Science Journal of Environmental Engineering Research
ISSN:2276-7495
http://www.sjpub.org/sjeer.html

Published By
Science Journal Publication
International Open Access Publisher

Research Article **Volume 2013, Article ID sjeer-265, 9 Pages, 2013. doi: 10.7237/sjeer/265**

COMPARISON OF FOUR METHODS TO FILL THE GAPS IN DAILY PRECIPITATION DATA COLLECTED BY A DENSE WEATHER NETWORK

Gianmarco Tardivo and Antonio Berti

Dipartimento di Agronomia Animali Alimenti Risorse Naturali e Ambiente.
Università degli Studi di Padova. Viale dell'Università 16 - 35020 Legnaro (PD) - ITALY.

Accepted 6th September, 2013

ABSTRACT

Daily precipitation data are often useful for running climatological models; nowadays these models make frequent use of computational and algorithmic approaches that require no missing values. Four straightforward methods to reconstruct gaps in precipitation databases have been considered and compared through a series of statistical indexes and applications to some practical issues using the daily precipitation database of the Veneto Region (Italy). The methods are compared from many points of view: estimating extreme errors of reconstruction; pairing observed rainfall values and respective reconstructing errors of each method; ability to predict monthly and annual accumulations, and monthly and annual rainy days; varying the network density. In the first case, a modified Normal Ratio method seems to have the best behaviour; in the second case, the modified Normal Ratio gives the same results as Linear Regression and Inverse Distance Weighting methods, while in the two last cases, Linear Regression seems to be the best performer, showing also a greater robustness when reducing the density of the network. The results highlight the inherent difficulty of dealing with data characterised by a strong spatial and temporal variability such as rainfall. The choice of the reconstruction method should be done considering both the purpose of the analysis (e.g. reconstruction of extreme events or identification of averages of subperiods) and the characteristics of the network and/or the climatic traits of the environment studied.

KEYWORDS: Gap filling; weather networks; precipitation data; reconstruction methods.

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).

These databases frequently presents gaps that should be reconstructed for subsequent analyses; the reconstruction method adopted is however one of the most important issues to obtain relevant information about rainfall, because they can affect the reliability of the results. Moreover, the lack of reconstructions could lead to underestimation errors of aggregated values (Eccel et al., 2012). Two important issues can be highlighted about the filling of a precipitation database: the ability of a reconstructing method to allow accurate computations on mean values (Xia, 1999a and Xia, 1999b) (e.g. on monthly or annual periods); and the ability to reconstruct extreme values (Allen and DeGaetano, 2001; Eischeid et al., 1995).

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: tipically, precipitation is characterized by higher 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 (Calza et al., 2008). 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. 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 (Jeffrey, 2001) and thin plate smoothing splines (Price et al., 2000). In many cases, such as basin flow evaluation (hydrology), Artificial Neural Network models (ANN) are very reliable (Kim and Pachepsky, 2010; Khorsandy, 2011; Coulubaly, 2007), but when more accurate evaluations of extreme values are required, ANN models are less effective in reproducing the events (Tirozzi et al., 2006). Other straightforward methods, such as Normal-Ratio (NR) (Paulhus and Kohler, 1952; Young, 1992), Multiple Linear Regression (MLR) (Eischeid, 1995), Multiple Discriminant Analysis (MDA) (Young, 1992), Nearest-Neighbour (NN) (Vicente-Serrano et al., 2009), Inverse Distance Weighting (IDW) (Vicente-Serrano et al., 2009) and Linear Regression (LR) (Vicente-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. The inverse-distance method, in particular, is recommended for interpolations using spatially dense networks (Dirks et al., 1998). Geostatistical methods are very interesting and frequently effective (Bajat et al., 2013), but a possible problem with their implementation is the need of some kind of supervision for the analysis of semi-variograms (Ly et al., 2013), thus limiting the automatization of these approaches. In this work, NN, IDW, LR and a slightly modified Young' s NR

Corresponding Author: Gianmarco Tardivo
Dipartimento di Agronomia Animali Alimenti Risorse Naturali e Ambiente. Università degli Studi di Padova. Viale dell'Università 16 - 35020 Legnaro (PD) - ITALY.
Email address: gian0812@gmail.com

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

2.0 MATERIALS AND METHODS

2.1 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 (Fig_01). 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.

2.2. Methods

Three of the four methods compared are presented in detail in Vicente-Serrano et al. (2009). Some differences from their application in this paper are due to the network and area involved and are described below.

Nearest-neighbour method (NN). The main aim of this method is to find the station with the highest correlation with the target one within a given search radius, to be used as predictor. Vicente-Serrano et al. (2009) used a search radius of 15 km and a minimum correlation threshold r=0.5; in the present paper these parameters were set to 40 km and r≥0.6 due to the differences in the spatial densities and distribution of the correlations between this dataset and the one used by Vicente-Serrano et al. (2009).

The gaps are filled directly with data from the closest station meeting the criteria. A further requirement of NN is the availability of a sufficient number of common data (i.e. non-missing days in both stations); in our case, the low percentage of missing days in the database makes it possible to not consider this requirement.

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, to avoid negative values and to retain the zeros. Only the slope coefficient was used to provide reconstructed data (Vicente-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. Vicente-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 and x_T the resulting target value.

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); 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: a range of exponents were tested in the neighbourhood of the value of 1.75. where a local minimum of extreme errors were found; the weight of Young was also tested, showing higher values of extreme errors.

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 with a large number of cross-validations, in which each reconstructed value was compared with the observed one.

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

3.0 RESULTS AND DISCUSSION

A preliminary analysis was carried out to evaluate the potential of ANN methods with the approach proposed by Kim and Pachepsky (2010). They used a regression tree method for the selection of the predictors and artificial neural network (feed-forward backpropagation) for data reconstruction. The ANNs were trained using the Levenberg–Marquardt algorithm (Kim and Pachepsky, 2010) to minimize the root mean square error between the actual and simulated daily precipitation at the target weather station. These tests showed extreme errors higher than 1000 mm per day, regardless of this method was easily automatizable and presented an optimal performance for not extreme errors. This fact lead us to not consider this method for further analysis.

