Estimation of missing data in a geophysical series of precipitation

Journal of Physics: Conference Series **1938** (2021) 012024 doi:10.1088/1742-6596/1938/1/012024

Estimation of missing data in a geophysical series of precipitation

H J Gallardo Pérez[1], **M Vergel Ortega**[1], **and J P Rojas Suárez**[1]
[1] Universidad Francisco de Paula Santander, San José de Cúcuta, Colombia

E-mail: henrygallardo@ufps.edu.co

Abstract. The analysis of dynamic systems is a topic of great interest in the basic sciences since it allows direct inference of the behavior of different systems. The study of physical phenomena provides large databases that, if recorded at regular time intervals, constitute time series. However, time series of geophysical data in many cases present missing data and their estimation requires the application of valid methods that allow estimating reliable information to complete the series since some analysis methods require these series to be complete. Two methods are used in this article to estimate the missing values of the precipitation series in the city of San José de Cúcuta, Colombia, the first one consists of considering the univariate data series and applying an adjustment of the sequential conditional expectation method of forecasting with restrictions, the second one refers to analyze the data series of a nearby station and through multivariate methods establish the cointegration between the series, and then use this as a basis for estimating the missing data in the analysis series. The two methods are recursive, a first estimation of the model is made ignoring the missing data, an initial estimation of the missing data is made, then a new estimation of the model parameters and a new estimation of the missing data is made, the algorithm continues running with the new values replacing the values estimated in the previous phase until the difference of the estimated values between successive iterations is less than a value fixed beforehand. Finally, a comparison is made between the estimates made by the two methods.

1. Introduction

Rainfall is one of the main variables in hydrogeological studies, since it is the fundamental source for the calculation of water balances and the generation of early warnings for drought risk in a region. In order to analyze precipitation, it is necessary to have a continuous, homogeneous database that covers the maximum possible time interval [1], since it is important to have a complete source of information to be successful in the modeling of any physical phenomenon.

The research was conducted with information from 2000 to 2020, in the city of San José de Cúcuta, capital of the department of Norte de Santander, located in northeastern Colombia, on the border with Venezuela, has a tropical climate characterized by short, very hot and cloudy summers and very short, hot and mostly cloudy winters, its temperature varies between 22 and 33 degrees Celsius, the most common type of precipitation during the year is only rain.

The weather station located at the Camilo Daza Airport in San José de Cúcuta, Colombia, records daily information on temperature, precipitation, hours of sunshine, wind speed and direction, among others, which it consolidates and reports on a monthly basis. However, for different external reasons, there are months in which the information is not recorded; these are the missing data in the associated time series. Stochastic events are unrepeatable in time and, therefore, it is not possible to obtain the

missing information in time series from the observation of the phenomenon again; however, given the importance of having complete series of physical phenomena, it is necessary to estimate the vector of missing data of the series through processes that yield accurate and reliable results from the periodic records that have been made of the series, in such a way that continuity is given to the observations through results adjusted to the reality of the phenomenon under study.

This paper shows two alternatives to estimate the vector of missing observations in the monthly rainfall series for the city of San José de Cúcuta, Colombia, using two different alternative methods. The first one consists of analyzing the univariate rainfall series in San José de Cúcuta, Colombia, and estimating the corresponding model from the observed data of the series, once the estimates of the model parameters are obtained, the vector of missing data is estimated and based on this complete information, a new estimate of the parameters of the underlying model is made and the process continues iteratively until the difference between the values of the vector estimated at each step is less than a previously established value. The second method is based on associating information from a data series that is cointegrated with the first one. In this case, the monthly rainfall series recorded at the Palonegro airport station in the city of Bucaramanga, Colombia, is used and, once the model that relates them is established, the missing values of the first series are estimated based on the observations recorded in the second series.

2. Materials and methods

The research is framed within the quantitative paradigm and uses deductive reasoning that allows the analysis of time series, the application of methods for estimating the parameters of the model underlying them, the identification of cointegration of time series and the estimation of the vector of missing data in a time series. The values of the realizations of the random variables at each time instant are exogenous to the work performed and come from secondary sources, since the researcher has no possibility to modify them, only to observe or record them. In this work two applications are made for the estimation of missing data of the same series, one using univariate series methods and the other using multivariate series methods.

