconstructing method of daily precipitation has to be effective when it is used to evaluate, for example, the number of rainy days or the values of annual or monthly accumulations. For that purpose calculations on these topics were made comparing the fitted values resulting from the four methods with the real values.

Tab.04-02_The numbers of outliers of errors and their mean values matched to each sub-indicated interval: p=0 mm; $0 < p \leq 2$ mm; $2 < p \leq 20$ mm; $20 < p \leq 40$ mm; $40 < p \leq 70$ mm; $70 < p \leq 88.4$ mm; $p > 88.4$ mm.

Number	Mean	Number	Mean	Number	Mean
p = 0 mm		0 < p ≤ 2 mm		2 < p ≤ 20 mm	
28889	1.16	9554	5.66	5373	18.48
30657	0.60	8894	3.88	5485	13.58
47967	0.72	9465	4.17	5074	11.65
63685	0.40	10311	3.65	5251	11.90
20 < p ≤ 40 mm		40 < p ≤ 70 mm		70 < p ≤ 88.4 mm	
457	40.55	104	64.31	18	91.11
567	28.63	152	45.67	17	65.38
608	25.48	172	42.72	33	59.58
590	26.48	174	43.35	27	61.33
p > 88.4 mm					
9	126.62				
22	85.85				
9	98.89				
17	85.01				

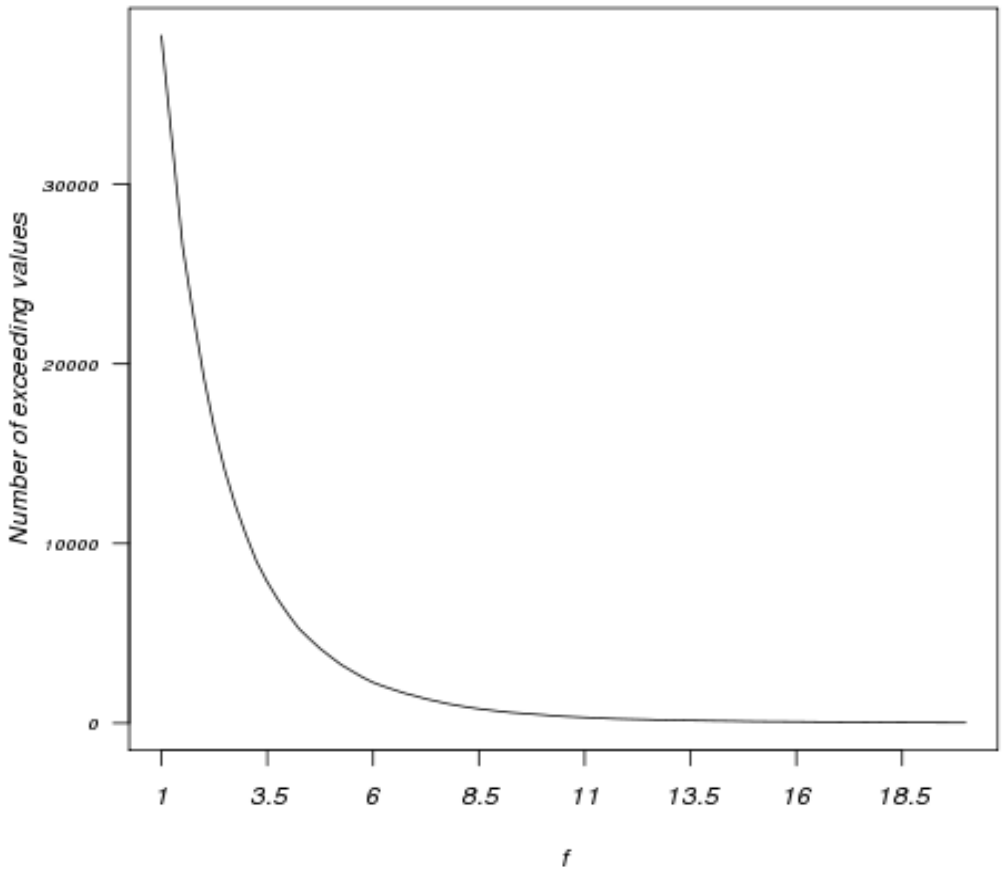

Fig.04-03_ The varying of the number of outliers flagged varying f.

Fig.04-04…04-11 can be analysed to reveal the goodness of the methods, comparing each one with the others.

For monthly accumulated values, **Fig.04-04** (scatterplots of the fitting values with the real ones) shows LR to be the more symmetric method. NN, IDW and NR have a tendency to underestimate these monthly values (especially for the high ones); **Fig.04-05** (boxplots of observed and fitting monthly accumulated values) shows a similar behaviour for all methods, but when outliers are considered, NN and LR seem to have better estimates. **Fig.04-06** and .**04-07** are similar to **Fig.04-04** and **.04-05**, respectively, but they show results from the estimates of the numbers of rainy days monthly. In this case IDW seems not to underestimate, on average, in comparison with the other

methods;

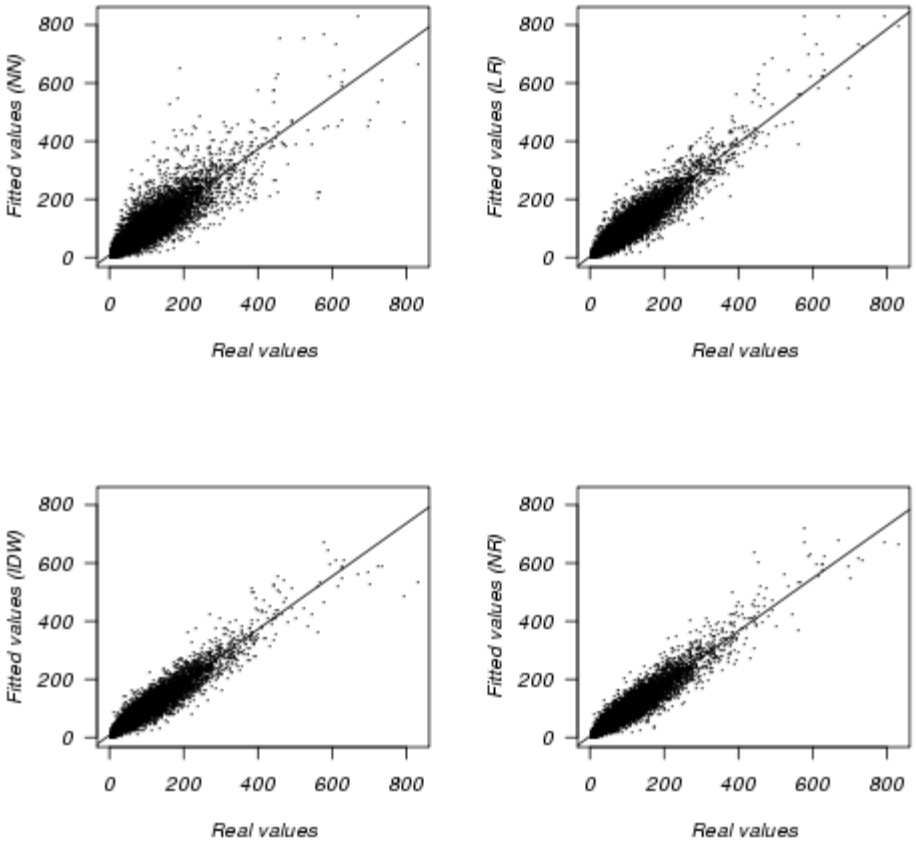

Fig.04-04_ Scatterplots of the fitting values with the real ones, for the four methods. Monthly accumulated values.

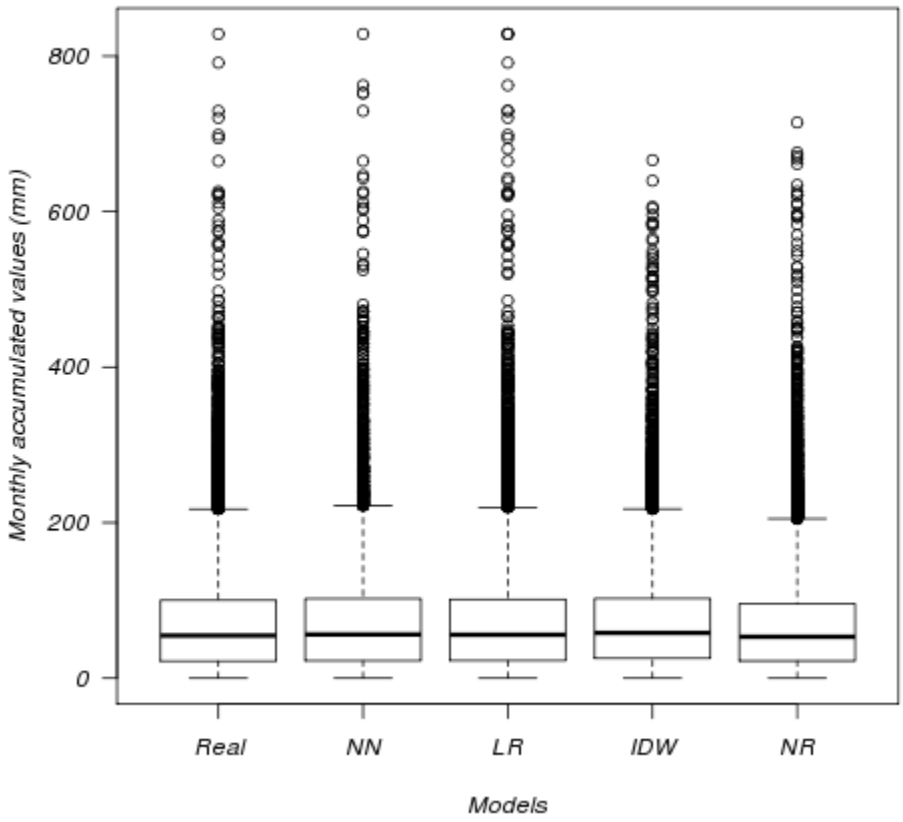

Fig.04-05_Boxplots of the fitting and the real values, for the four methods. Monthly accumulated values.