A total of 515117 cross-validations, corresponding to all the non-missing days (both rainy and dry), were done to evaluate and compare the four methods.

For all four methods, the upper range of error centiles (from 88^{th} to 100^{th} centiles) are shown in Fig_02. Up to the 98^{th} 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.

How to Cite this Article: Gianmarco Tardivo and Antonio Berti "Comparison of Four Methods to Fill the Gaps in Daily Precipitation Data Collected by a Dense Weather Network" Science Journal of Environmental Engineering Research, Volume 2013, Article ID sjeer-265, 9 Pages, 2013. doi: 10.7237/sjeer/265

Fig 1: Distribution of meteorological stations across Veneto Region.

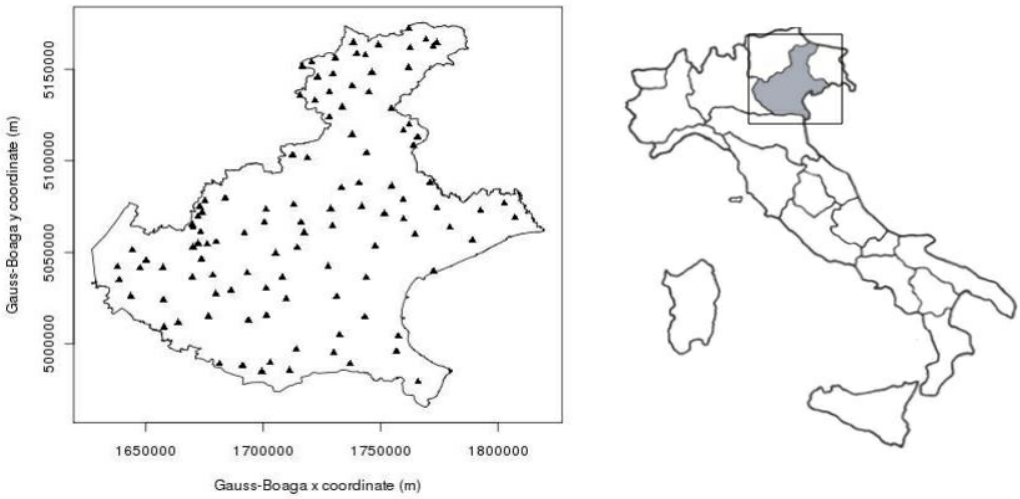
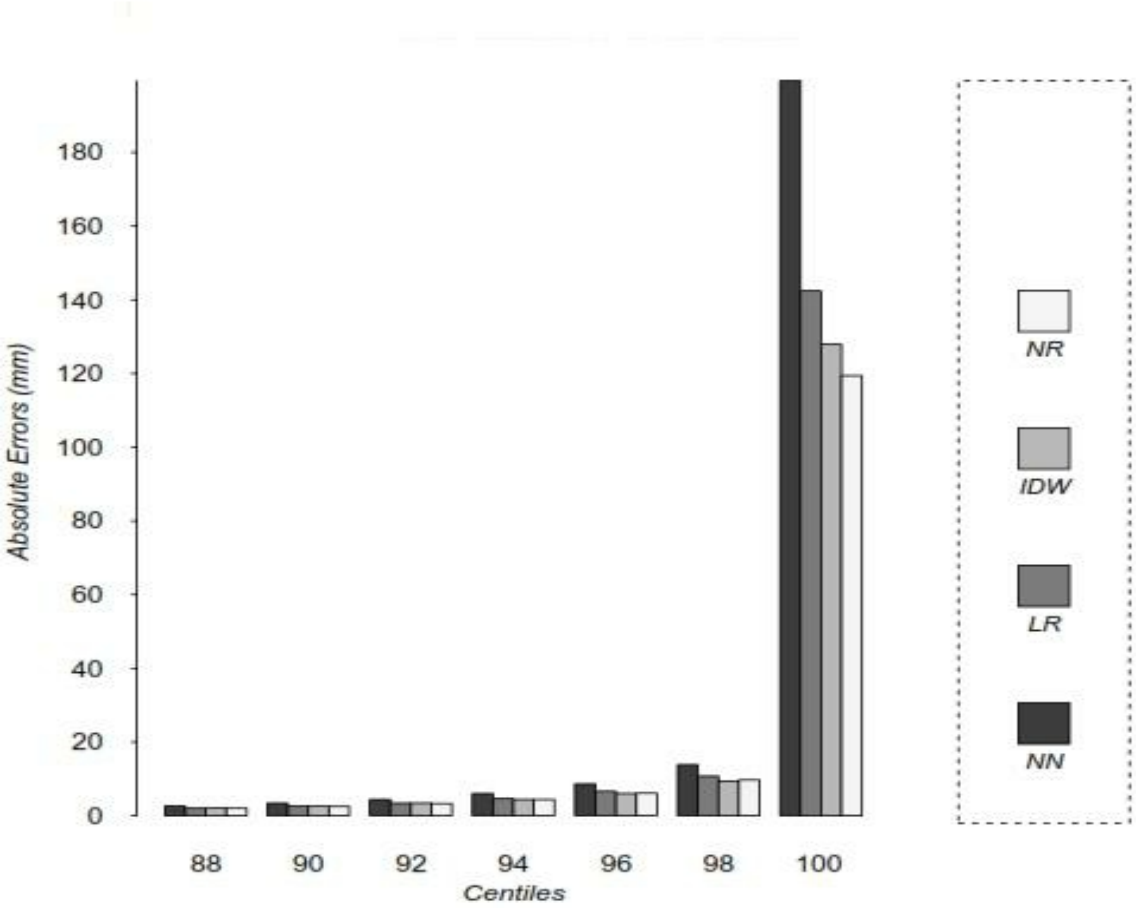


Fig 2: Centiles of absolute errors (from the 88th to the 100th centile) through the whole set of cross-validations. For the four methods.



