COMPARISON OF MISSING VALUE IMPUTATION METHODS FOR TURKISH MONTHLY TOTAL PRECIPITATION DATA [1]

S. ASLAN*, C. YOZGATLIGİL*, C. İYİGÜN**, İ. BATMAZ*,
M. TÜRKEŞ***, H. TATLI***

**Department of Statistics, Middle East Technical University*
Ankara, TÜRKİYE
e-mail: sipan@metu.edu.tr
***Department of Industrial Engineering, Middle East Technical University*
Ankara, TÜRKİYE
****Department of Geography, Çanakkale Onsekiz Mart University*
Çanakkale, TÜRKİYE

Abstract

Climate related studies are required complete time series data to be used. On the other hand, considerable number of observations is missing in meteorological time series due to several reasons. This conflicting problem, however, can be overcame by imputing missing values using observations of correlated nearby climate stations. The main aim of this study, therefore, is to compare the performances of six different methods for imputing monthly total precipitation series obtained from stations located in two different climate regions of Türkiye. These include Single Arithmetic Average (SAA), Normal Ratio (NR), NR Weighted with Correlations (NRWC), Multi Layer Perceptron type Neural Network (MLPNN) and Expectation-Maximization Algorithm based on Monte Carlo Markov Chain (EMMCMC). In addition, we propose a modification in the EMMCMC method which uses the results of different imputation methods as reference series. Results show that both EMMCMC methods perform better than the other imputation methods considered in the study.

Keywords: EM Algorithm, MCMC, Multiple Imputation, Missing Value, Precipitation Data

1 Introduction

Missing data are frequently encountered in climate variables due to many reasons including failure in the observatory instruments, meteorological extremes and observation recording errors. Climatological studies such as determining the effects of climate change on climate variables, homogeneity analysis of series, cluster and multivariate statistical analysis need complete data which are recorded in many stations spread out the whole region of interest in the long run. For that reason, it is crucial for the success of these studies that the missing data exist in series should be handled carefully [5].

[1] This study has been supported by the Middle East Technical University under the project number BAP-2008-01-09-02.

There are several methods which consider only the temporal behavior of the series for filling out missing values. However, considering temporal correlations only may cause loss of spatial information [3]. So that missing values in a series can be imputed by also using correlated nearby climate stations' complete observations.

In this study, missing values which are created artificially in monthly total precipitation data obtained from the stations of Turkish State Metorological Service (TSMS) located at two different climate regions are imputed by using six methods. Single Arithmetic Average (SAA), Normal Ratio (NR) and NR Weighted with Correlations (NRWC) are the three simple methods used. On the other hand, Multi Layer Perceptron type Neural Network (MLPNN) and Expectation-Maximization Algorithm based on Monte Carlo Markov Chain (EMMCMC) methods incorporate complex computations that mostly due to spatial features of the series. In addition to these, we propose a modification in the EMMCMC method in which results of above imputation methods are used as reference series.

In Section 2, data used in the study are described. Imputation methods used are explained in Section 3. In Section 4, findings of the study, and in Section 5, conclusion and further studies are presented.

2 Data Description

To better evaluate the capability of imputation methods, data in the period of 1965-2006 are obtained from the stations located at two different climate regions of Türkiye. Accordingly, four stations are selected from Northwest Coastline (NWC) region of Türkiye, which has influenced by the rainy regime overall, and five stations are selected from Southeast (SW) region of Türkiye, which has considerably low rainy regime. Note here that these series are tested for homogeneity [2]. Bartin and Goksun stations are selected as the target stations, in which missing values are imputed, in the NWC and SW of Türkiye, respectively. The correlations among the selected target and reference series are calculated by a robust method. Note that correlations between the stations in the NWC are relatively low(at most 0.75) compared to the correlations between the stations in the SW(at least 0.75).

3 Imputation Methods Studied

Choosing the most appropriate imputation method depends on the mechanism that creates missing values in data [4]. There are three general class of missingness mechanism described in the literature: Missing at Random (MAR), Missing at Completely Random (MCAR) and Missing not at Random (MNAR). In our context, MAR is considered to be the most appropriate mechanism because missingness that occurs in precipitation series does not explicitly depend on the precipitation itself but it may depend on the other variables [5].

