








## Methodological approaches for imputing missing data into monthly flows series

ARTICLES doi:10.4136/ambi-agua.2795

Received: 15 Sep. 2021; Accepted: 07 Mar. 2022

Michel Trarbach Bleidorn<sup>1</sup>; Wanderson de Paula Pinto<sup>2</sup>  
Isamara Maria Schmidt<sup>1\*</sup>; Antonio Sergio Ferreira Mendonça<sup>1</sup>  
José Antonio Tosta dos Reis<sup>1</sup>

<sup>1</sup>Departamento de Engenharia Ambiental. Universidade Federal do Espírito Santo (UFES), Avenida Fernando Ferrari, n° 514, CEP: 29075-910, Vitória, ES, Brazil. E-mail: michelbleidorn@gmail.com, anserfm@terra.com.br, jatreis@gmail.com

<sup>2</sup>Núcleo Integrado de Pesquisa em Engenharia Ambiental. Faculdade da Região Serrana (FARESE), Rua Jequitibá, n° 121, CEP: 29645-000, Santa Maria de Jetibá, ES, Brazil. E-mail: wandersondpp@gmail.com

\*Corresponding author. E-mail: isamaraschmidt@gmail.com

### ABSTRACT

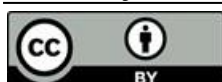
Missing data is one of the main difficulties in working with fluvimetric records. Database gaps may result from fluvimetric stations components problems, monitoring interruptions and lack of observers. Incomplete series analysis generates uncertain results, negatively impacting water resources management. Thus, proper missing data consideration is very important to ensure better information quality. This work aims to analyze, comparatively, missing data imputation methodologies in monthly river-flow time series, considering, as a case study, the Doce River, located in Southeast Brazil. Missing data were simulated in 5%, 10%, 15%, 25% and 40% proportions following a random distribution pattern, ignoring the missing data generation mechanisms. Ten missing data imputation methodologies were used: arithmetic mean, median, simple and multiple linear regression, regional weighting, spline and Stineman interpolation, Kalman smoothing, multiple imputation and maximum likelihood. Their performances were compared through bias, root mean square error, absolute mean percentage error, determination coefficient and concordance index. Results indicate that for 5% missing data, any methodology for imputing can be considered, recommending caution for arithmetic mean method application. However, as the missing data proportion increases, it is recommended to use multiple imputation and maximum likelihood methodologies when there are support stations for imputation, and the Stineman interpolation and Kalman Smoothing methods when only the studied series is available.

**Keywords:** Doce river, imputation, missing data.

### Abordagens metodológicas para imputação de dados faltantes de vazões médias mensais

### RESUMO

A falta de dados é uma das principais dificuldades no trabalho com registros fluviométricos. As lacunas no banco de dados podem resultar de problemas nos componentes das estações fluviométricas, interrupções no monitoramento e falha dos observadores. A análise



This is an Open Access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

de séries incompletas gera resultados incertos, impactando negativamente a gestão dos recursos hídricos. Assim, a consideração adequada dos dados faltantes é muito importante para garantir a qualidade de informação. Este trabalho teve como objetivo analisar, comparativamente, metodologias de imputação de dados faltantes em séries temporais de vazões fluviais mensais, considerando, em um estudo de caso, o Rio Doce, localizado no Sudeste do Brasil. Os dados faltantes foram simulados nas proporções de 5%, 10%, 15%, 25% e 40% seguindo um padrão de distribuição aleatória e ignorando os mecanismos de geração de falhas. Foram utilizadas dez metodologias de imputação de dados faltantes: média aritmética, mediana, regressão linear simples e múltipla, ponderação regional, interpolação spline e Stineman, suavização de Kalman, imputação múltipla e máxima verossimilhança. Seus desempenhos foram comparados por meio dos indicadores viés, raiz do erro quadrático médio, erro absoluto médio percentual, coeficiente de determinação e índice de concordância. Os resultados indicam que para 5% de dados faltantes, qualquer metodologia de imputação pode ser considerada, recomendando cautela na aplicação da média aritmética. No entanto, à medida que a proporção de dados faltantes aumenta, recomenda-se o uso das metodologias imputação múltipla e máxima verossimilhança quando houver estações de suporte para imputação, e os métodos de interpolação Stineman e suavização de Kalman quando apenas as séries estudadas estiverem disponíveis.

**Palavras-chave:** dados faltantes, imputação, Rio Doce.

## 1. INTRODUCTION

Knowledge about the water regime in a river basin is fundamental in hydrological studies, being an indispensable factor for adequate water-resource management (Moreira, 2006). In this sense, Brazilian Law 9,433 (Brasil, 1997), which instituted the National Water Resources Policy, indicates, among the management instruments, the National Water Resources Information System. In this context, the hydrometeorological monitoring network maintained by the National Water Agency (ANA) and the availability of the databases generated in the Hydrological Information System (HIDROWEB) become relevant.

The instrument National Water Resources Information System, however, is still incipient in relation to the others, mainly due to the limited number of monitoring stations and the incompleteness of the data generated (Fioreze and Oliveira, 2010). The existence of gaps in the historical series is due to technical or maintenance problems, non-ideal climatic conditions, instrumental failures or device errors during data collection, human error during data entry, calibration processes and/or data damage due to malfunction of machines storage, construction and organization of hydrometric databases (Gao, 2017; Johnston, 1999; Peña-Angulo *et al.*, 2019; Tencaliec, 2017; Tucci, 1997).

According to McKnight *et al.* (2007), “in general, the term missing data means that some type of information about the phenomenon in which we are interested is missing” and, therefore, the sample is called incomplete. The analysis of incomplete flow-time series produces negative impacts, especially on stochastic decompositions of series, compromising information such as trend, stationarity, cycle and seasonality (Box and Cox, 1964).