How to Cite this Article: Gianmarco Tardivo and Antonio Berti "Comparison of Four Methods to Fill the Gaps in Daily Precipitation Data Collected by a Dense Weather Network" Science Journal of Environmental Engineering Research, Volume 2013, Article ID sjeer-265, 9 Pages, 2013. doi: 10.7237/sjeer/265

It is worth noting that IDW, while presenting a slightly higher MAE than NR, has the lowest RMSE (Tab_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_03).

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.

Fig3: Box-plots of the values deduced from the bootstraps carried out to estimate the significance of the performance of NR method. The values are show for NN, LR, IDW and NR methods. Bootstraps are carried out for extreme errors, MAE and RMSE.

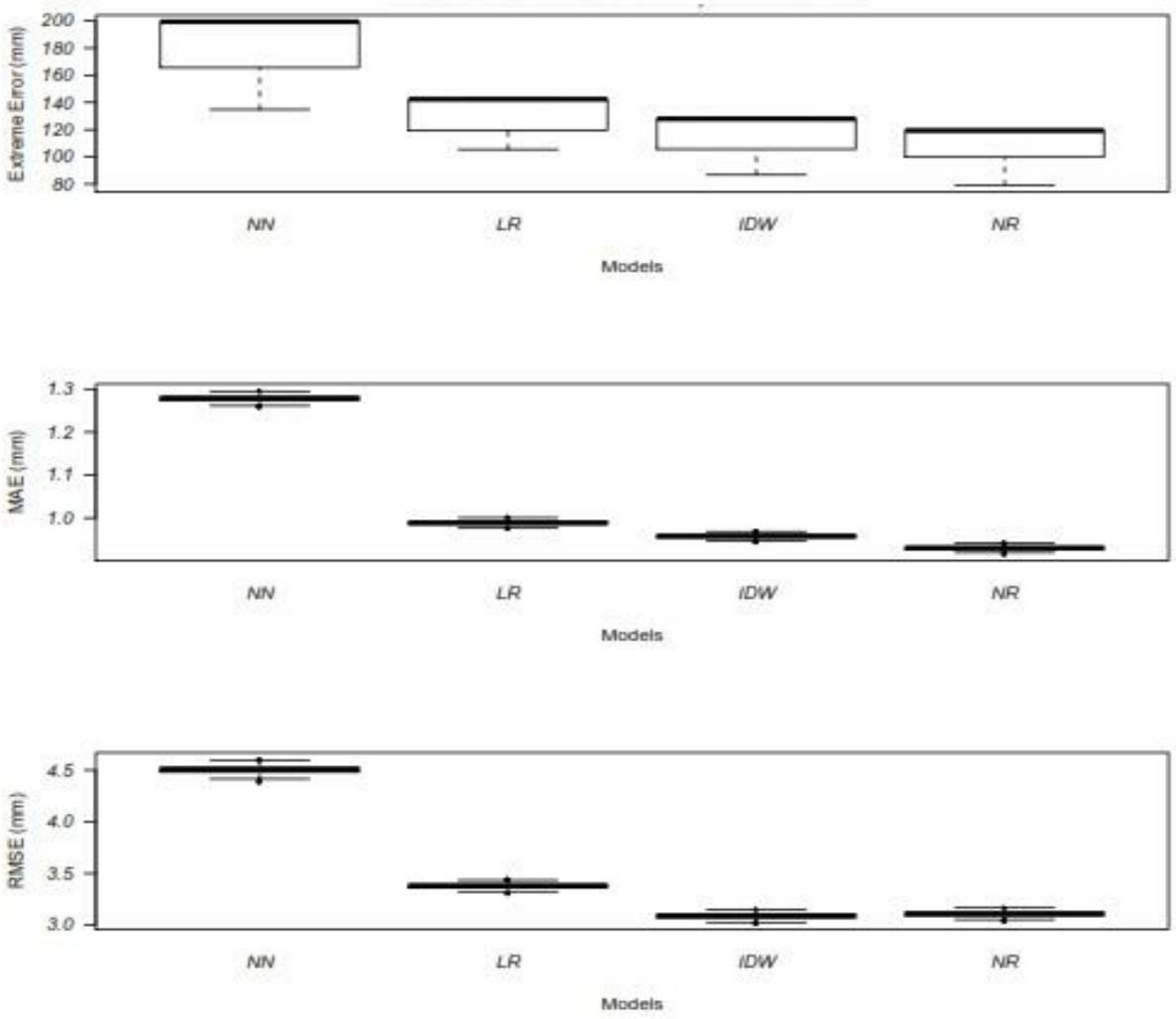


Table 1: Maximum absolute error (MAE) and Root Mean Square Error (RMSE) of the four methods compared. These values are obtained for all the cross-validations

Method	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 in the case of outliers, 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) $+75^{th}$ centile or below 25^{th}- 1.5×IRQ of the distribution of errors, are considered outliers of errors. Tab_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 = 2 mm; 2 < p = 20 mm; 20 < p = 40 mm; 40 < p = 70 mm; 70 < p = 88.4 mm; p > 88.4 mm. It can be seen that when p is equal to 0, the NR-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 NR-method is always among the lowest.

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 used by Eischeid et al., 1995), all values greater than 88.4 mm were found as outliers in the whole series (see Fig_04).

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.

Fig_05-08 can be analysed to reveal the goodness of the methods, comparing each one with the others.

For monthly accumulated values, Fig_05 (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). Considering the monthly number of rainy days (Fig. 06), IDW seems not to underestimate, on average, in comparison with the other methods; NR shows a slight underestimation for the high values. When annual accumulated values were considered (Fig_07), 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.