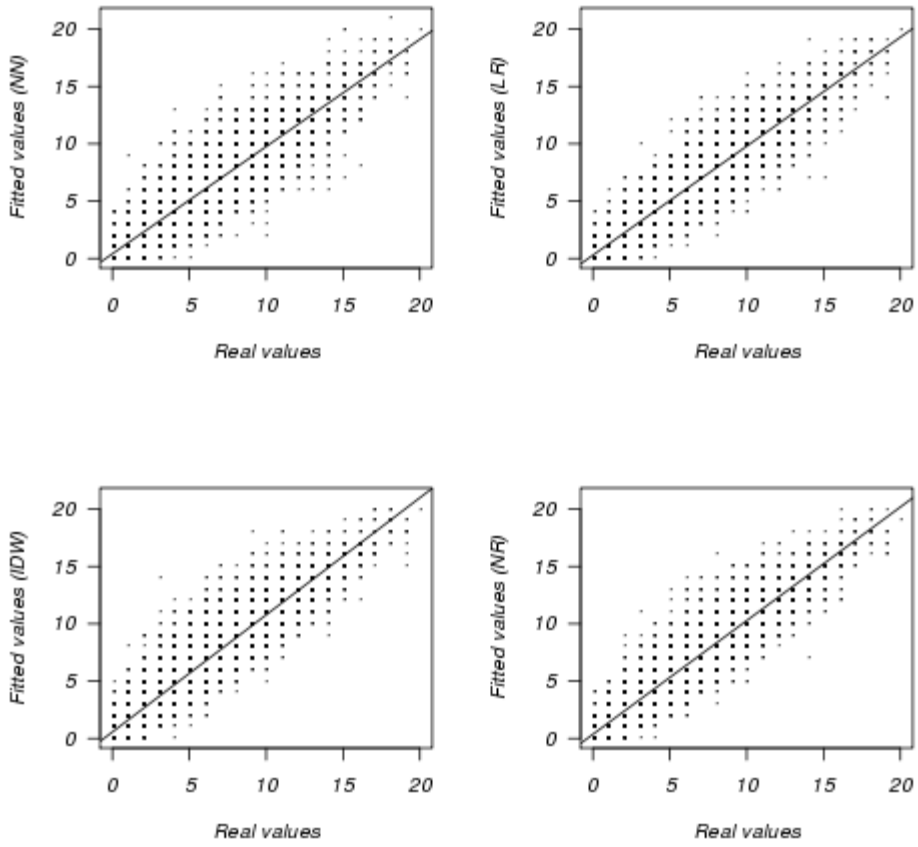

Fig.04-06_Scatterplots of the fitting values with the real ones, for the four methods. Monthly rainy days.

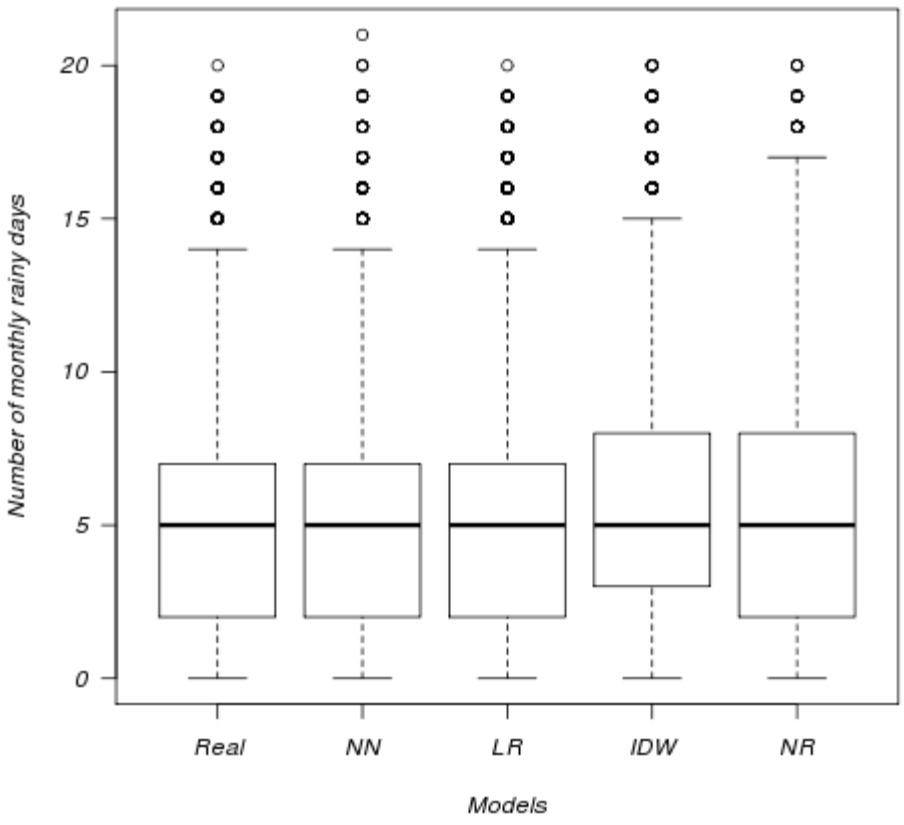

Fig.04-07_Boxplots of the fitting and the real values, for the four methods. Monthly rainy days.

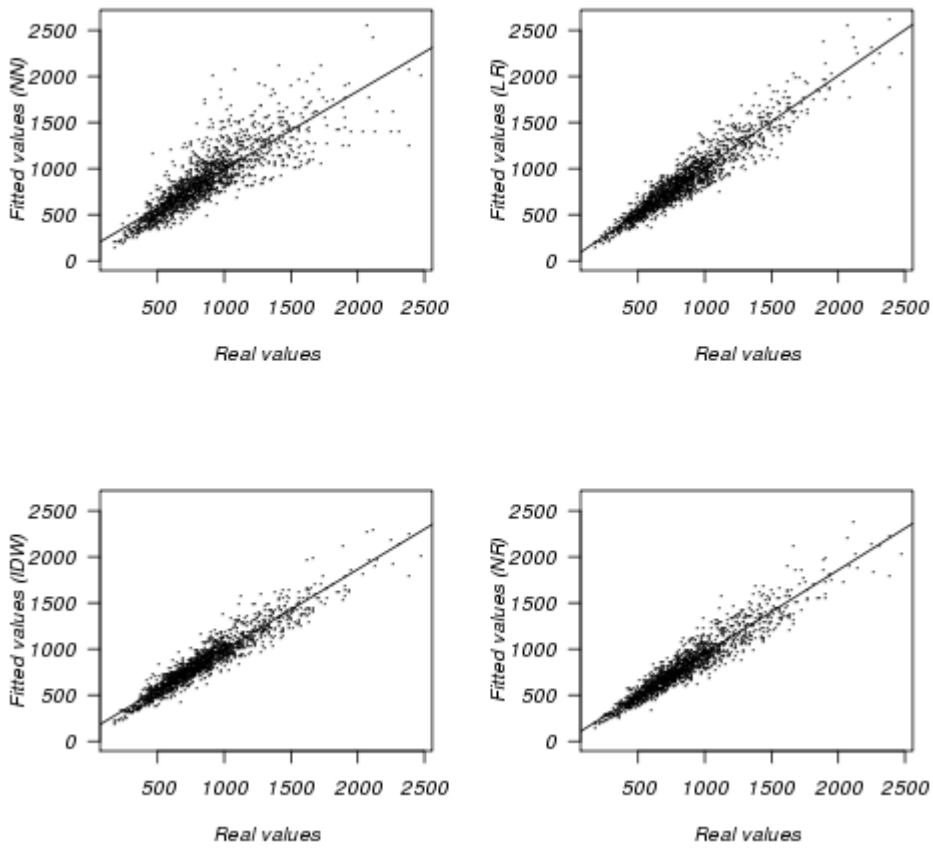

Fig.04-08_Scatterplots of the fitting values with the real ones, for the four methods. Annual accumulated values.