As discussed by Roth *et al.* (1999) and Pigott (2001), the existence of missing values in time series generally decreases the capacity and precision of statistical analysis approaches and contributes to biased estimates of the relationship between variables, which may cause inaccurate assumptions in data set exploration which can negatively impact water resources management, for example, in determination of maximum permissible uptake and ecological flows, extreme flows estimation, flows forecasting, hydraulic systems designs, among others. Due to this fact, the reconstruction of incomplete series and the treatment of missing data must be seen as a priority in the data preparation procedure (Hamzah *et al.*, 2020).

Because missing data imputation is a useful tool in water-resource management studies (Barnette and Kobiyama, 2006), several authors have worked on the application of techniques for imputing missing data in hydrological studies resulting in a variety of methods ranging from simple imputation by mean or median to widely used statistical methods such as Regional Weighting (Ely *et al.*, 2021); interpolations (linear, quadratic and cubic) (Gyau-Boakye and Schultz, 1994, Hamzah *et al.*, 2020); methods based on linear regressions (single and multiple) (Kamwaga *et al.*, 2018; Khalifeloo *et al.*, 2015); Self Organizing Map (SOM) and Soil and Water Assessment Tool (SWAT) (Kim *et al.*, 2015); to more advanced and robust methods, such as different Artificial Neural Network approaches (Canchala-Nastar *et al.*, 2019; Elshorbagy *et al.*, 2000; Nkiaka *et al.*, 2016; Starrett *et al.*, 2010; Vega-Garcia *et al.*, 2019); machine learning methods (Heras and Matovelle, 2021; Rado *et al.*, 2019); satellite radar altimetry and multiple imputation (Ekeu-Wei *et al.*, 2018); combination of regression and autoregressive integrated moving average (ARIMA) models called dynamic regression (Tencaliec *et al.*, 2015); Singular Spectrum Analysis (SSA) and Multichannel Singular Spectrum Analysis (MSSA) (Semiromi and Koch, 2019); among many others. The many methods that can be used for hydrological missing data imputation resulted in literature reviews as can be seen in Ben Aissia *et al.* (2017) and Hamzah *et al.* (2020).

Ventura *et al.* (2016) carried out a study to compare statistical methods for filling gaps and to verify which method presents better results for meteorological data series. Three weather stations located in Porto Alegre, Rio de Janeiro and Manaus cities, in Brazil, were chosen. Failures were simulated in real data series and the performances of four methods were compared: simple average, moving average, simple linear regression and multiple linear regression. To verify the obtained results, the mean absolute error and the correlation coefficient were used. The results showed excellent performance of the multiple linear regression method for the variables temperature, humidity and dew point, while the simple average had the best result for the variable atmospheric pressure. None of the four methods presented good results for the variable solar radiation.

Nunes *et al.* (2009) carried out a study with the objective of publicizing the Multiple Imputation (MI) method. The authors selected a 470 surgical patient death outcome data set and adjusted logistic models to it. Two incomplete data sets were generated, one presenting 5% and the other 20% of missing data for the variable albumin. Models were adjusted to the complete series, to the series presenting missing data and the series filled by using MI. The estimates obtained by the analysis of the series presenting missing data and with the filled series were different, mainly for those presenting 20% of missing data. The utilized MI was efficient, because the results achieved with the series filled by imputations were close to those obtained with the complete series. The results obtained considering series filled by using MI were superior to those obtained for series with missing data.

Junger and Leon (2015) presented an imputation method via Maximum Likelihood (ML) that is suitable for multivariate time series using the EM algorithm (Expectation and Maximization) under the assumption of normal distribution. The authors used a database related with tem  $PM_{10}$  monitoring stations located in São Paulo city, Brazil. Different approaches were considered to filter the temporal component. A simulation study was carried out to compare the proposed and some frequently used methods of quality and performance. The simulations showed that when the amount of missing data was less than 5%, the complete data analysis generated satisfactory results, regardless of the mechanism that generated the missing data. Imputation quality began to degenerate when missing data proportions exceeded 10%. The proposed imputation method presented good accuracy and precision in different configurations with respect to the missing observation patterns. Most imputations obtained valid results, even under the non-random losses mechanism.

Sattari *et al.* (2017) evaluated different methods of imputing missing data in monthly

rainfall time series collected at six stations in southern Iran. Imputation methodologies analyzed include arithmetic mean, inverse distance interpolation, linear regression, multiple imputations, multiple linear regression analysis, non-linear iterative partial least squares algorithm, NR method, single best estimator, UK traditional method and M5 decision model tree. Results showed that arithmetic averaging method, multiple linear regression method and nonlinear iterative partial least squares algorithm perform best. Multiple regression methods provided a successful missing precipitation data estimation. Multiple imputation methods produced the most accurate results for precipitation data from five dependent stations. Finally, the decision-tree algorithm is explicit, and therefore it is used when insights into decision making are needed.

Chen *et al.* (2019) verified the impact of using different methods for imputing missing data in rainfall series on the forecasting hydrological and non-point (H/NPS) pollution performance using the Soil and Water Assessment Tool (SWAT) model. Multiple imputation (MI) and maximum likelihood methods using expectation-maximization bootstrap algorithm (EMB) were considered. Different imputed data sets effects were investigated through a case study in the Daning River Basin, Three Gorges Reservoir Region, China. Results indicate that rainfall data imputation and H/NPS model performance obtained by EMB algorithm are superior to MI performance. Authors highlight the important implications for choosing appropriate imputation methods in H/NPS models to solve data scarcity problems for watershed studies.

Hamzah *et al.* (2022) evaluated the performance of multiple imputations by chained equations (MICE) approach to predicting recurrence in streamflow datasets. To evaluate and verify MICE approach effectiveness in treating missing streamflow data, complete historical daily streamflow series from 2012 to 2014 were used. Later, MICE methods coupled with multiple linear regression (MLR) were applied to restore streamflow rates in Malaysia's Langat River Basin from 1978 to 2016. The best estimation methods are validated with tests such as adjusted R-squared ( $Adj R^2$ ), residual standard error (RSE) and mean absolute percentage error (MAPE). Findings revealed that the classification and regression tree (CART) method combined with MLR outperformed the other approaches tested, with highest  $Adj R^2$  value and lowest RSE and MAPE values observed regardless of missing conditions.