Fig_08 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.

In all these four tests (monthly and annual, accumulated and rainy days) the distributions of the reconstructed values were compared with that of the real ones (data not shown): generally LR and NN distributions seemed to be the most similar to the real one.

The behaviour of the fourth method (NR) 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.

To evaluate the robustness of the methods with more sparse networks, a further analysis was done, reducing the number of available stations. From the original grid of 109 stations, other three sub-grids were obtained eliminating 30, 60 and 90 stations randomly.

For each sub-grid the cross-validation process was then carried out and maximum error, MAE and RMSE were obtained for each method and number of stations (Fig_09). The NN methods confirms to be less adapted to the environmental conditions tested, having higher values of all the parameters, independently from the number of stations considered.

The other three methods performs similarly, even if LR seems to be less prone to extreme errors with sparse networks. On the other hand, NR, which gives the smaller extreme errors with the original grid, appears to be heavily affected by the decrease of the number of available stations.

Table 2: The numbers of outliers of errors and their mean values matched to each sub-indicated interval: p=0 mm; 0 < p ≤ 2 mm; 2 < p ≤ 20 mm; 20 < p ≤ 40 mm; 40 < p ≤ 70 mm; 70 < p ≤ 88.4 mm; p > 88.4 mm. Over a total of 515117 cross-validations.

Method	Number	Mean	Number	Mean	Number	Mean
	p = 0 mm		0 < p ≤ 2 mm		2 < p ≤ 20 mm	
NN	28889	1.16	9554	5.66	5373	18.48
LR	30657	0·60	8894	3.88	5485	13.58
IDW	47967	0·72	9465	4.17	5074	11·65
NR	63685	0·40	10311	3.65	5251	11·90
	20 < p ≤ 40 mm		40 < p ≤ 70 mm		70 < p ≤ 88.4 mm	
NN	457	40.55	104	64.31	18	91.11
LR	567	28.63	152	45.67	17	65.38
IDW	608	25.48	172	42.72	33	59.58
NR	590	26.48	174	43.35	27	61.33
	p > 88.4 mm					
NN	9	126.62				
LR	22	85.85				
IDW	9	98.89				
NR	17	85.01				

How to Cite this Article: Gianmarco Tardivo and Antonio Berti "Comparison of Four Methods to Fill the Gaps in Daily Precipitation Data Collected by a Dense Weather Network" Science Journal of Environmental Engineering Research, Volume 2013, Article ID sjeer-265, 9 Pages, 2013. doi: 10.7237/sjeer/265

Fig 4: Relationship between the number of outliers and the threshold factor 'f' (Eischeid et al., 1995).

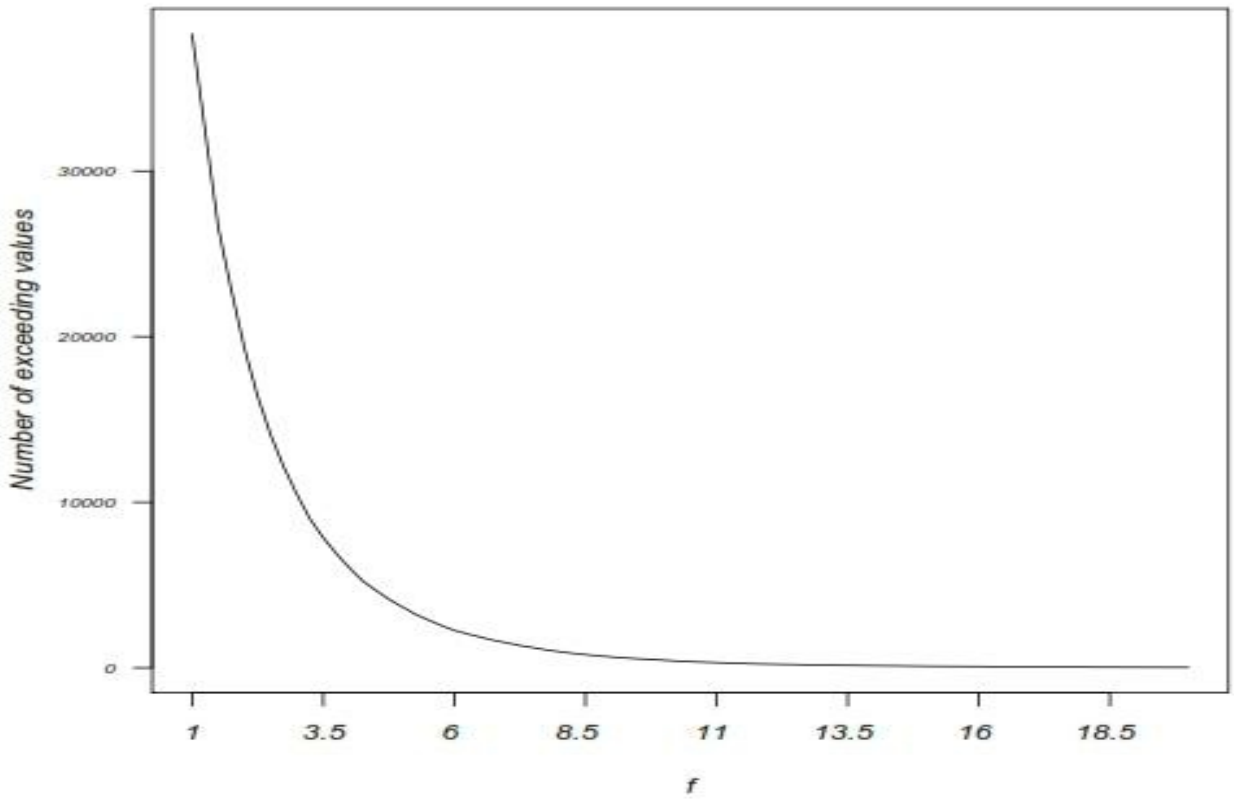


Fig 5: Scatterplots of the fitting values with the real ones, for the four methods. Monthly accumulated values.