2.1. Time series

The study of time series in the explanation of dynamic phenomena, whether economic, environmental or business, allows, among other things, to identify the factors that cause variability in the time series, to describe regular and non-regular fluctuations, their characteristics and the processes and phenomena that cause them [2].

A time series is defined as an ordered succession of realizations of random variables observed at different regular periods of time, from a mathematical statistical point of view, it consists of realizations of a stochastic process in discrete time: $\{Z_t\}_T = \{Z_1, Z_2, Z_3, \dots, Z_t, \dots\}$. Each component of the vector is a random variable and in each time period only one realization of the variable is obtained. The observed value of the vector of random variables in period t is denoted Z_t; in each time period only one realization of the random variables is observed. It is assumed that there is equispacing between observations and that these correspond to discrete points in time, so the collected data correspond to finite successions of realizations of stochastic variables [3,4]. Then, a time series consists of a sequence of ordered and chronologically equidistant observations on a characteristic (univariate series) or on several characteristics (multivariate series) of an observable unit at different points in time [5].

There are different methods for time series analysis [6], however, in this research we will work with the methodology of Box and Jenkins [7] which proposes the Autoregressive Moving Average Integrated model, ARIMA (Equation (1)) where B is the usual lag operator, $\theta(B)$ is the moving average polynomial, $\varphi(B)$ is the autoregressive polynomial having no common factors with $\theta(B)$, and $\delta(B)$ is the difference polynomial (inducing stationarity) with its roots above the unit circle; this model explains the value of the series as a function of the combination of two polynomials: the autoregressive polynomial, AR, and

the moving average, MA, (Equation (2)); estimation of the ARIMA model parameters is performed based on the exploration of the autocorrelation and partial autocorrelation functions that are obtained once adjusted to a stationary form through simple or seasonal differentiations to stabilize the mean and exponential or logarithmic transformations for variance stabilization [8].

$$Z_t = \frac{\theta(B)}{\delta(B)\phi(B)} a_t, \quad (1)$$

$$Z_t = c\mu + \emptyset_1 Z_{t-1} + \emptyset_2 Z_{t-2} + \dots + \emptyset_p Z_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} - \dots - \theta_q a_{t-q}. \quad (2)$$

The purpose of time series analysis is to develop a statistical model that adequately describes the series in such a way that the theoretical implications of the model are compatible with the sampling patterns observed in the time series [9]. For the estimation of the model, it is assumed that the time series is a particular finite realization of a stochastic process. Once the parameters have been estimated and the model validated, it is used to describe the behavior of the series over time, forecast its future evolution and test theories about the characteristics or variables that make up the series by means of empirical evidence [7].

2.2. Missing data estimation

Missing data estimation methods seek to establish a good approximation of the vector of unobserved observations of the series. In the time series literature, methods have been proposed to deal with missing data, outliers and interventions on the series [10]. Other methods start from the assumption that the model is known or that there is a subset of observations that allow the identification of its structure and from there the vector of missing observations is estimated through an iterative procedure [11-15], multivariate series methods are also used [16-18] in which, after establishing cointegration between series, linear regression models are used to estimate the vector of missing observations in a series from the other related ones.

3. Results

Figure 1 presents the information corresponding to rainfall in the city of San José de Cúcuta, Colombia, during the years 2000 to 2020, according to the records obtained at the Camilo Daza airport station in San José de Cúcuta, Colombia, [19]. It is observed that there are missing data for the years 2004, 2005, and 2008.

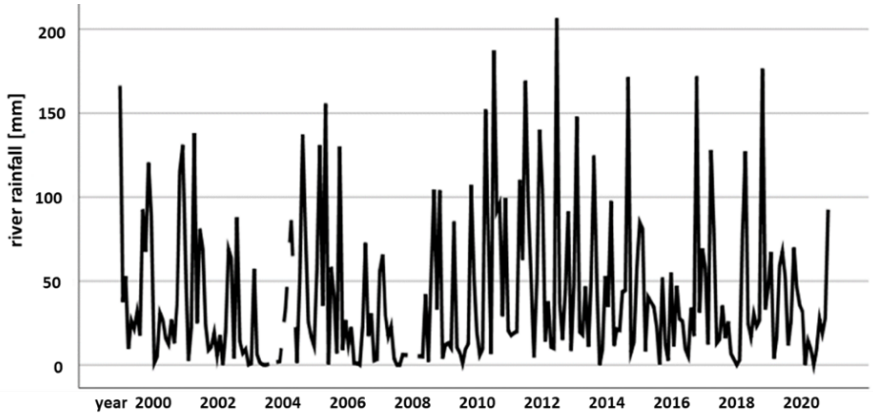

Figure 1. Monthly rainfall in San José de Cúcuta, Colombia.