Abu Romman *et al.* (2021) compared ten imputation methods that were used to impute rainfall depth data in an arid region of the Mediterranean. Series mean, linear interpolation, linear trend, arithmetic mean, normal ratio, inverse distance weighting, linear regression with GPCC data, linear regression with satellite data, stepwise multiple linear regression and multiple imputation were used for these imputations. The results showed that for intervals between 5 and 20% of failures, the stepwise multiple linear regression method produced best results with a root mean square error (RMSE) and mean absolute error (MAE) less than 7 and 2 mm, respectively. This was followed by the Monte Carlo Markov chain expectation-maximization-based multiple imputation method, which had an RSME and MAE of 1.01 and 0.08 mm, respectively, when the series had 20% failures. On the other hand, satellite data use for imputation was adequate for failures between 10 and 15%.

Other studies can be highlighted, especially related to medicine and health (Camargos *et al.*, 2011; Carreras *et al.*, 2021; Khan *et al.*, 2021; Nunes, 2007; Payrovnaziri *et al.*, 2021), air pollution (Choi *et al.*, 2021; Ghazali *et al.*, 2021; Pinto, 2013), engineering, mainly civil and traffic (Abdelgawad *et al.*, 2015; Jiang *et al.*, 2021), meteorology (Afrifa-Yamoah *et al.*, 2020; Bier and Ferraz, 2017; Costa *et al.*, 2021; Ferrari and Ozaki, 2014; García-Peña *et al.*, 2014), agriculture (Jiao *et al.*, 2016; Nishina *et al.*, 2017; Swenson, 2014), energy (Barbosa *et al.*, 2018; Pelisson, 2021) and education (Vinha and Laros, 2018).

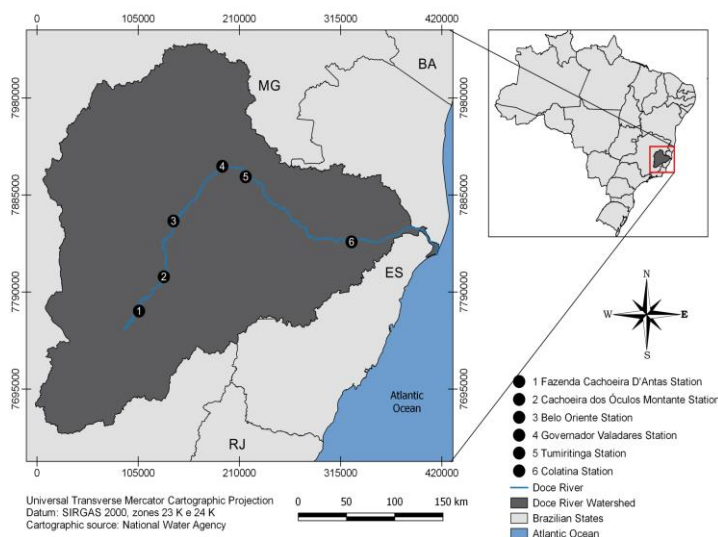
Currently, it is unclear in the literature which method is the most appropriate to deal with missing value imputation in river-flow time series. Considering this scenario, this research evaluates ten imputation techniques, based on single and multiple imputation methods: Arithmetic Mean (AM), Median (M), Simple Linear Regression (SLR), Multiple Linear

Regression (MLR), Regional Weighting (RW), Spline Interpolation (SPLINE), Stineman interpolation (STINE), Kalman Smoothing with (KALMAN), Multiple Imputation (MI) and Maximum Likelihood (ML). It is important to emphasize that there are still few studies that use MI and ML imputation methodologies dealing with missing value imputation in river-flow time series, evidencing the need to develop works analyzing their performance, and these results can help sustainable water resource management. In this context, the present research objective is to comparatively analyze methodologies for imputing missing data by an application to Doce River, Brazil, monthly average fluviometric flow-time series.

## 2. MATERIAL AND METHODS

### 2.1. Study Area

The Doce River watershed, Figure 1, is located in Southeast Brazil, occupying portions of Minas Gerais and Espírito Santo states between the parallels 17°45' and 21°15' South latitude and the meridians 39°55' and 43°45' West longitude. The Doce River presents 853 km total extension and 83,465 km<sup>2</sup> drainage area (Coelho, 2007). Of this area, 86% belongs to Minas Gerais state and the remaining 14% to Espírito Santo state, being, therefore, a federal dominion watershed.



**Figure 1.** Doce River Watershed.

### 2.2. Data

Six ANA hydrometeorological network Doce River fluviometric stations were considered in this work. The monthly flow data was obtained from the HIDROWEB Hydrological Information System (Agência Nacional de Águas, 2018). Station-selection criteria were related to the existence of consistent historical average monthly flow-time series for at least 20 years, according to literature recommendations (Longhi and Formiga, 2011; Sarmiento, 2007).

Historical series stationarity analysis, that verifies mean and variance identity of two distinct hydrological series subperiods, utilized Fisher's  $F$  and  $t$ -Student tests, in order to evaluate time series hydrological regime behavior changes, which can be caused by several factors, such as reservoir construction upstream of the fluviometric station, water withdrawal for irrigation agricultural activities use and even local climate regime changes over time (Sousa *et al.*, 2009). For the analysis, the software SisCAH 1.0 - Computational System for Hydrological Analysis - was used (Sousa *et al.*, 2009).

The selected stations are shown in Figure 1. The list of stations considered is presented in Table 1, with respective ANA identification codes, identification nomenclature adopted in this

study, geographic coordinates and drainage areas. Colatina station is located in Espírito Santo state, and the others in Minas Gerais state. After analyzing the available data, the study base period 1987 to 2014 was determined.