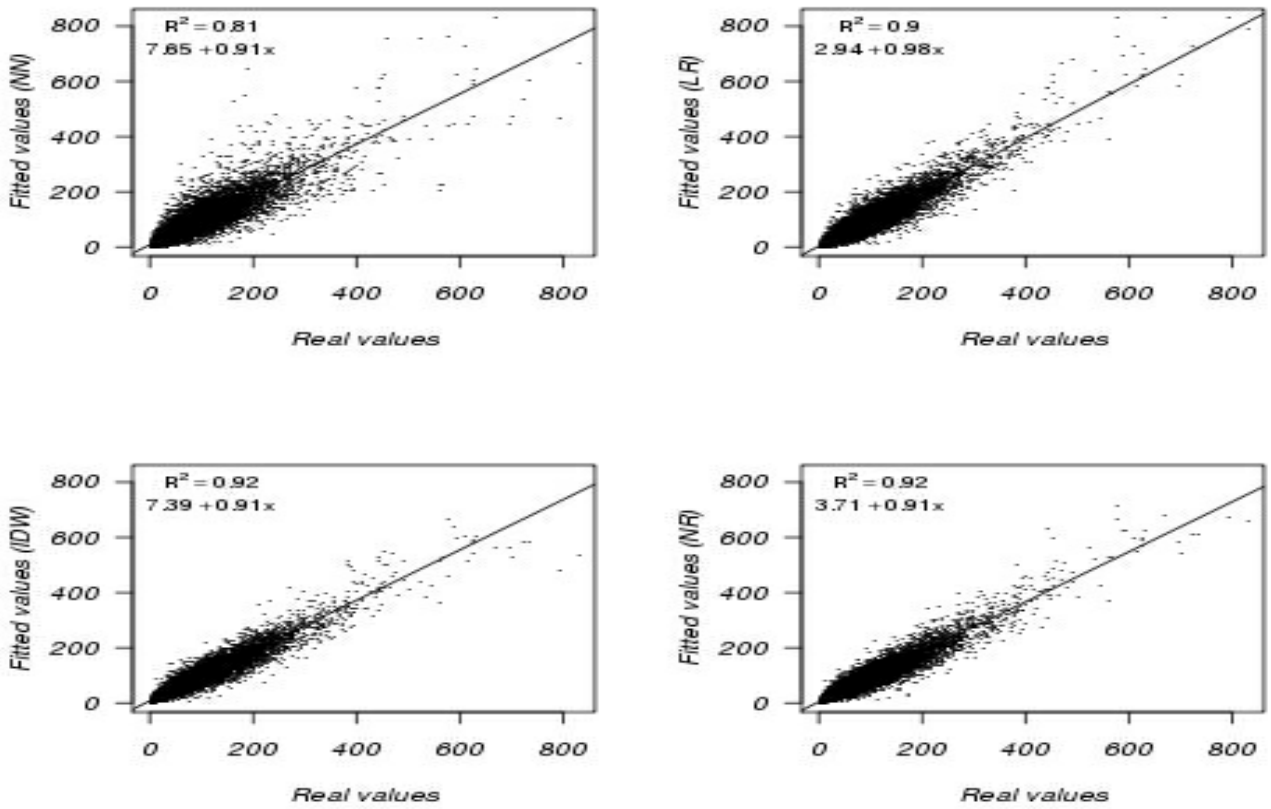


Fig 6: Scatterplots of the fitting values with the real ones, for the four methods. Monthly rainy days.