Using the methodology of Box and Jenkins [7], the parameters of the ARIMA model are estimated using the information from the years 2000 to 2003 and then using the information obtained during the

years 2009 to 2020; the estimated models are similar; the two subseries are integrated type of order 1, I(1), that is, they require a differentiation of degree one to achieve stationarity and, in both, the autocorrelation function suggests the estimation of an AR(1) model. The estimated model with information from 2000 to 2003 is: $\widehat{W}_t = -1.958 - 0.329\widehat{W}_{t-1}$ and after performing the inverse integration to the lag operator we obtain: $\hat{Z}_t = -1.958 + 0.671\hat{Z}_{t-1} + 0.329\hat{Z}_{t-2}$; and the estimated model with the 2009 to 2020 information is: $\widehat{W}_t = 0.395 - 0.424\widehat{W}_{t-1}$, which after performing order one integration leads to the estimated model: $\hat{Z}_t = 0.395 + 0.576\hat{Z}_{t-1} + 0.424\hat{Z}_{t-2}$. The initial estimation of the vector of missing observations is similar for both models, then the first adjustment is made, and, after two iterations, the estimated series presented in Figure 2 is obtained. The general model estimated for the data series from 2000 to 2020, including the estimated data is: $\widehat{W}_t = -0.220 - 0.434\widehat{W}_{t-1}$, which leads to model AR(1): AR(1): $\hat{Z}_t = -0.220 + 0.566\hat{Z}_{t-1} + 0.434\hat{Z}_{t-2}$

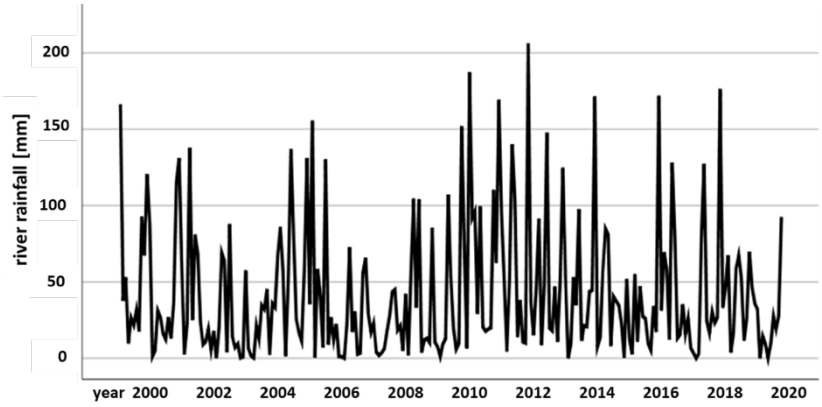

Figure 2. Monthly rainfall in San José de Cúcuta, Colombia. Estimated missing data.

In order to use multivariate models, the series of rainfall observations recorded at the Palonegro airport station in the city of Bucaramanga, Colombia, [19] is used. The two series are I(1) and when fitting the linear regression model that relates them, the residuals comply with the unit root test, then it can be deduced that they are cointegrated [2]. Figure 3 shows the two series. Note that even though the two cities differ in height above sea level by about a thousand meters and are located on either side of a branch of the eastern cordillera, their variational behavior over time is quite similar; in Bucaramanga, Colombia, the AR(1) model fitted for the monthly rainfall series is: $\widehat{W}_t = -0.668 - 0{,}353\widehat{W}_{t-1}$, leading to the model: $\hat{Z}_t = -0.668 + 0.647\hat{Z}_{t-1} + 0.353\hat{Z}_{t-2}$