**Table 1.** Selected fluviometric stations information.

Code	ID	Station	X (m)	Y (m)	Zone	Drainage Area (km <sup>2</sup> )
56425000	E1	Fazenda Cachoeira D'Antas	743325	7766305	23 K	10,100
56539000	E2	Cachoeira dos Óculos - Montante	764419	7762315	23 K	15,900
56719998	E3	Belo Oriente	775691	7760371	23 K	24,200
56850000	E4	Governador Valadares	189099	7753279	24 K	40,500
56920000	E5	Tumiritinga	221838	7752450	24 K	55,100
56994500	E6	Colatina	329007	7749346	24 K	76,400

### 2.3. Missing data patterns

According to Silva (2012), the presence of missing data in a database can be characterized by observed failure behavior patterns which is of paramount importance to describe missing value locations in the series. First, Collins *et al.* (1991) described and divided the pattern of missingness into two groups: *general* (random) and *special* patterns. Special patterns including univariate missing data, unit nonresponse, and monotone missing data. “Missing general” or “random pattern” is where missing data occur in any of the variables in any position. In a “special pattern” case, if there is only one variable with missing data while the other variables are completely recorded, the pattern is called “univariate missing data”. Additionally, when the multivariate pattern is detected, means that the missing values arise in more than one variable; if there are missing values on a block of variables for the same set of cases, and the remaining of the variables are all complete, the missing data pattern is called “unit nonresponse”; and, the pattern is said to be “monotone” whenever observations are ordered and item  $k$  is missing, and all  $k + 1, \dots, n$  cases are also missing (Collins *et al.*, 1991; Silva, 2012). For the present study, the general pattern will be considered, because it assumes that the missing data follow a random distribution.

### 2.4. Missing data mechanisms

Little and Rubin (2019) classified three possible ways that data may go missing: Missing Completely at Random (MCAR), Missing at Random (MAR) and Missing Not at Random (MNAR). As discussed by Hamzah *et al.* (2020) and Gao *et al.*, (2018), MCAR describes data where the gaps are distinct from any of the variables in the dataset. In any event, the missing values probably correlated to other observed values, yet not to missing values; in that case, the missingness is assumed to be MAR. Missing data which are dependent on the observed value is known as MNAR. Missing data can be presented in the form of a probabilistic process that describes the association among the measured variables and the probability of missing value. For more details about Missing Data Mechanisms, see Little and Rubin (2019); based on these authors, missing value in streamflow study is determined as MCAR because the missingness episode in streamflow data of an area is not influenced by data in that area or any area. This mechanism of missing data is classified as ignorable, that is, there isn't need to model the mechanism as part of the estimation process (Allison, 2001).

### 2.5. Missing data imputation methods

The missing data imputation methodologies can be classified in three general classes: Single Imputation (SI), Multiple Imputation (MI) and Estimation. The basic principle of SI methods is to impute one value to each database missing data and, then, analyze it as if there

weren't missing data (McKnight *et al.*, 2007). The MI is characterized by being a method that consists of imputing a certain missing data by a data set and, after analyzing it, determining the best value to be taken to impute it. The Estimation method consists in estimating parameters that govern the missing values distribution from the observed data. For the present study, the Maximum Likelihood (ML) methodology will be employed. In addition to this general classification and more widely used in the literature to deal with missing data, imputation methodologies can also be classified as univariate or multivariate. The first occurs when the series itself provides information that can be used by imputation methodology. The second one is when it is necessary to use support stations to impute interest series. In this study, AM, M, SPLINE, STINE and KALMAN were classified as univariate imputation methods. In turn, SLR, MLR, RW, MI and ML are multivariate methods.

### 2.5.1. Single Imputation Methods

#### Arithmetic Mean (AM)

This is the most commonly used single imputation technique where the missing values are replaced with the mean value of the time series. The mean of a series of values  $x_1, x_2, \dots, x_n$  is given by Equation 1.

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad (1)$$

#### Median (M)

Median imputation is another simple method often appropriate for highly skewed data (Junger and Leon, 2015). This method calculates the median of the variable based on all cases that have data for any variable and replaces the series missing values with the median of the variable (Kabir *et al.*, 2019). The median of a series of values  $x_1, x_2, \dots, x_n$  is can be obtained by Equation 2 after sorting the dataset in ascending order:

$$m(x) = \begin{cases} x_{\left(\frac{n+1}{2}\right)} & \text{if } n \text{ an even number} \\ \frac{x_{\left(\frac{n}{2}\right)} + x_{\left(\frac{n}{2}+1\right)}}{2} & \text{if } n \text{ an odd number} \end{cases} \quad (2)$$

#### Simple (SLR) and Multiple Linear Regression (MLR)

Imputation with linear regression uses information from complete series to fill the missing values of the interest series. It can be of two types: (i) imputation by simple linear regression, when the purpose is to predict the dependent series behavior as a function of only one independent series (Equation 3) or (ii) imputation by multiple linear regression, when the dependent series behavior is a function of more than one independent series (Equation 4).

$$X = \beta_0 + \beta Y + \varepsilon \quad (3)$$

In Equation 3, X represents the linear regression equation dependent series;  $\beta_0$  a linear coefficient vector;  $\beta$  the angular coefficient; Y the independent station and  $\varepsilon$  the model residuals. As in SLR case, only one independent series is used to adjust the regression, Pearson correlation coefficient was used as a criterion for choosing this series and, after the test was applied, the station time series that presented highest correlation coefficient with E6, which is the interest station, was then chosen to perform imputation.

$$X = \beta_0 + \sum_{i=1}^n \beta_i Y_i + \varepsilon \quad (4)$$

In Equation 4, X represents the linear regression equation dependent series;  $\beta_0$  a linear coefficient vector;  $\beta_i$  the angular coefficients;  $Y_i$  the independent station and  $\varepsilon$  the model

residuals. For MLR, stations E1, E2, E3, E4 and E5 series were used as independent series to fit the regression with E6.