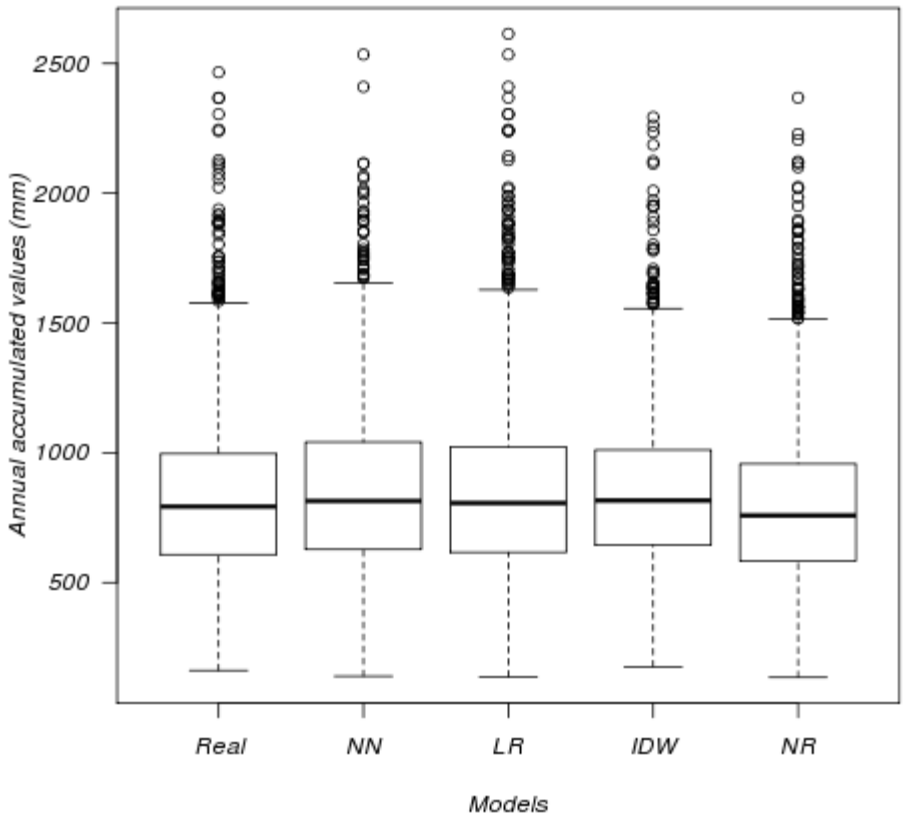

Fig.04-09_Boxplots of the fitting and the real values, for the four methods. Annual accumulated values.

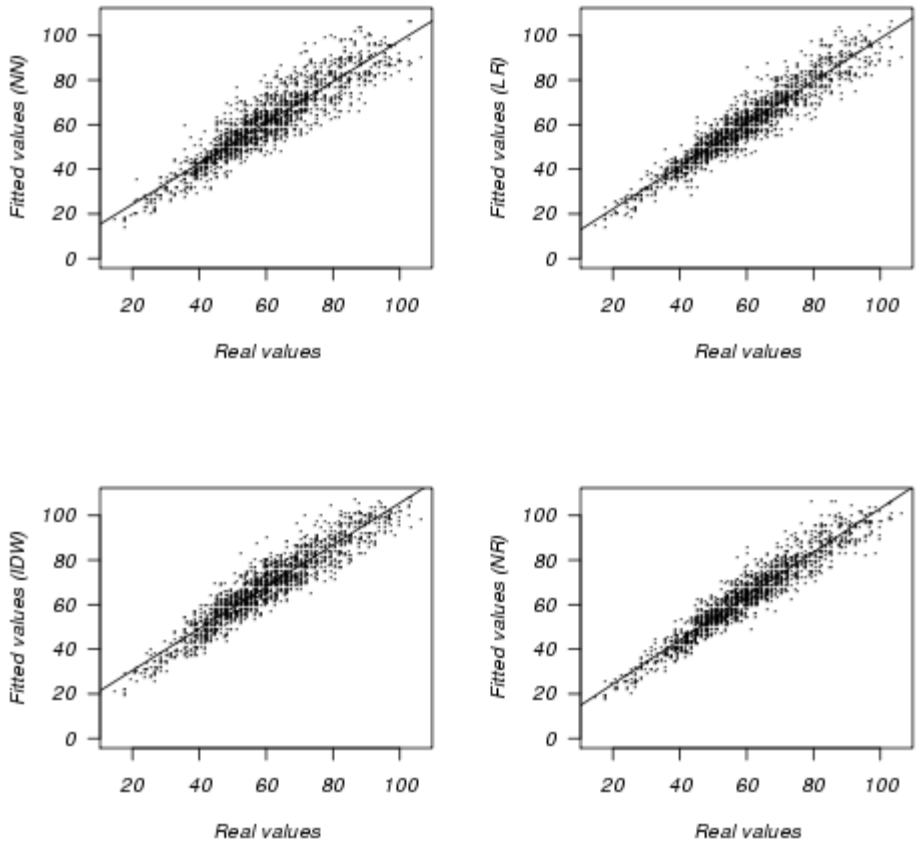

Fig.04-10_Scatterplots of the fitting values with the real ones, for the four methods. Annual rainy days.

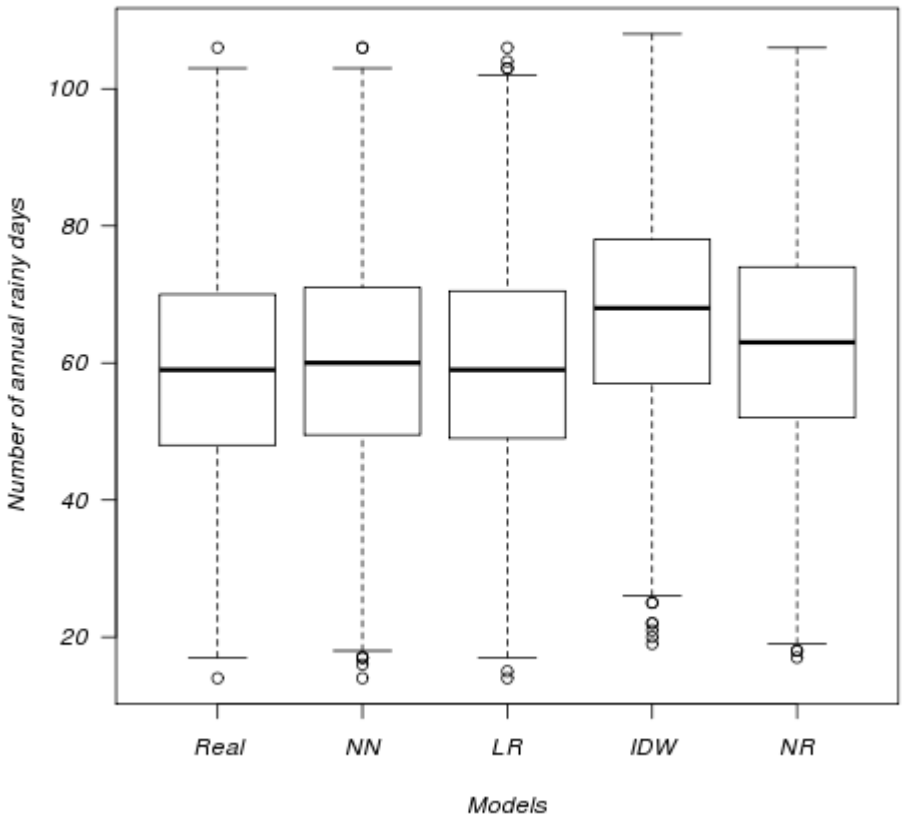

Fig.04-11_Boxplots of the fitting and the real values, for the four methods. Annual rainy days.

methods; NR shows a slight underestimation for the high values. Looking at the boxplots (**Fig.04-07**) NN and LR show the best accordance with the real estimations.

When annual accumulated values were considered (**Fig.04-08**), a more symmetric behaviour was noted (on average) for the LR method, but an overestimation is evident for all the methods when low values are considered. Boxplots in **Fig.04-09** show an overestimation of extreme values especially for LR method and a slight underestimation for IDW, but there is not a great difference between methods (the best seems to be NR).

Fig.04-10 presents scatterplots for the number of annual rainy days. Graphs of NN and LR methods show overestimation and underestimation for low and high values respectively; overestimation of both low and high values can be noted for IDW and overestimation of low values for NR ; however the best distributions were presented by NN and LR methods (**Fig.04-11**).

The behaviour of the fourth (NR) method with outliers could then be considered matching the specific structure and density of the network, even if the differences between methods 2 to 4 appears to be very small. Indeed, Borrough and McDonnell (1998) stated that when data are abundant most interpolation techniques give similar results.

Another analysis was therefore conducted, reducing the number of stations in the network; first decreasing the total number by 30 stations, then by 60 and 90 stations.

Tab.04-03 shows MAE, RMSE and extreme errors for each method and each of the four numbers of available stations. Decreasing the number of stations, the second method shows ~~a~~ greater robustness and it is able to reconstruct missing values even when few stations are available.

Tab.04-03_MAE, RMSE and extreme errors for each method and each of the four numbers of available stations: 109, 79, 49 and 19 stations.

Number of stations	Models	Extreme errors	MAE	RMSE
109	NN	199.4	1.3	4.5
	LR	142.4	1.0	3.4
	IDW	127.9	1.0	3.1
	NR	119.4	0.9	3.1
79	NN	152.0	1.1	3.9
	LR	112.2	1.0	3.6
	IDW	116.2	1.0	3.3
	NR	118.4	1.0	3.3
49	NN	163.8	1.3	4.3
	LR	98.4	1.1	3.8
	IDW	126.5	1.2	3.7
	NR	121.5	1.2	3.8
19	NN	131.8	1.7	5.7
	LR	93.7	1.4	4.6
	IDW	123.5	1.5	4.8
	NR	153.9	1.5	4.8

Conclusions

The results obtained depict a different behaviour of the four methods considered. When dealing with extreme values, the NR methods seems to be the most effective, while considering average values the performances of LR and IDW methods are almost equal to those of NR. However, the results obtained with a reduced set of stations showed that LR method presents a greater robustness when stations are more spread.