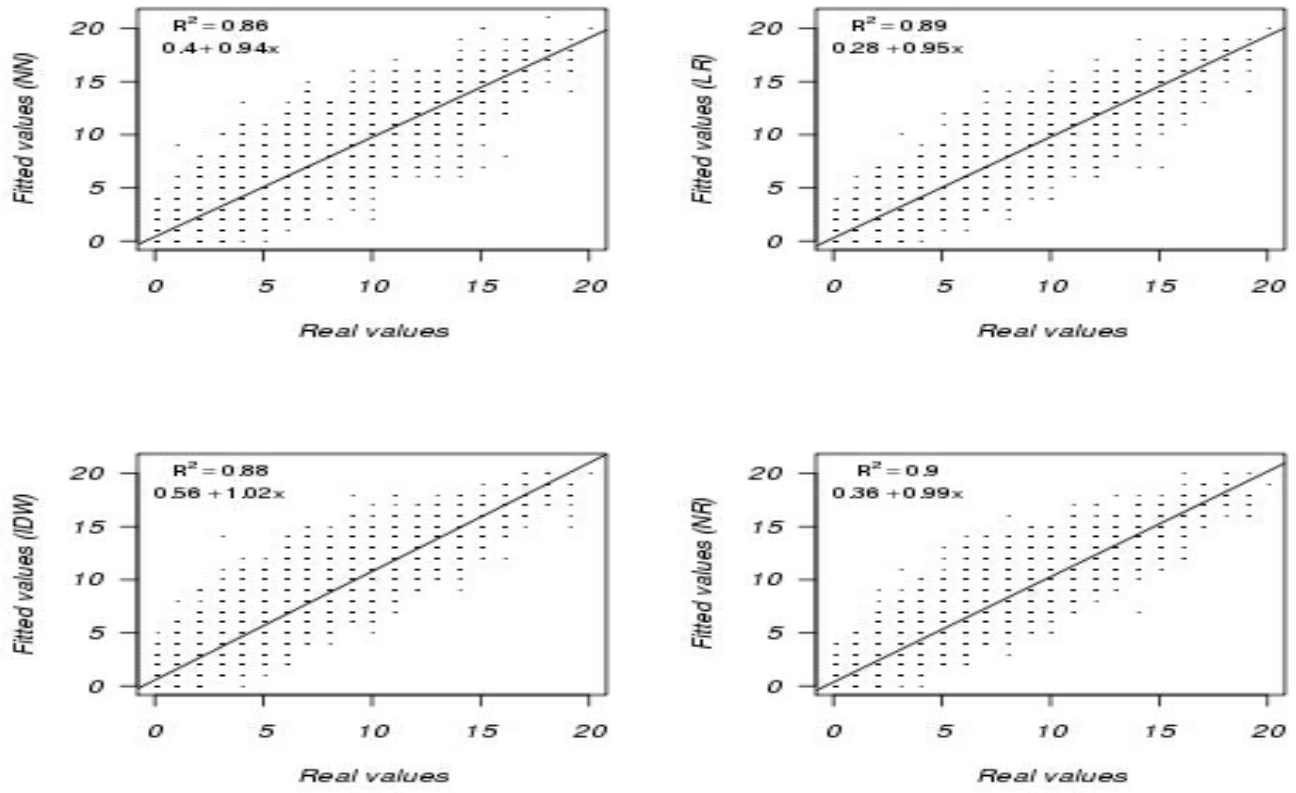


Fig 7: Scatterplots of the fitting values with the real ones, for the four methods. Annual accumulated values.

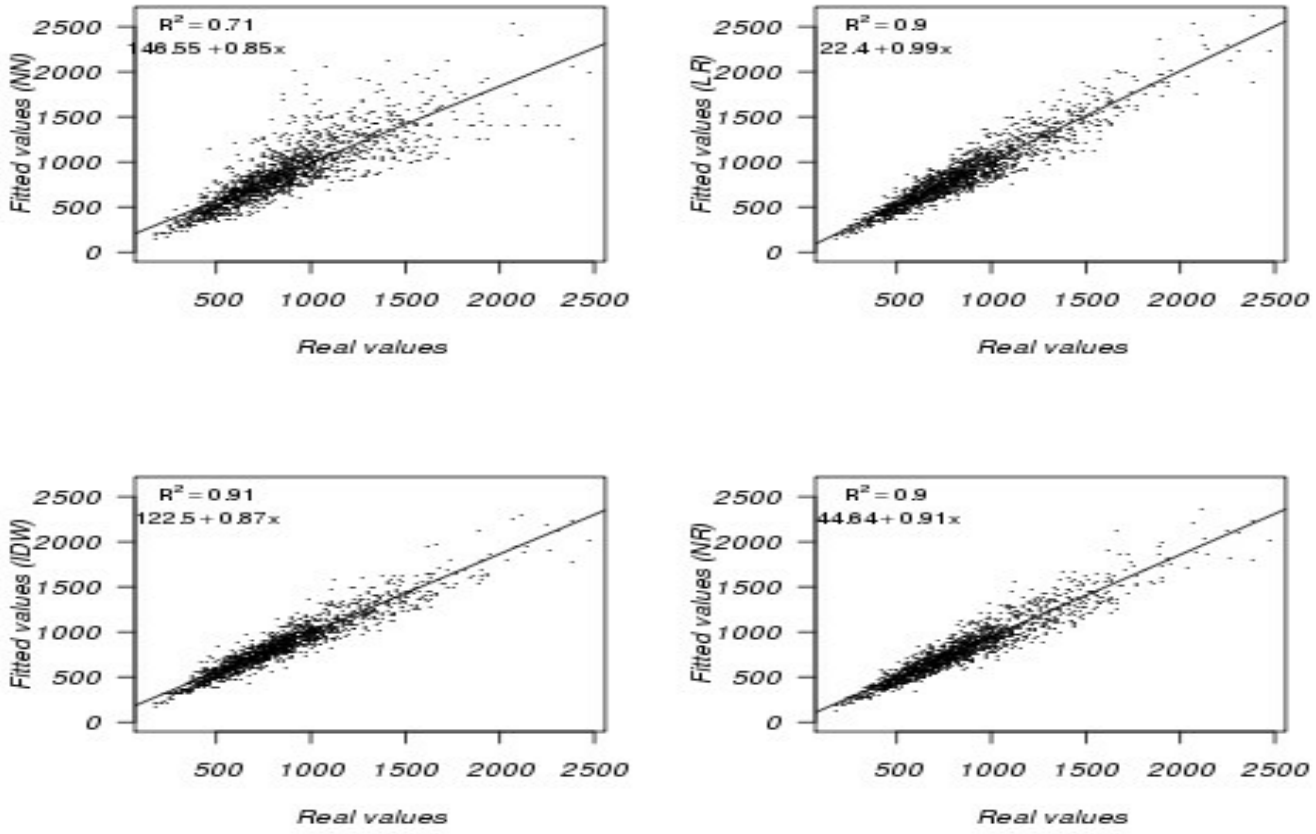


Fig 8: Scatterplots of the fitting values with the real ones, for the four methods. Annual rainy days.