### Regional Weighting (RW)

Imputation by regional weighting method establishes linear regressions between the station that has series with missing data,  $X$ , and each neighboring station series,  $Y_1, Y_2, \dots, Y_n$ , incorporating distance information (Mello *et al.*, 2017; Pruski *et al.*, 2004). From each linear regression performed, a correlation coefficient is obtained for the data to be estimated. Equation 5 denotes the regional weighting method.

$$X = \frac{1}{n} \sum_{i=1}^n \frac{N_X}{N_i} D_i \quad (5)$$

In Equation 5,  $N_X$  and  $N_i$  represents the monthly average flows data for the station with missing data to be imputed and the order “ $i$ ” neighboring station monthly average flow, respectively ( $\text{m}^3 \cdot \text{s}^{-1}$ );  $D_i$  denotes the values observed in the order “ $i$ ” neighboring stations during the month of occurrence in the station with the data to be imputed ( $\text{m}^3 \cdot \text{s}^{-1}$ ); and  $n$  is the number of neighboring stations considered.

### Spline Interpolation (SPLINE)

According to Wijesekara and Liyanage (2020), for  $n + 1$  pair of observations  $\{(t_i, x_i): i = 0, 1, \dots, n\}$ , the shape of spline is modeled by interpolating between all the pairs of observations  $(t_{i-1}, x_{i-1})$  and  $(t_i, x_i)$  with polynomials described in Equation 6.

$$x = q_i(t), i = 1, 2, \dots, n \quad (6)$$

### Stineman interpolation (STINE)

This is an advanced interpolation method where interpolation occurs based on (i) whether values of the specified points ordinates change monotonically and (ii) the slopes of the line segments joining specified points change monotonically (Turicchi *et al.*, 2020).

### Kalman Smoothing (KALMAN)

The Kalman filter calculates the mean and variance of the unobserved state, given the observations. This filter is a recursive algorithm; the current best estimate is updated whenever a new observation is obtained. Kalman Smoothing takes the form of a backwards recursion and it can be used to compute a smoothed estimator of the disturbance vector (Wijesekara and Liyanage, 2020).

In the present work, R package *imputeTS* was used for Spline Interpolation, Stineman Interpolation and Kalman smoothing imputations.

### 2.5.2. Multiple Imputation

In an attempt to develop a method that could reflect uncertainty over missing data imputations, Rubin (1987) created the MI method, in which each missing value is replaced by a set of plausible values representing this uncertainty about the value to be imputed. MI consists of the following three steps: (i)  $m$  complete databases  $Y_{\text{obs}}, Y_{\text{mis}}$  are obtained through appropriate imputation techniques; (ii) separately,  $m$  data banks are analyzed by a traditional statistical method, as if they were indeed complete data groups; (iii) the  $m$  results found in step (ii) are combined in a simple and suitable way for obtaining the so-called repeated imputation inference.

In this research, the imputation model used (i) is adjusted by the Bayesian Paradigm: from the result of the posterior distribution, a set of random extractions is made for the missing data from the observed data, thus obtaining the complete database, that is, multiple imputations are



made using the linear regression method ( $Y = \alpha + \beta X$ ),  $Y \sim N(X\beta; I\sigma^2)$ , where the response variable  $Y$  will be the variable to be imputed for which the parameters are estimated from its own posterior distribution. The predicted values for  $Y_{obs}$  and  $Y_{mis}$  are calculated and, for each predicted  $Y_{mis}$ , the observed unit with the closest predicted value is sought using it as the value to be imputed. The variability between imputations is generated through the steps used to estimate  $\beta$  and  $\sigma$  and which are repeated  $m$  times; in the next step (ii)  $Q$  of each imputed data set is estimated; finally, in step (iii), the  $m$  results obtained can be combined using the rules proposed by Rubin (1987). The idea is that from each analysis the estimates for the parameter of interest  $Q$  are obtained, that is,  $Q_i$  for  $i = 1, 2, 3, \dots, m$ . According to Schafer (1999),  $Q$  can be any scalar measure to be estimated, such as mean, correlation, regression coefficient or odds ratio. Then, the combined estimate will be the average of the individual estimates  $\bar{Q} = \frac{1}{m} \sum_{i=1}^m \hat{Q}_i$ . For the combined variance, the variance is first calculated within the imputations  $\bar{U} = \frac{1}{m} \sum_{i=1}^m U_i$  and the variance between imputations  $B = \frac{1}{m-1} \sum_{i=1}^m (\hat{Q}_i - \bar{Q})^2$ . Then, the total variance, which is the combined variance, will be  $T = \bar{U} + \left(1 + \frac{1}{m}\right) B$ .

The literature standardizes  $m = 5$  imputations, a value that, as indicated by Nunes (2007) emerged from the researcher's experiences. MI is implemented in the *mice* R software library, developed by Buuren and Groothuis-Oudshoorn (2010).

### 2.5.3. Maximum Likelihood

The ML method's basic principle is to choose as an estimation of parameters those values that, if true, would maximize the probability of observing what was actually observed (Allison, 2001). The parameters estimation is performed by maximizing the likelihood function, which, in most cases, can't be performed analytically. It is necessary to employ numerical methods, such as the EM algorithm. This maximization method is very popular when the data considered in the estimation aren't complete (Allison, 2001).

The EM algorithm is an iterative procedure consisting of two steps repetition: Estimation ( $E$ ) and maximization ( $M$ ). The process begins with mean vector and covariance matrix estimation by using only the complete data. According to Junger and Leon (2015), being  $\mathbf{x}_t$ , ( $t = 1, \dots, m$ ), the  $t^{th}$  random vector  $\mathbf{X}$  realization, with multivariate normal distribution and  $m$  components not observed. The vector  $\mathbf{x}_t$  can be arranged in such a way that the  $m$  missing components are placed in the first positions, that is,  $\mathbf{x}_t = (x_{t1}, \dots, x_{tm}, x_{t(m+1)}, \dots, x_{tp})^T$ , and represented as  $\mathbf{x}_t = (\mathbf{x}_{t1}, \mathbf{x}_{t2})^T$ . Consider  $B$  windows with different covariance regimes over time. The estimation of the mean vector at the instant  $t$  and window  $b$ , ( $b = 1, \dots, B$ ), can be partitioned following the same configuration as that corresponding to  $\mathbf{x}_t$  components, according to equations 7 and 8.