These results highlight the inherent difficulty of dealing with data characterised by a strong spatial and temporal variability such as rainfall and that the selection of the 'best' method should be done considering the purpose of the analysis (e.g. reconstruction of extreme events or identification of averages of subperiods) and the characteristics of the network, thus requiring a proper analysis on available data prior to the phase of data reconstruction.

References

Burrough P.A. and McDonnell R.A. 1998. Principles of Geographical Information Systems. New York: Oxford University Press.

Kim J.W. and Pachepsky Y.A.. Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. Journal of Hydrology. DOI : 10.1016/j.jhydrol.2010.09.005.

Tirozzi B, Puca S, Pittalis S, Bruschi A, Morucci S, Ferraro E, Corsini S. 2006. *Neural Networks and Sea Time Series: Reconstruction and Extreme-Event Analysis*. Modelling and Simulation in Science, Engineering and Technology. Birkhauser. P. 162. ISBN 10-8176-4347-8.

Paulhus J.L.H. and Kohler M.A. 1952. Interpolation of missing precipitation records. Mon. Wea. Rev., **80**, 129–133. doi: 10.1175/1520-0493(1952)

Young K.C. 1992. A three-way model for interpolating for monthly precipitation values. Monthly Weather Review. DOI : http://dx.doi.org/10.1175/1520-0493(1992)120

Vicente-Serrano S. M., Beguería S., López-Moreno J.I., García-Vera M.A. and Stepanek P. 2010. A complete daily precipitation database for northeast Spain: reconstruction, quality control, and homogeneity. Int. J. Climatol., 30: 1146–1163. doi: 10.1002/joc.1850

Eischeid J.K., Baker C.B., Karl T.R. and Diaz H.F. 1995. The quality control of long-term climatological data using objective data analysis. Journal of Applied Meteorology. http://dx.doi.org/10.1175/1520-0450(1995)

Chapter 5

Spatial and time correlation of thermometers and pluviometers in a weather network data-base

Introduction

The most familiar measure of dependence between two random variables is the Pearson's correlation coefficient (or linear correlation coefficient). It is obtained by dividing the covariance of two variables by the product of their standard deviations.

This index is very useful in the field of the applied climatology, for example when the aim of the work is to validate, homogenize or reconstruct the data base of a weather network, or to characterize the climatic zones concerning the studied area (WMO, 2011).

Regarding the reconstructing methods used to homogenize or fill missing values of a meteorological data base, e.g. data of thermometers and pluviometer grids, there are many examples where the correlation coefficient is used. The nearest neighbour, the linear regression (Serrano et al., 2009), the normal ratio (Young, 1995) and Eischeid's method (Eischeid, 1995) are some simple examples, where the correlation coefficient plays a leading role.

In these last examples, two important issues can be highlighted to make the correlation between a target station and the other ones easier to understand: a) the relation between the distance of a station from the target station and its correlation coefficient (with the target station) and b) the variation of the order of the stations within the unfolding of the history of the network (since its birth), when they are ranked by the correlation coefficients with the target station.

The first issue is important, for example when a method such as the nearest neighbour is considered. In this case a relation between the distance of a station from a target station and their correlation has to be evaluated to allow its application.

Though the main aim of the paper of BenHamida et al. (2009) was to propose an original synchronous average-based decomposition of the time series, the first issue is mentioned in this paper, where spatial correlation was studied, as a function of the distance.

An analysis concerning the second issue should allow, for example, to establish a station to be steadily coupled with another from a certain point in the history of the network onwards. This pair could definitely be used to check the data mutually, thereby enabling direct validation and reconstruction.

The analysis of this second issue can have another more general use; it allows to understand

if some studies carried out on a network for the whole period of the time series can be applied also to the future. In fact, if the correlations orders are stable over the time, they could be fixed without other future calculations.

In this paper, the analysis was carried out over the area of the Veneto Region (Italy), which was subdivided into three main climatic zones: the mountains, the plain and the coast.

This study deals with daily precipitation and daily maximum, mean and minimum temperature. Temperature and precipitation were considered because of their very different distribution in space and time. Moreover, sea and altitude are two important factors that influence climate conditions, and 29 % of the surface of the Veneto Region is mountainous. Its coast extends for 140 km of the Adriatic Sea.

Materials and methods

The data.

Data span from 1st January 1993 to 31th December 2008 (5844 days), for both temperature and precipitation. The network has a total of 112 pluviometers and 114 thermometers, distributed over 18400 km^2 of the Veneto Region. All 112 pluviometers are in the same site of a thermometer. Each instrument produced a number of missing data that did not exceed 5 per cent of the spanned period.
The aim of this work was to carry out analyses on daily data of precipitation, and maximum, mean and minimum temperatures (Tmax, Tmean and Tmin).

Data reconstruction.

Missing data have been reconstructed prior to the present data analysis. All kinds of temperatures (Tmax, Tmean and Tmin) were reconstructed with the new dynamic method presented in Chapter 2, while missing precipitation data, on the basis of the results obtained in Chapter 4, were reconstructed with the linear regression method (LR) (Serrano, 2009).

Methods

The method can be subdivided into two phases.
The first is to understand the relationship between the distance of a station from the target station and its correlation coefficient, considering the whole history of the network as sampling size of the correlation formula.
The second phase is to study the variation of the order of the stations, in the unfolding of the history of the network (since its birth), when they are ranked by the correlation

coefficient with the target station. In this case the sampling size is not only the whole history of the network but varies from a minimum to a maximum value (the whole history of the network).

In the first phase, for each target station (spanning the whole set of stations) the correlation coefficients of any other station with the target one are obtained. All these stations are ranked by the value of the coefficient and sorted in descending order. The distance from target station are matched to each of these correlation values.
Three matrices are obtained. A matrix of correlation coefficients, $\mathbf{c}_{ij}$, where i=1,...,112 (precipitations) or 1,...,114 (temperature) spanning the whole set of stations (target stations), and j=1,...,111 or 1,...,113 spanning the number of the other stations except the target; it can be noted that $\mathbf{c}_{ij}>\mathbf{c}_{ik}$ when $j<k$.
The distances matched to the $\mathbf{c}_{ij}$ coefficients are described by the other matrices, $\mathbf{d}_{ij}$; note that distances are generally not sorted. A third matrix is defined as $\mathbf{d'}_{ij}$, where $\mathbf{d'}_{ij}<\mathbf{d'}_{ik}$ when $j<k$.
To obtain a relationship between coefficients and related distances, the averages $\mathbf{m}_j$=mean($\mathbf{d}_{ij}$) are calculated, obtaining for each j the mean distance of the coefficients that are in the j-th column of the $\mathbf{c}_{ij}$ matrix. This vector will be compared with the vector $\mathbf{m'}_j$=mean($\mathbf{d'}_{ij}$). The purpose will be to analyse the distance, from the target station, of the most correlated stations with the target station.

In the second phase, a similar matrix of the first phase is obtained by computing the correlation coefficients with the first n days of the history of the network: n varied from 4 (due to the natural application of the Pearson's formula) to 5844 spanning the 16 years of the series, yielding a number of 5841 matrices. Finally, a three-dimension array, $\mathbf{c'}_{ijk}$ (where $\mathbf{c'}_{ijk}$ is a matrix for each k=1,...,5841), is obtained. In the same way as the first phase, an array $\mathbf{d'}_{ijk}$ is matched to the $\mathbf{c'}_{ijk}$ array; in this second phase, for each k, $\mathbf{d'}_{ijk}$ is sorted by ascending order: at j=1 position there is the distance of the nearest station to the target station, i.

Considering all $\mathbf{d'}_{ijk}$ entries of the matrix i, with j=1,2 and k=w,w+1, it can be said that the

order of correlations changes, passing from $k=w$ to $k=w+1$, if $\mathbf{c}'_{i1w}<\mathbf{c}'_{i2w}$ and $\mathbf{c}'_{i1(w+1)}>\mathbf{c}'_{i2(w+1)}$ (or $\mathbf{c}'_{i1w}>\mathbf{c}'_{i2w}$ and $\mathbf{c}'_{i1(w+1)}<\mathbf{c}'_{i2(w+1)}$).
More generally, given m, for each u ranging from 1 to m (the m stations closest to the target station), and $k=w,w+1$, it can be said that the order of correlations changes, passing from $k=w$ to $k=w+1$, if $\mathbf{c}'_{iuw}<\mathbf{c}'_{iuw}$ (or $\mathbf{c}'_{iuw}>\mathbf{c}'_{iuw}$) for some u.

For each $m=2,...,113$ (or 111 for precipitations) and $w=1,...,5840$, the presence of some changes in order is monitored.

To investigate the degree of permanence of the correlation order in time, per size of the groups of nearest stations, a definition of "stability" is proposed: the group of the m closest stations to the target station is considered "stable" if given a value of q days, in 100 percent of the cases, in which $w \geq q$, there are no changes in order (for the m stations considered).

The value of q days is set because of the transient period characterizing all sequences of correlation coefficients, calculated between any pair of stations, from the beginning to the end of the time series. **Fig.05-01** and **Fig.05-02** show two examples for precipitations and maximum temperature, for plain, mountain and coast.
It can be noted that the curves of these figures, up to a specific value of sampling size, have irregular oscillations with a large amplitude especially for precipitation; therefore, in this period (transient) talking about stability does not make any sense.
Not even, an excessive value of q must be considered, otherwise once again the concept of stability does not make sense. Moreover, the number of $w \geq q$ has to be as large as possible, to permit statistical evaluations.