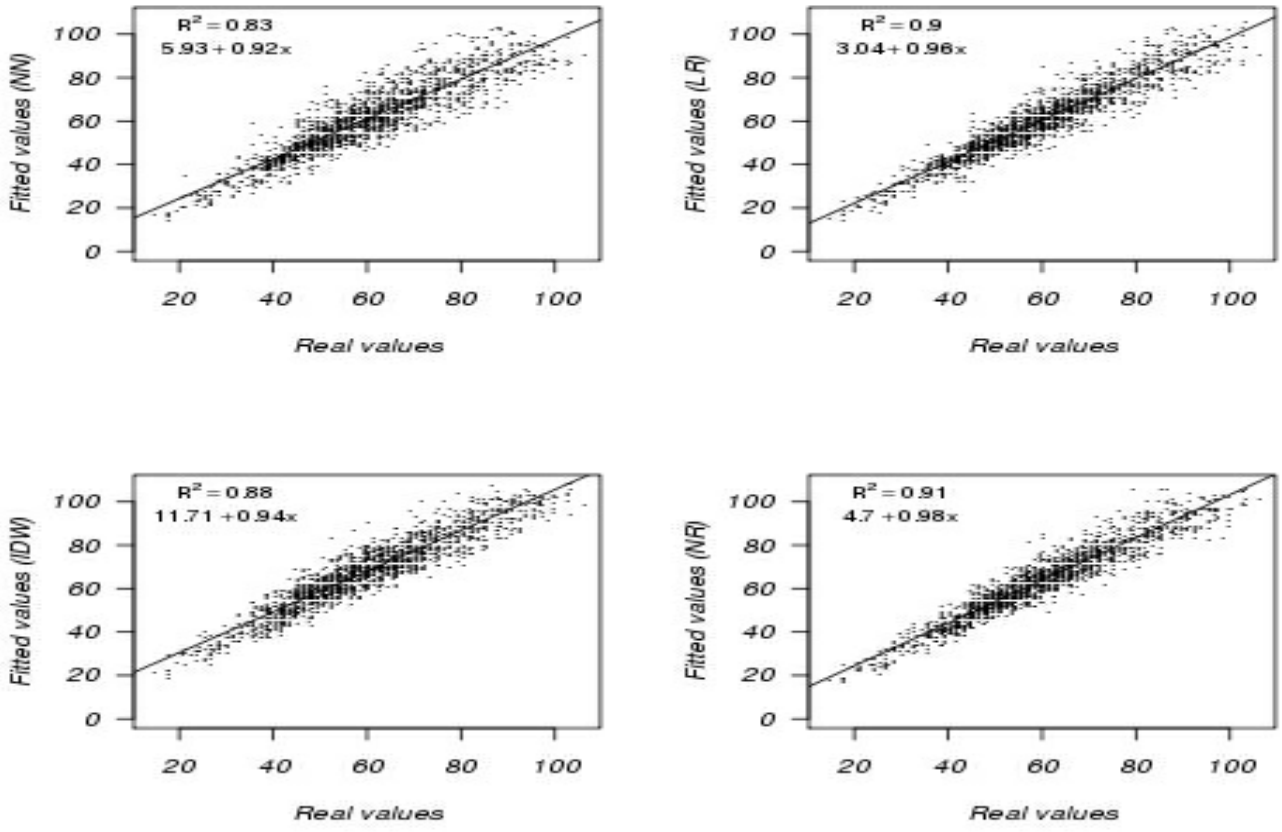
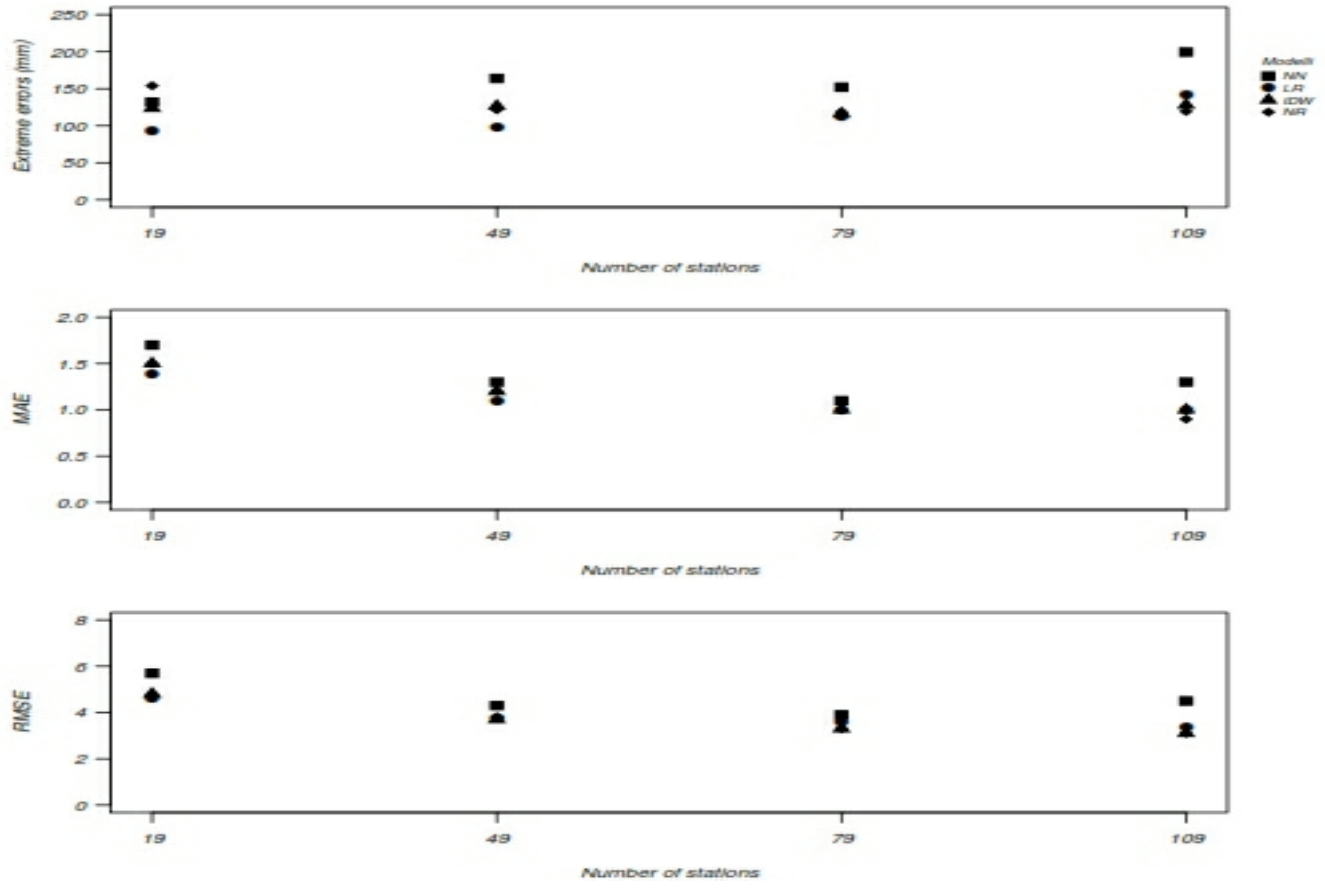


Fig 9: MAE, RMSE and extreme errors for each method and each of the four numbers of available stations: 109, 79, 49 and 19 stations.