$$\tilde{\mu}_t = \begin{bmatrix} \tilde{\mu}_{t1} \\ \tilde{\mu}_{t2} \end{bmatrix} \quad (7)$$

$$\Sigma_b = \begin{bmatrix} \tilde{\Sigma}_{b11} & \tilde{\Sigma}_{b12} \\ \tilde{\Sigma}_{b21} & \tilde{\Sigma}_{b22} \end{bmatrix} \quad (8)$$

The imputation algorithm consists of (i) replace the missing values for estimates values; (ii) estimate the normal model parameters  $\mu$  and  $\Sigma$  and each univariate time series  $\mu_t$  level; (iii) re-estimate the missing values considering just the updated parameters and each time series level. This process is repeated until the estimated values stop to vary (Junger and Leon, 2015). Junger and Leon (2015) report that initial estimates  $\tilde{\mu}_0$  and  $\tilde{\Sigma}_0$  are, respectively, the mean vector and the sample covariance matrix, considering only the observed data. In iteration  $(k + 1)$  of

the step E for the EM algorithm, the missing values are imputed as the conditional mean to the observed values and the parameters estimated in the previous interaction appropriated using equation (9), with the contributions to covariance estimated by Equations 10 and 11.

$$\tilde{\mathbf{x}}_{t1}^{(k+1)} = E \left[ \mathbf{X}_{t1} \mid \mathbf{x}_{t2}, \tilde{\mu}_t^{(k)}, \tilde{\Sigma}_b^{(k)} \right] = \tilde{\mu}_{t1}^{(k)} + \tilde{\Sigma}_{b12}^{(k)} \tilde{\Sigma}_{b22}^{(k)-1} (\mathbf{x}_{t2} - \tilde{\mu}_{t2}^{(k)}) \quad (9)$$

$$\widetilde{\mathbf{x}}_{t1} \mathbf{x}_{t1}^T \quad (k+1) = E \left[ \mathbf{X}_{t1} \mathbf{X}_{t1}^T \mid \mathbf{x}_{t2}, \tilde{\mu}_t^{(k)}, \tilde{\Sigma}_b^{(k)} \right] = \tilde{\Sigma}_{b11}^{(k)} - \tilde{\Sigma}_{b12}^{(k)} \tilde{\Sigma}_{b22}^{(k)-1} \tilde{\Sigma}_{b21}^{(k)} + \tilde{\mathbf{x}}_{t1} \tilde{\mathbf{x}}_{t1}^T \quad (10)$$

$$\widetilde{\mathbf{x}}_{t1} \mathbf{x}_{t2}^T \quad (k+1) = E \left[ \mathbf{X}_{t1} \mathbf{X}_{t2}^T \mid \mathbf{x}_{t2}, \tilde{\mu}_t^{(k)}, \tilde{\Sigma}_b^{(k)} \right] = \tilde{\mathbf{x}}_{t1} \tilde{\mathbf{x}}_{t2}^T \quad (11)$$

According to Junger and Leon (2015), in step  $M$  the  $\mu_b$  and  $\Sigma_b$  reviewed maximum likelihood estimates are computed, considering implicit the interaction index  $(k + 1)$ ,  $\tilde{\mu}_b = \sum_{t=1}^{n_b} \frac{\tilde{x}_{bt}}{n_b}$  and  $\tilde{\Sigma}_b = \sum_{t=1}^{n_b} \frac{\tilde{x}_{bt} \tilde{x}_{bt}^T}{n_b} - \tilde{\mu}_b \tilde{\mu}_b^T$ . The  $\tilde{\mu}_b$  estimate is used only for  $\tilde{\Sigma}_b$  calculation.

Junger and Leon (2015) emphasize the need to use additional models to estimate the contribution of the time component for each univariate series, that is,  $\mu_t$  value. The temporal trajectory of the series considered in this study was modeled with use of cubic non-parametric splines because, according to the same authors, this trajectory was the one that presented the best performance in relation to the regression models and *ARIMA* (Autoregressive Integrated Moving Average) models for air pollution time series missing data imputation.

Therefore, consider that  $\mu_t$  can be estimated by a smooth function  $g_j$  with  $j = 1, \dots, p$ . The curve  $g_j$ , in turn, is estimated in such a way that the functional  $S(g_j) = \sum_{k=1}^K [X_t - g(v_k)]^2 + \lambda \int_a^b (g'')^2 dx$  be minimized. The points  $v_1, v_2, \dots, v_k$ , ordered in the interval  $[a, b]$ , are the nodes and  $\lambda$  is the curve smoothing parameter (Junger and Leon, 2015). This results in a natural cubic spline (Green and Silverman, 1993). Each variable  $X_j$  has its level given by  $\mu_{jt} = g(X_{jt})$ .

In this paper, the EM algorithm proposed by Junger (2008) was used, which makes use of the *mtsdi* platform in the R software, which was implemented by the author.

## 2.6. Performance Indexes

For the performance evaluation of different missing data, imputation methods in effecting the estimates were employed as can be seen below (Equations 12, 13, 14, 15 and 16).