These two phases are performed for precipitations, the three kinds of temperature and for the three climatic zones (mountain, plain and coast).

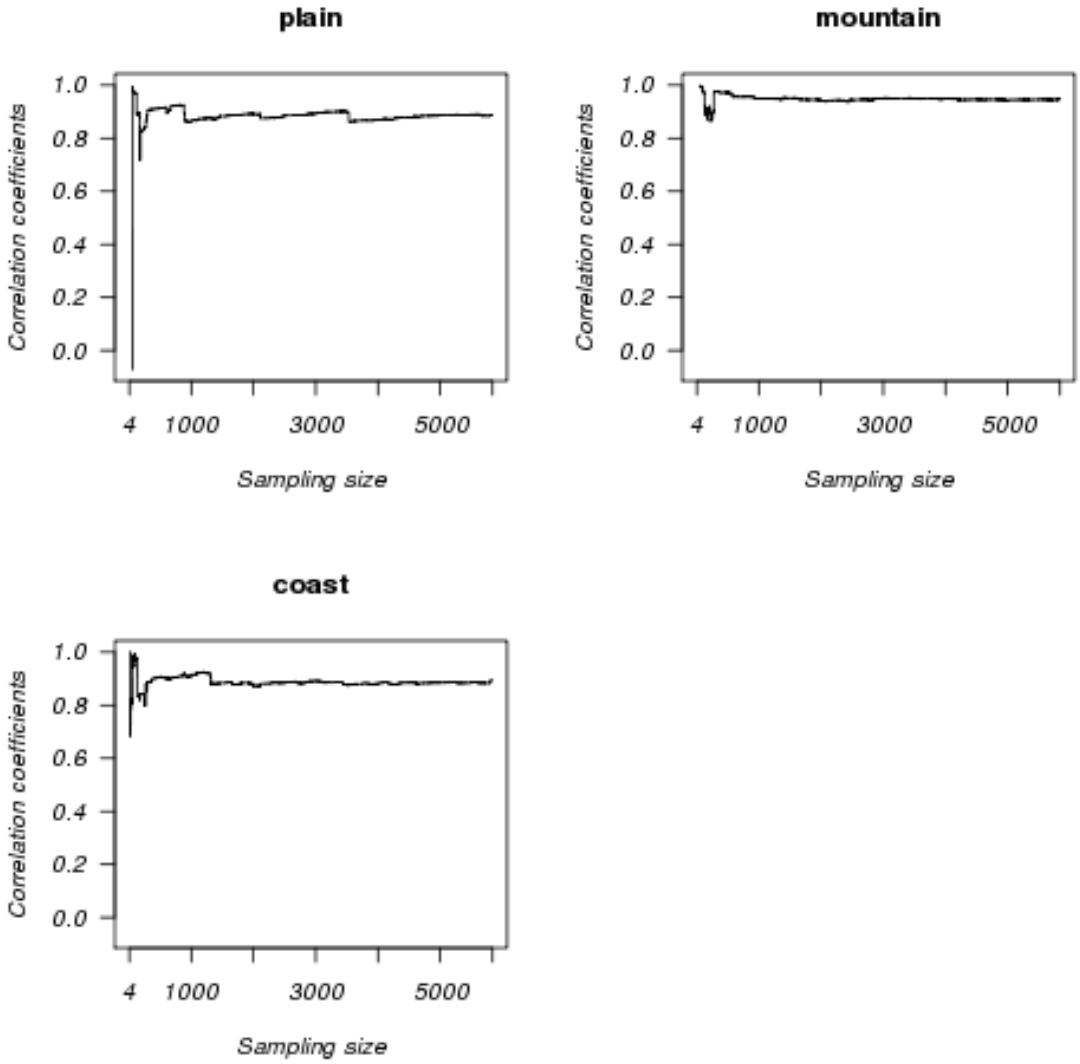

Fig.05-01_Correlation coefficients as a function of sampling size from the beginning of the time series. Three examples for precipitation data: Plain, Mountain and Coast.

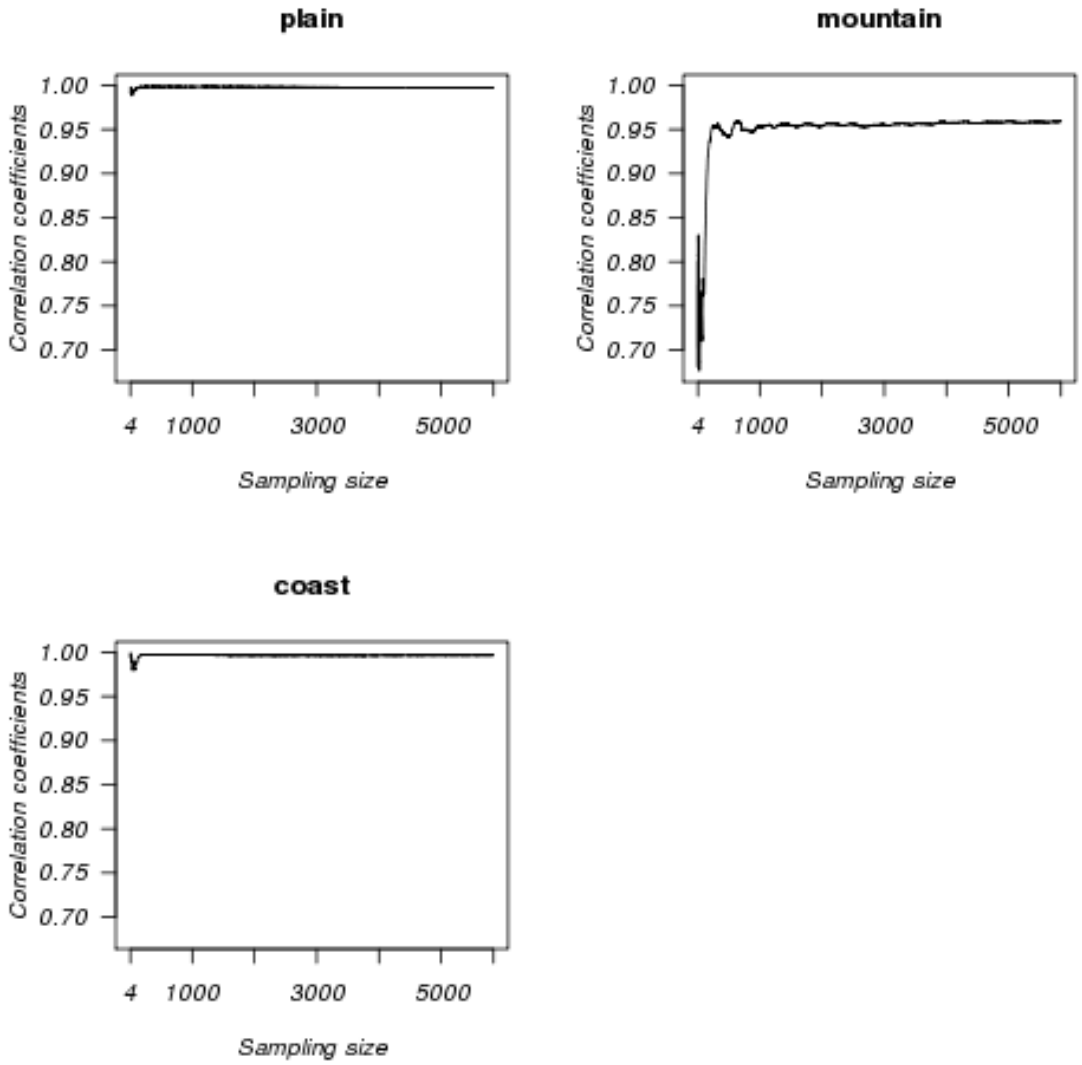

Fig.05-02_Correlation coefficients as a function of sampling size from the beginning of the time series. Three examples for maximum temperature data: Plain, Mountain and Coast.

Results and discussion

It must be noted that no more than 5 percent of the data base used for this work were missing, so that, a series of reconstructions were carried out to fill in these gaps, permitting calculations that would otherwise have been impossible.

The relationship between correlation and distance can be seen in **Fig.05-03** and **Fig.05-04**, where correlation, as a function of distance from target station (see BenHamida, 2009), is presented for precipitation and maximum temperature (the same behaviour was found for Tmean and Tmin): an example for each climatic zone.

In comparison with temperature, precipitation data present a closer dependence on the distance of the stations from the target station. For precipitations, at great distances the r value can reach very low values, while the correlation for temperature remain very high even at a great distance.

This fact can be related to the precipitation system that characterizes the climate of the Veneto Region: the database of rainfall shows high time and spatial gradients; there is a marked difference between the rainfall phenomenology of the mountain and the plain.

As described by the first phase of the method, the difference $\mathbf{m}_j$-$\mathbf{m'}_j$, varying j, is presented by **Fig.05-05** for precipitation and **Fig.05-06** for minimum temperature (the same behaviour was noted for Tmax and Tmean); graphs show data for the whole Region, plain, mountain and coast.

It can be observed that, in comparison with temperature, the mean distance of the stations (pluviometers) sorted by coefficients ($\mathbf{m}_j$) are rather similar to the mean distances sorted by order of distance ($\mathbf{m'}_j$).

In **Fig.05-05** (precipitations) it is important to note that, unlike the mountain and plain, for the coast the most correlated stations are, on average, at a greater distance from the target station: approximately 10 km more than the mean distance, $\mathbf{m'}_j$.