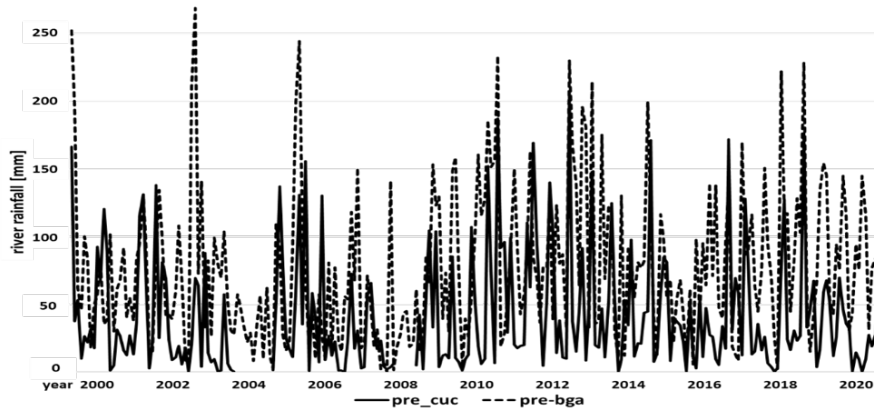

Figure 3. Monthly rainfall precipitation in San José de Cúcuta and Bucaramanga, Colombia.

Since the two series are cointegrated, a linear regression model can be estimated that allows estimating the values of the monthly rainfall data for the city of San José de Cúcuta, Colombia, as a function of those of the Bucaramanga, Colombia, series. The estimated model is: $\text{pre\_cuc} = 19.72 + 0.29 * \text{pre\_buc}$. Using this statistical model, the vector of missing observations for the series of the city of Cúcuta is estimated, the results of which are presented in Figure 4. It is important to note that when comparing Figure 2 and Figure 4, both methods allow obtaining similar estimates of the vector of missing data of a time series.

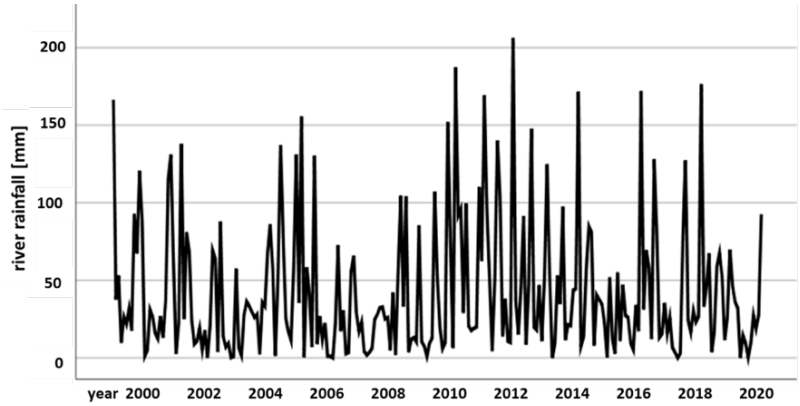

Figure 4. Monthly rainfall in San José de Cúcuta, Colombia. Missing data estimated by multivariate method.

4. Discussion

The estimation of missing data in time series has been approached from different methods since it constitutes a fundamental basis for the reconstruction of unrecoverable information over time. This paper presents two estimation options for the rainfall series in the city of San José de Cúcuta, Colombia. Both methods lead to estimates that are in accordance with other research in this area [1,10] and allow reaching similar results in the estimates, which is a good indicator of the consistency of both procedures.

To compare the missing data vectors estimated by both methods, paired differences are established and two procedures are performed. One consists of estimating the coefficient of variation, the arithmetic ratio between the standard deviation and the arithmetic mean, finding a percentage variability of 4.7%, which indicates high homogeneity in the data. The second is performed by means of a Kolmogorov Smirnov test for comparison of distributions, which yields a significance level of 0.31% that allows rejecting the hypothesis of significant differences between the two vectors of observations.

This work is important and novel because its results provide reliable and complete information on one of the variables used in the explanation of geophysical phenomena and complement research conducted in the area [1,10,18]. It also allows establishing that the use of univariate models in the estimation of missing data in time series is statistically equivalent to the multivariate method used, which generates simplicity in the use of resources to estimate missing data in series related to physical phenomena.