BIAS

$$\frac{1}{N} \sum_{i=1}^N (x_i - \tilde{x}_i) \quad (12)$$

Root Mean Square Error (RMSE)

$$\frac{1}{N} \sqrt{\sum_{i=1}^N (x_i - \tilde{x}_i)^2} \quad (13)$$

Mean Absolute Percentage Error (MAPE)

$$\frac{1}{N} \sum_{i=1}^N \left| \frac{x_i - \tilde{x}_i}{x_i} \right| \times 100 \quad (14)$$

Coefficient of Determination ( $R^2$ )

$$\frac{1}{N} \sum_{i=1}^N \frac{\sum_{i=1}^N [(x_i - \bar{x})(\tilde{x}_i - \bar{\tilde{x}})]}{\hat{\sigma}(x_i) \hat{\sigma}(\tilde{x})} \quad (15)$$

Concordance Index ( $d_2$ )

$$1 - \left[ \frac{\sum_{i=1}^N (x_i - \tilde{x}_i)^2}{\sum_{i=1}^N (|x_i - \bar{x}| + |\tilde{x}_i - \bar{x}|)^2} \right] \quad (16)$$

The BIAS measure quantifies underestimation and overestimation estimates with respect to the average observations (Bier and Ferraz, 2017; Junger, 2008); RMSE and MAPE are accuracy estimates measures. In equations 12-16,  $N$  represents the number of missing data in the modeled data set,  $x_i$  the observed data,  $\tilde{x}_i$  the imputed data,  $i = 1, \dots, m$ ,  $\bar{x}$  the observed values mean and  $\bar{\tilde{x}}$  the imputed data mean (Pinto, 2013).

## 2.7. Application

From the Doce River monthly average flow-time series, an algorithm was created to simulate five incomplete data banks, with 5%, 10%, 15%, 25% and 40% missing data. The routines for simulating the missing data percentages, imputations and analyzes were implemented by using the R software (R Development Core Team, 2018). The website for that software access is <http://www.r-project.org>.

## 3. RESULTS AND DICUSSION

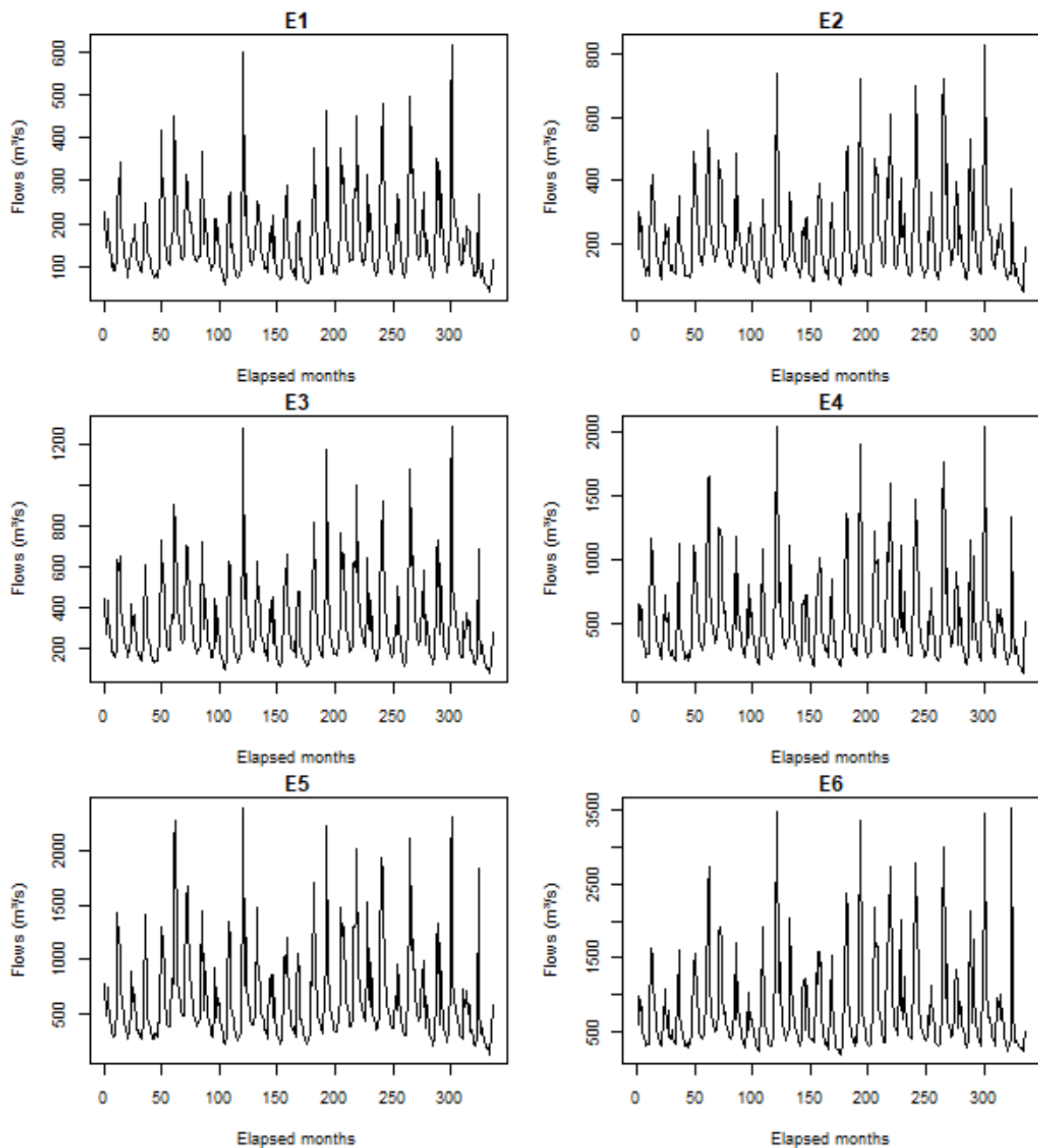
For a better understanding of the studied average monthly flow variable, each station's descriptive statistics are presented in Table 2. It's observed that, during the study period, the monthly average flow rates ranged from  $162.48 \text{ m}^3 \cdot \text{s}^{-1}$  in E1 to  $797.77 \text{ m}^3 \cdot \text{s}^{-1}$  in E6. These values represent the lowest and highest average monthly flows, considering all stations. A high standard deviation and coefficient of variation values can be observed, an aspect that allows us to conclude that the means aren't representative, a fact that, according to Bayer *et al.* (2012), can be associated with the large data intra-annual variability, indicating a seasonal component. It is also noted that the observed extreme monthly average flow values, maximum and minimum, were  $3,528.32$  and  $41.58 \text{ m}^3 \cdot \text{s}^{-1}$ , at Stations E6 and E1, respectively. The first was registered on January 12, 2013, while the second was observed on October 1, 2014.