This could be related to the influence of the sea on the rainy system of the coast.

Looking at the Regional graph of temperature (**Fig.05-06**), a greater distance from the target station of the most correlated stations is shown, and this fact is mainly influenced

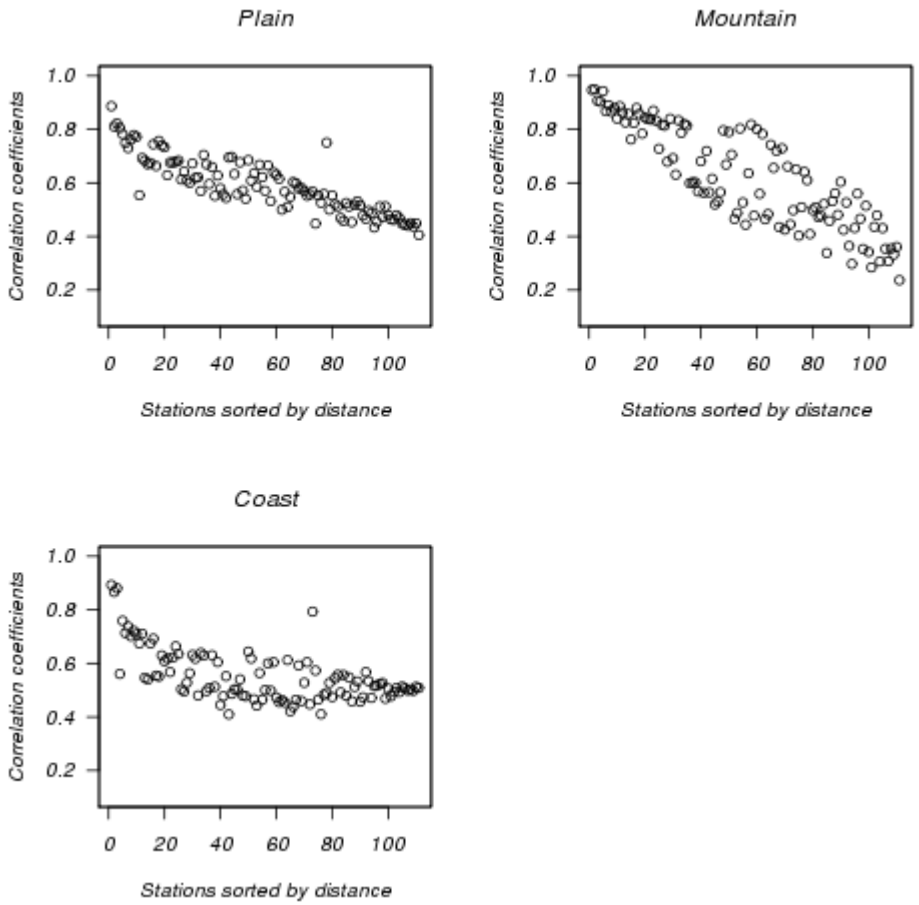

Fig.05-03_Correlation coefficients (calculated over the whole time series) as a function of the distance of stations from target station. Three examples for precipitation data: Plain, Mountain and Coast.

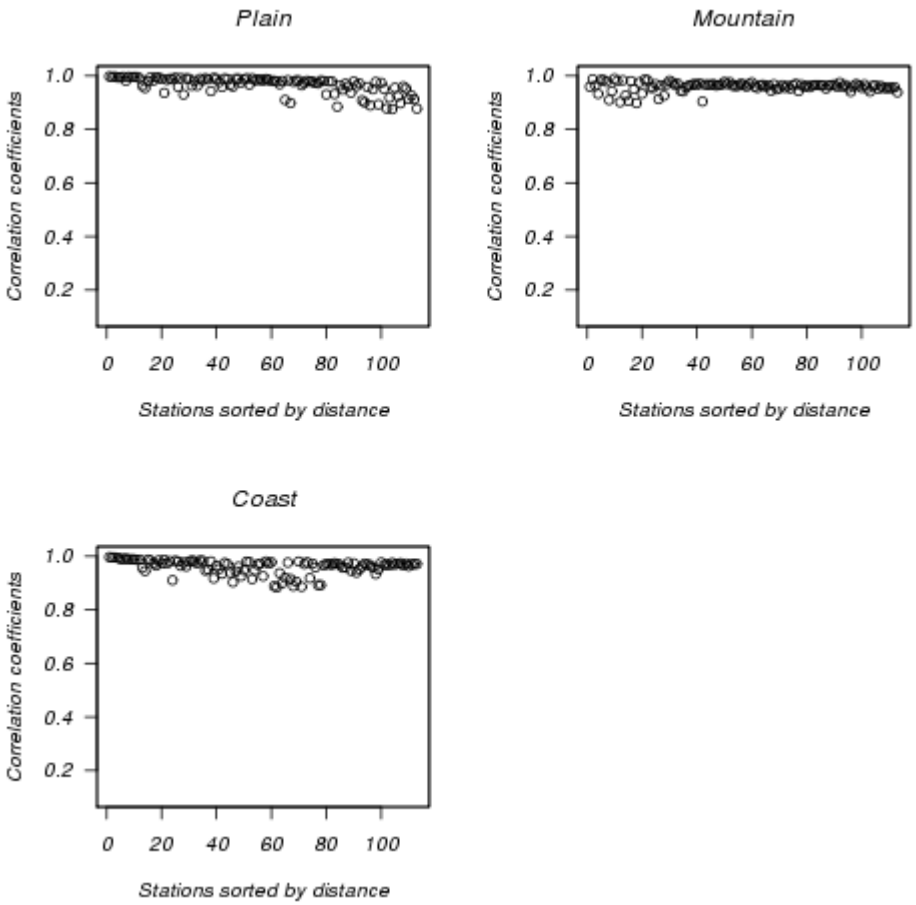

Fig.05-04_Correlation coefficients (calculated over the whole time series) as a function of the distance of stations from the target station. Three examples for maximum temperature data: Plain, Mountain and Coast.

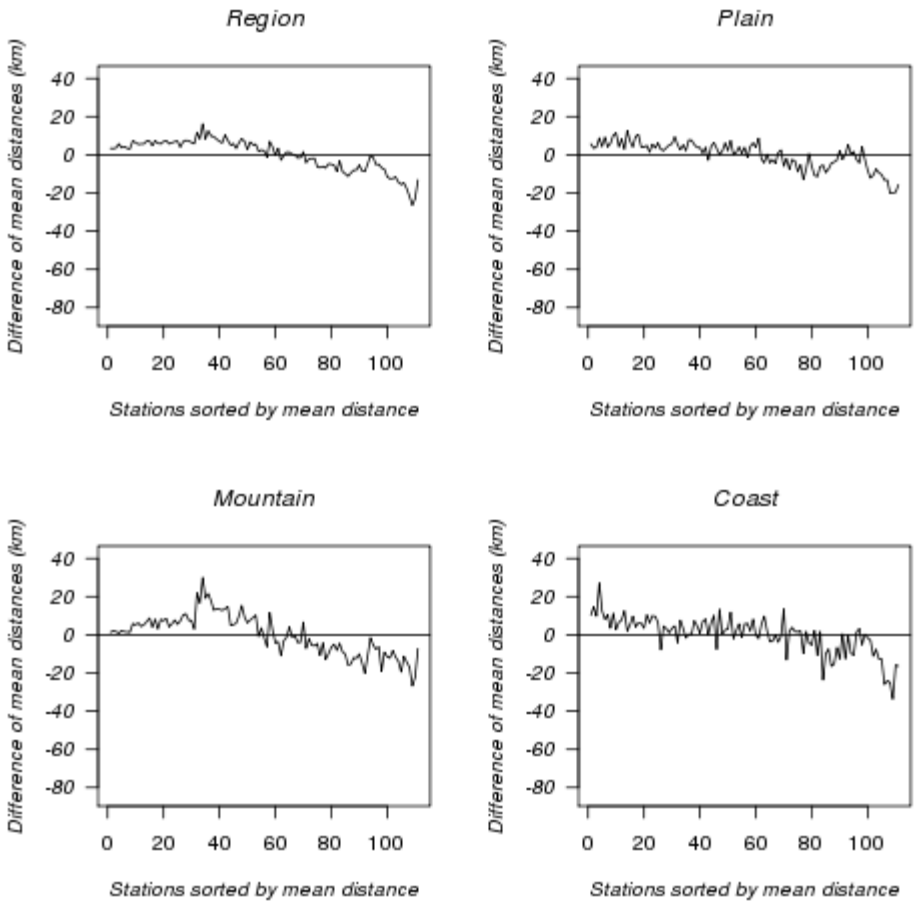

Fig.05-05_Difference between the mean distance of the *j*-th best coefficients and the mean distance of the *j*-th closest stations from their respective target stations ($\mathbf{m}_j$-$\mathbf{m}'_j$, varying *j*). For precipitation: Region, Plain, Mountain and Coast.