5. Conclusions

Stochastic variables in time produce different phenomena that can be studied from physics and statistics. The values observed in a time series are unrepeatable and if the value is not taken or an error is made in its measurement, there is no way to go back in time to repeat the observation; this leads, if it is the case, to have a vector of missing data in a time series. The methods illustrated here allow the estimation, with high reliability, of the vector of missing observations in a time series. In the second case, it is necessary to determine that the series are cointegrated, otherwise the estimation would lose validity.

The application presented in this paper illustrates the possibility of estimating, by valid and reliable methods, the missing data in a time series of rainfall recorded at a meteorological station. It is possible to extend the second method to the estimation of missing data using a larger number of time series constructed from information from other meteorological stations; in this case, it was not possible to extend it since there are no meteorological stations located in nearby sites with similar climatological conditions.

The main purpose of this work was the application of statistical methods of univariate and bivariate time series to the estimation of missing data in a climatological time series; however, the work did not include the intervention of external phenomena, such as the Nino and Nina phenomena; but this is the subject of another phase of the research, since, in this case, we must work with models based on the presence of a random perturbation whose effect on the structure of the series may be permanent or temporary. The relevance of the work is given by the fact that the methodology developed allows the estimation of mathematical models for the explanation of the behavior of stochastic variables that vary in time from empirical data, which is very useful to obtain complete time series from the estimation, with high reliability, of the vector of missing data.

References

[1] Herrera-Oliva C, Campos-Gaytán J, Carrillo-González F 2017 Estimación de datos faltantes de precipitación por el método de regresión lineal: Caso de estudio Cuenca Guadalupe, Baja California, México *Investigación y Ciencia* **25(71)** 33
[2] Gallardo H, Vergel M, Rojas J 2020 Análisis dinámico de series multivariadas *Mundo FESC* **10(20)** 34
[3] Gallardo H, Rojas J, Gallardo O 2019 *Modelación de Series Temporales en el Sector Productivo del Norte de Santander* (Bogotá: ECOE)
[4] Gallardo H, Vergel M, Rojas J 2020 Dynamic and sequential update for time series forecasting *Journal of Physics: Conference Series* **1587(1)** 012016:1
[5] Mauricio J 2007 *Introducción al Análisis de Series Temporales* (Madrid: Universidad Complutense de Madrid)
[6] Abril J 2011 Análisis de la evolución de las técnicas de series tiempo. Un enfoque unificado *Estadística* **63(181)** 5
[7] Box G, Jenkins G 1969 *Time Series Analysis, Forecasting and Control* (San Francisco: Holden–Day)
[8] Gallardo H, Gallardo O, Rojas J 2019 Estimation of models and cycles in time series applying fractal geometry *Journal of Physics: Conference Series* **1329(1)** 012018:1
[9] Guerrero V 1989 Optimal conditional ARIMA forecasts *Journal of Forecasting* **8** 215
[10] Medina-Rivera R, Montoya-Restrepo E, Jaramillo-Robledo A 2008 Estimación estadística de valores faltantes en series históricas de lluvia *Cenicafé* **59(3)** 260
[11] Box G, Tiao G 1975 Intervention analysis with applications to economic and environmental problems *Journal of the American Statistical Association* **70** 335
[12] Chow G, Lin A 1976 Best linear unbiased estimation of missing observation in a economic time series *Journal of the American Statistical Association* **71** 719
[13] Anderson B, Moore B 1979 *Optimal Filtering* (Englewood: Prentice-Hall)
[14] Jones R 1980 Maximum likelihood fitting of ARMA Models to time series with missing observations *Technometrics* **22** 389
[15] Peña D, Maravall A 1990 Interpolation, outliers and the inverse autocorrelations *Communications in Statistics* **20(10)** 3175
[16] Velásquez M, Martínez J 2009 Estimación de observaciones faltantes en series de tiempo usando métodos multivariados con restricciones *Comunicaciones en Estadística* **2(1)** 1
[17] Guerrero V, Peña D 2003 Combining multiple time series predictors: a useful inferential procedure *Journal of Statistical Planning and Inference* **116** 249
[18] Alfaro E, Javier F 2009 Descripción de dos métodos de rellenado de datos ausentes en series de tiempo meteorológicas *Revista de Matemáticas: Teoría y Aplicaciones* **16(1)** 59
[19] WeatherOnline Ltd 2020 *WeatherOnline Ltd. - Meteorological Services* (London: WeatherOnline Ltd) Consulted on: www.woespana.es