4.0 CONCLUSIONS

The results obtained depict a different behaviour of the four methods considered. For the number of rainy days, the four methods gave a quite similar behaviour, while differences were clear considering the reconstruction of rainfall data.

When dealing with extreme values, the NR methods seems to be the most effective, while considering averages, 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. A further positive aspect of this method is that it is very simple to implement.

Our results are quite different from those of Serrano et al. (2009), which identified the NN method as the one giving statistics closest to the original record, maintaining the distribution characteristics of the original series. It is worth noting that the network used by these Authors is more dense than ours (on average 1 station every 51.3 km^2 against 1 station every 168.7 km^2 for the Veneto Region network) and the NN method appeared to be the more influenced by the density of the network with a steep increase of errors decreasing the number of stations considered.

These results highlight the inherent difficulty of dealing with data characterised by a strong spatial and temporal variability such as rainfall. The choice of the reconstruction method should be done considering on one hand the purpose of the analysis (e.g. reconstruction of extreme events or identification of averages of subperiods) and in the other the characteristics of the network considered and/or the climatic traits of the environment studied, thus requiring a proper analysis on available data prior to the phase of data reconstruction.

REFERENCES

1. Allen RJ and DeGaetano AT (2001). Estimating missing daily temperature extremes using an optimized regression approach. Int. J. Climatol. 21, 1305-1319.

2. Bajat B, Pejović M, Luković J, Manojlović P, Ducić V and Mustafić S (2013): Mapping average annual precipitation in Serbia (1961-1990) by using regression kriging, Theoretical and Applied Climatology, 112(1-2), 1-13. DOI: 10.1007/s00704-012-0702-2.

3. Burrough PA and McDonnell RA (1998). Principles of Geographical Information Systems, second ed. Oxford University Press, New York.

4. Calza M, DallaFontana A, Domenichini F, Monai M and Rossa AM (2008). 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.

5. Coulibaly P and Evora ND (2007). Comparison of neural network methods for infilling missing daily weather records. J. Hydrol. 341, 27-41.

6. Dirks KND, Hay JE, Stow CD and Harris D (1998). High-resolution studies of rainfall on Norfolk Island: Part II: Interpolation of rainfall data. J. Hydrol. 208, 187-193.

7. Eccel E, Cau P and Ranzi R (2012) Data reconstruction and homogenization for reducing uncertainties in high-resolution climate analysis in Alpine regions. Theor Appl Climatol 110(3):345-358.

8. Eischeid JK, Baker CB, Karl TR and Diaz HF (1995). The quality control of long-term climatological data using objective data analysis. J. Appl. Meteorol. 34, 2787-2795.

9. Jeffrey SJ, Carter JO, Moodie KB and Beswick AR (2001). Using spatial interpolation to construct a comprehensive archive of Australian climate data. Environ. Modell. Softw. 16, 309-330.

10. Khorsandi Z, Mahdavi M, Salajeghe A and Eslamian SS (2011). Neural network application for monthly precipitation data reconstruction. J. Environ. Hydrol. 19, 1-12.

11. Kim JW and Pachepsky YA (2010). Reconstructing missing daily precipitation data using regression trees and artificial neural networks for SWAT streamflow simulation. J. Hydrol. 394, 305-314.

12. Ly S, Charles C and Degre A (2013): Different methods for spatial interpolation of rainfall data for operational hydrology and hydrological modeling at watershed scale. A review, Biotechnologie Agronomie Societe et Environnement, 17(2), 392-406.

13. Paulhus JLH and Kohler MA (1952). Interpolation of missing precipitation records. Mon. Weather Rev. 80, 129-133.

14. Price DT, McKennedy DW, Nalder IA, Hutchinson MF and Kesteven JL (2000). A comparison of two statistical methods for spatial interpolation of Canadian monthly mean climate data. Agr. Forest. Meteorol. 101, 81-94.

15. Tirozzi B, Puca S, Pittalis S, Bruschi A, Morucci S, Ferraro E and Corsini S (2006). Neural Networks and Sea Time Series: Reconstruction and Extreme-Event Analysis. Modelling and Simulation in Science, Engineering and Technology. Birkhauser.

16. Vicente-Serrano SM, Beguería S, Lόpez-Moreno JI, García-Vera MA and Stepanek P (2009). A complete daily precipitation database for northeast Spain: reconstruction, quality control, and homogeneity. Int. J. Climatol. 30, 1146-1163.

17. Xia Y, Fabian P, Stohl A and Winterhalter M (1999a). Forest climatology: reconstruction of mean climatological data for Bavaria, Germany. Agr. Forest. Meteorol. 96, 117-129

18. Xia Y, Fabian P, Stohl A and Winterhalter M (1999b). Forest climatology: Estimation of missing values for Bavaria,Germany. Agr. Forest. Meteorol. 96, 131-144

19. Young KC (1992). A three-way model for interpolating for monthly precipitation values. Mon. Weather Rev. 120, 2561-2569.

How to Cite this Article: Gianmarco Tardivo and Antonio Berti "Comparison of Four Methods to Fill the Gaps in Daily Precipitation Data Collected by a Dense Weather Network" Science Journal of Environmental Engineering Research, Volume 2013, Article ID sjeer-265, 9 Pages, 2013. doi: 10.7237/sjeer/265