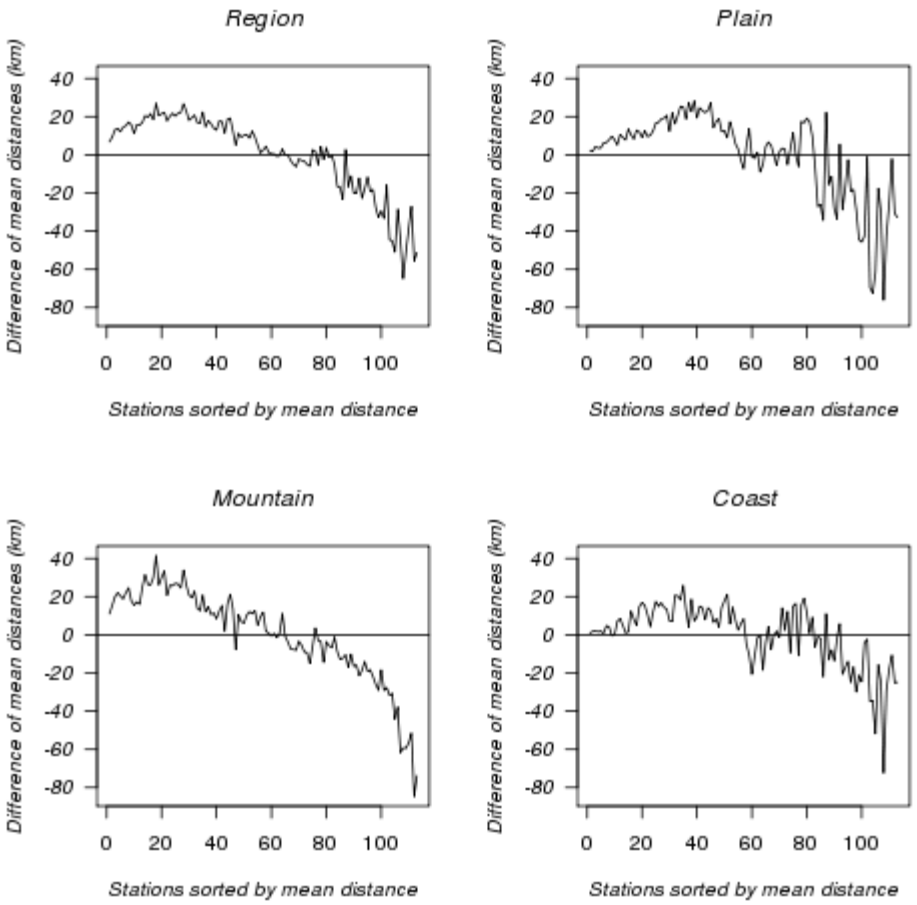

Fig.05-06_Difference between the mean distance of the *j*-th best coefficients and the mean distance of the *j*-th closest stations from their respective target stations ($\mathbf{m}_j$-$\mathbf{m'}_j$, varying *j*). For Tmin: Region, Plain, Mountain and Coast.

by the behaviour of the mountain area: in this climatic zone the most correlated stations can be, on average, at 15 km more distant than the mean distance from the target station ($\mathbf{m'}_j$).

This behaviour could be related to the great diversity of altitude values and sun exposure that characterize the placement of thermometers in the mountainous area (Collins and Bolstad, 1996).

Time correlation is analysed in the second phase of the method. As mentioned previously, setting *q* is very important for this phase. The transient period and the magnitude of 5840-*q* (as large as possible), must be taken into account. Drawing on experience, *q* was set at 2000 days.

The values of *m* (defined before), for all target pluviometers, meeting the defined criterion of stability, are shown through **Fig.05-07**. These boxplots indicate that over the whole area, 25 % of stations do not present any *m* of stability, 75 % have the two closest stations that are stable, at least; with a maximum value of *m* equal to 7. Comparing the three zones, it was noted that the best behaviour was for the plain that shows the same percentage of stations without stability (25 %).

Referring to temperature, **Fig.05-08** shows the same as **Fig.05-07**, for Tmax, Tmean and Tmin. It was found that over the whole area, 17 %, 15 % and 7 %, respectively for Tmax, Tmean and Tmin, do not present stability for any *m* value. In this case, the best behaviour was reached in the mountain zone that presents 5 and 9 % (of unstable stations) for Tmax and Tmean, and 0 % for Tmin.

A maximum value of *m* is reached with Tmean: *m*=21 stations (mountain).

Comparing **Fig.05-08** with **Fig.05-07**, it is clear that temperature is generally more stable than precipitation, and as might be expected, this fact may be due to the cases in which precipitation does not show a limited transient, as is clearly shown in the plain-graph of **Fig.05-01**.

This observation makes it possible to deduce that, from a certain point of the history of the network onwards, the data base of temperature maintains the correlation order within a wide radius from the target station.

The complete graph of the average of the distances of the stations from a target station ($\mathbf{m'}_j$)

is reported in **Fig.05-09**. The radius of preservation of the order of correlation in time, for precipitations and Tmean (**Tab.05-01**) is of 11 and 28 km respectively for precipitations, and 18 and 41 km respectively for Tmean.

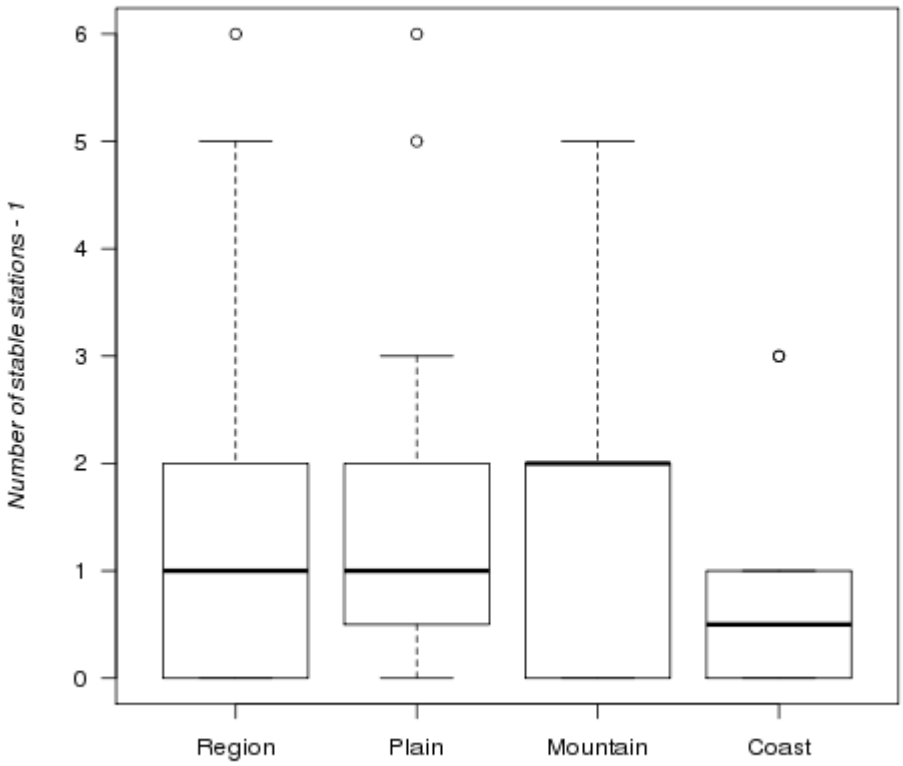

Fig.05-07_Boxplot of the values of m for which stations have stability (the correlation order of the closest *m* stations is stable). In the graph, 0 means no stability. This graph is for precipitation: Region, Plain, Mountain and Coast.

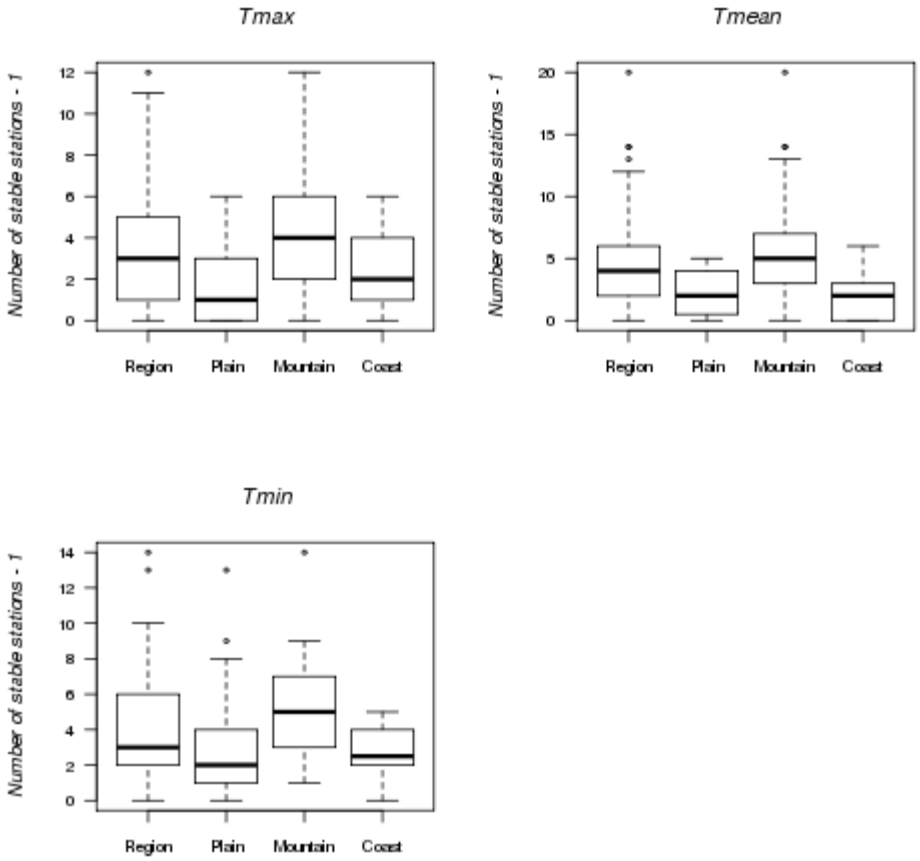

Fig.05-08_Boxplots of the values of m for which stations have stability (the correlation order of the closest *m* stations is stable). In the graph, 0 means no stability. These graphs are for temperature (Tmax, Tmean and Tmin): Region, Plain, Mountain and Coast.