The imputation methods used in the study are as follows. SSA imputes missing values by the arithmetic average of concurrent observations in the neighbor stations having similar features with the target station [7]. NR imputes missing data with the help of weights obtained from the ratios of precipitation totals of nearby stations [6]. NRWC is similar to NR method described above. However, in this method, weights are obtained from the correlations between the target and reference stations [7, 8]. In MLPNN method, on the other hand, the series obtained from the reference stations and the target station are treated as the input and the output of the MLP model, respectively [1]. In EMMCMC, assuming multivariate normal distribution of data and MAR missingness mechanism, multiple imputations are done by using the parameter estimates obtained from the EM algorithm [4] as the initial estimates for MCMC. Note here that although the precipitation data do not usually validate the normality assumption, missing values can be imputed successfully by this method [9]. In addition, we propose a modification in EMMCMC algorithm in which the results obtained from the other methods are used as the observations of reference stations.

4 Findings

In this study, the performances of imputation methods are assessed and compared by filling out missing values which are artificially created in two series of precipitation data. For this purpose, NRMSE values are obtained as the ratio of the RMSE to the mean. In the calculations, Ms Excel, JMP, SAS and SPSS softwares are used. These results are given in Table 1. Results indicate that there is no statistically significant difference between the means of the original and imputed series for both of the target stations studied (p-value>0.1).

Table 1: NRMSE values of two series obtained by using different imputation methods.

Imputation Methods	**Goksun Station**	**Bartin Station**
SAA	0.06959	0.10731
NRWC	0.06965	0.11523
NR	0.06875	0.10694
MLPNN	0.06721	0.10515
EMMCMC	**0.06666**	0.10301
modified EMMCMC	0.06723	**0.10239**

5 Conclusion and Further Studies

Above findings show that EMMCMC algorithm performs better than the others for both series with respect to the NRMSE criterion. Furthermore, distributional characteristics and spatio-temporal movements of series are preserved after the imputation of missing values by the EMMCMC method. As a result, we can say that methods which take the spatial features of data into account may improve the performance of imputations. Moreover, EMMCMC method can also handle the missing data that may exist in the reference series. Nevertheless, it should be kept in mind that the number of reference stations as well as their selection process is the key to the success of imputation methods. To improve the performance of the EMMCMC method further, some other modifications are going to be considered in the following studies. The methods studied will also be evaluated in different other stations of Türkiye by using several other comparison criteria.

Acknowledgements
We would like to thank here to our research project group members for their valuable contributions and comments on this study.

References

[1] Coulibaly P. (2007). Comparison of neural network methods for in filling missing daily weather records. *Journal of Hydrology.* Vol. **341**, pp. 27-41.

[2] Göktürk O.M., Bozkurt D., Şen Ö.L., Karaca M. (2008). Quality Control and Homogeneity of Turkish Precipitation Data. *Hydrological Processes.*. Vol. **22**, pp. 3210-3218.

[3] Ramos-Calzado P., Gomez-Camacho J., Perez-Bernal F., F. Pita-Lopez M. (2008). A novel approach to precipitation series completion in climatological datasets: Application to Andalusia. *International Journal of Climatology.* Vol. **28**, pp. 1525-1534.

[4] Schafer J.L. (1997). *Analysis of Incomplete Multivariate Data.* Chapman and Hall/CRC Press, London.

[5] Schneider T. (2001). Analysis of incomplete climate data: Estimation of mean values and covariance matrices and imputation of missing values. *Journal of Climate.* Vol. **14**, pp. 853-871.

[6] WMO. (1983). *Guide to Climatological Practices., 2nd edn.Guide to Climatological Practices., 2nd edn.* Word Meteorological Organization: WMO no:100, Secretariat of the World Meteorological Organization: Geneva.

[7] Xia Y., Fabian P., Stohl A., Winterhalter M. (1999). Forest climatology: Estimation of missing values for Bavaria. *Germany Agricultural and Forest Meteorology.* Vol. **96 (1-3)**, pp. 131-144.

[8] Young K.C. (1992). A three-way model for interpolating for monthly precipitation values. *Monthly Weather Review.* Vol. **120**, pp. 2562-2569.

[9] Yucel R.M., He Y., Zaslavsky A.M. (2008). Using calibration to improve rounding in imputation. *American Statistician.*. Vol. **62 (2)**, pp. 125-129.