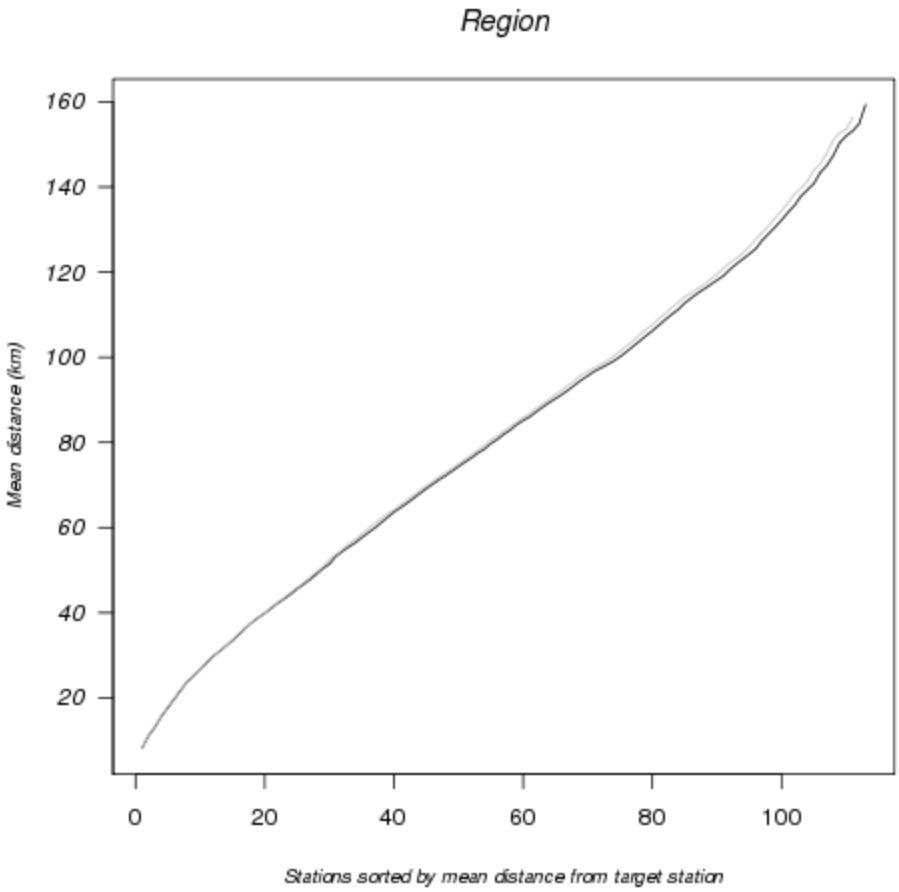

Fig.05-09_The mean distance of the stations, sorted by distance from the target station ($\mathbf{m}'_j$). The grey line correspond to the precipitation data. Precipitation and temperature values are almost the same.

Tab.05-01_Median and maximum values of the values of m (the m stations closest to the target station) and the matched m'_j (mean distance of the j-th station closest to the target station). Table shows precipitation, Tmax, Tmean and Tmin values.

	m		Mean distance (km)	
	Median	Max	Median	Max
Prec	2	7	11	28
Tmax	4	13	16	31
Tmean	5	21	18	41
Tmin	4	15	16	33

Conclusions

The correlation coefficients of the precipitation database are, on average, inversely proportional to the mean distances from the target station. The same has not been shown by temperature data that, however, always show a high correlation coefficient regardless of the distance.

The sentence: “the correlation coefficients are higher for the closest stations”, is generally true for precipitation data, except for the stations of the coastline, while it is frequently not true for temperature, particularly in the mountain zone.

However, from 5.5 years (q=2000 days) from the beginning of the time series, the temperature variable is characterized by a high stability of the correlation order in time. The highest values are shown by the mean temperature; whereas the mountain area .of the minimum temperature does not have any unstable station.

From a more general point of view this result also indicate that, from a certain year of the history of the network (in our case: about the fifth year), the calculation of the Pearson' s coefficient acquires an important stability, regardless of the type of variable (precipitation or temperature).

References

BenHamida E, Borgnat P, Esaki H, Abry P, Fleury E (2009). Live E! Sensor Network: Correlations in Time and Space. In the *XXIIe Colloque* GRETSI 2009 - *Traitement du Signal et des Images*, Dijon, France, September 8-11, 2009.

Collins FC, Bolstad PV (1996). A comparison of spatial interpolation techniques in temperature estimation. In: Proceedings of the Third international Conference/Workshop on Integrating GIS and Environmental Modeling, Santa Fe, New Mexico, January 21-25, 1996. Santa Barbara, California: National Center for Geographic Information Analysis (NCGIA). CD-ROM.

Eischeid JK, Baker CB, Karl TR, Diaz HF (1995). The Quality Control of Long-Term Climatological Data Using Objective Data Analysis. *J. Appl. Meteor.* 34: 2787-2795. doi: 10.1175 /1520-0450(1995).

Tardivo G, Berti A (2012). A dynamic method for gap filling in daily temperature datasets. *J. Appl. Meteor. Climatol.*, **51:** 1079–1086. doi: 10.1175/JAMC-D-11-0117.1.

Vicente-Serrano SM, Beguería S, López-Moreno JI, García-Vera MA, Stepanek P (2010). A complete daily precipitation database for northeast Spain: reconstruction, quality control, and homogeneity. *Int. J. Climatol.*, **30**:1146–1163. doi:10.1002/joc.1850.

WMO (2011). Guide to Climatological Practices. 3rd ed. WMO No.100, 180 pp. [Available online at http://www.wmo.int/pages/prog/wcp/ccl/documents/WMO_100_en.pdf].

Young KC (1992). A Three-Way Model for Interpolating for Monthly Precipitation Values. *Mon. Wea. Rev.*, **120**, 2561–2569. doi: 10.1175/1520-0493(1992).

Chapter 6

General Conclusions

Conclusions

The first method (temperature data reconstructions) could seem quite complex at first sight and the analysis is described by two independent chapters (articles), though the kernel of this method makes use of one of the simplest statistical models, namely, the multiple linear regression. The novelty underlying this dynamic technique is the system of cross-validations trials that allows the algorithm to adapt the parameters used to "shape" the multiregression formula in the best way for each gap. When the series of cross-validations are carried out to evaluate the performance of gap filling, it is known in advance that the method will have the highest performance possible. In fact, the trials carried out initially to "shape" the multiregression formula (for each gap) work with the algorithm and in the same space-time location of the subsequently running cross-validation system.
Through these two chapters, a better performance of the adaptive regression methods in comparison with non-adaptive regression ones can be demonstrated, at least as far as temperature data are concerned.
Looking to the method more closely, some issues that are not mentioned in the papers can be outlined. For example, a deeper analysis on the variation of performance using reconstructed data to fill other gaps has not been already done. However, the experience and unpublished calculations show this use of reconstructed data does not affect the system significantly. In future papers, issues of this nature will be formalised and analysed in depth.

The method selected for reconstructing of precipitation data is very simple (LR method). Considering the whole network, this methods performs similarly to other approaches, but the trials carried out with a reduced set of stations proved its effectiveness and robustness, at least for the network considered.
When the number of stations of this network is reduced, the philosophy of seeking the "shape" of a rainfall event, using more than one station near the target station (combined with a specific and proper formula), is superseded by the more simple approach of using the data of the most correlated station only: less attention to the spatial and greater attention to

the time relationship is, therefore, afforded.
We can note that in both cases (temperature and precipitation), a linear model was used to reconstruct the data. This model is very close to our way of reasoning: trying to forecast some event of an unknown phenomenology by linking it directly with other events whose phenomenology is known seems to be natural. Consequently, linear and multilinear regression models appear to be more simple and easy to manage to our minds in comparison with other tangled statistical techniques. It is worth noting that reconstruction of missing data is done for relatively short periods, where the assumption of stability of meteorological conditions are applicable.

The chapter referring to the Pearson's coefficient (chapter 5) demonstrates that some years after the birth of the network (about 5 years, in our case), the correlation system became relatively stable, and many calculations linked to this system (correlation system) can be considered to be relatively definitive in the future, too. As a consequence, studies on spatial correlation, as in our case, can be carried out without the availability of long time series, as required by the concept of Climate Normal of 30 years (WMO).

Furthermore the sentence “the correlation coefficients are higher for the closest stations”, can be not true, especially when talking about the temperature variable which can anyway be considered to be relatively continuous.

A comparison between the third and the fifth chapters emphasize the difference between linear or multilinear environments. In the first case the most correlated predictors are always quite close to the target station, while in a multiregression environment, predictors with a high multilinear correlation can be found even at very long distances from the target station, even up to 80 km away.

Acknowledgements

I am particularly grateful for the assistence and teaching given by my supervisor Prof. Antonio Berti.

I would like to express my deep gratitude to Claudia, Marco, my family (Sergio, Paola, Rossano, Francesca, Alessio, Francesco, Carolina, any babies, etc…).

I wish to thank the Meteorological Centre of Teolo, Fabio, Gianni, and all my collegues (DAFNAE).