The fluviometric stations' monthly average flow distributions present positive asymmetry values ranging from 1.79 to 2.04. For kurtosis, the variation was from 3.24 to 4.63. The mean values are greater than the respective median and mode values, and the data distribution values' degree of concentration is classified as leptokurtic, according to Fonseca and Martins (2011).

Figure 2 shows the six stations' average monthly flow-time evolution. As can be observed, the time series present intra-annual variability pattern, with floods periods followed by drought periods, which characterizes seasonality property, already indicated by the descriptive statistics and confirmed in Pinto *et al.* (2015) studies for E6 monthly mean flows and by Bleidorn *et al.* (2018) studies, for minimum monthly flows for all stations considered in this study. It's relevant to note that both studies suggest a seasonal component with a 12-month period, with a flood period from November to May, and drought period from June to October.

**Table 2.** Average monthly flow variable descriptive measures.

Descriptive measures	E1	E2	E3	E4	E5	E6
Mean ( $\text{m}^3 \cdot \text{s}^{-1}$ )	162.48	214.60	325.18	524.07	646.85	797.77
Median ( $\text{m}^3 \cdot \text{s}^{-1}$ )	132.74	171.45	256.43	395.25	499.78	554.74
Standard Deviation ( $\text{m}^3 \cdot \text{s}^{-1}$ )	92.34	129.15	204.40	345.43	418.09	604.33
Coefficient of Variance (%)	56.83	60.18	62.85	65.91	64.63	75.75
Minimum Value ( $\text{m}^3 \cdot \text{s}^{-1}$ )	41.58	52.49	81.63	112.20	131.03	194.20
Maximum Value ( $\text{m}^3 \cdot \text{s}^{-1}$ )	616.22	827.27	1,219.04	2,051.50	2,391.05	3,528.32
Kurtosis	4.03	3.58	3.80	3.50	3.24	4.63
Asymmetry	1.79	1.74	1.77	1.78	1.77	2.04



**Figure 2.** Fluviometric stations average monthly flow time series.

According to the Fischer  $F$  and  $t$ -Student tests, the time series under study are homogeneous, that is, they do not present variances and mean changes over time, considering five-year analysis periods. This fact is important to validate studied stations' data consistency. The imputation methods considered in this work require that the data follow a Normal distribution; however, according to Shapiro and Wilk (1965) and Bera and Jarque (1981) tests, the data normality null hypothesis is rejected. According to Junger (2008) and Sabino *et al.* (2014), environmental data, commonly, do not follow Normal distribution. Thus, statistical tests are performed to check the application of a transformation that could stabilize the variance of the observations. In the case under study, the smoothing parameter  $\lambda$  was estimated, as suggested by Box and Cox (1964) to define the type of transformation. According to Reisen *et al.* (2008), often the transformation not only stabilizes the variance but also improves the data distribution approximation to the Normal distribution. Thus, all imputations were performed using the natural logarithm transformation in order to improve the approximation with the

Normal distribution and to stabilize the variance. The imputed data was back transformed for subsequent analysis.

Linear Pearson correlation coefficients for the average monthly flows  $r$  between stations under study indicate a highly homogeneous distribution, because among the pairs of stations the lowest value was 0.9240, as shown in Table 3. This condition was expected, because the stations E1 and E6 present the highest distance between stations identified in the study area, which suggests that better performance can be achieved by using the multivariate imputation methods (SLR, MLR, RW, MI and ML).

**Table 3.** Average monthly flow data: Pearson correlations between stations.

	E1	E2	E3	E4	E5	E6
E1	1					
E2	0.9875	1				
E3	0.9797	0.9869	1			
E4	0.9604	0.9698	0.9897	1		
E5	0.9419	0.9524	0.9762	0.9928	1	
E6	0.9240	0.9368	0.9616	0.9782	0.9794	1

In order to validate the missing data imputation methods considered in this study, the E6 database underwent 1000 replications for each failure ratio. Table 4 shows the performance-indicator means for the missing data imputation methods. In general, there is a decreasing imputation quality gradient presented by the performance indexes (PI) because of missing data increase. The SI (AM, M, SLR, MLR and RW) methods showed considerably low BIAS values and high RMSE and MAPE values in relation to SI (SPLINE, STINE e KALMAN) and mainly to multivariate attribution (MI and ML) methods. SI (AM, SLR, MLR and RW) methods showed a considerable increase in BIAS when missing data proportion increased, making imputed series underestimated in relation to observed one. M showed a similar behavior, however, with lower intensity. The increase in bias values was observed to a lesser extent for SI methods (SPLINE, STINE and KALMAN) and mainly for MI and ML, suggesting a small imputed data variance loss in relation to observed ones. In addition, the methods that lost most quality for  $R^2$  and  $d_2$  indicators were SI (AM, M, SLR, MLR and RW) followed by SI (SPLINE, STINE and KALMAN) and with little loss for MI and ML. This confirms MI and ML methods good performance, being that the last corroborates with the efficiency proven by Junger (2008) and reforced by Burger *et al.* (2018).

Both AM and M methodologies are central tendency measures. However, the second is a better alternative for variables that do not follow a Normal distribution, that is, the median better represents the central tendency of a distribution that presents large deviations from the Normal distribution (McKnight *et al.*, 2007; Pinto, 2013). Therefore, for variables that do not follow a Normal distribution, such as the river average flow rate, under study, the imputation methodology by AM is not efficient, especially for faults above 5% even under natural logarithm transformation. This conclusion was also found in Pinto (2013), where the detrimental effect of this method on the imputation of the variable  $PM_{10}$  (Inhalable Particulate Matter with aerodynamic diameter less than  $10 \mu m$ ) is reduced only in samples with a small percentage of missing data (5%). The efficiency of the imputation methodology by M decreases slightly for faults above 10%. Hence, this method is not indicated for these situations.

**Table 4.** Performance indicators for missing data imputation methodologies.

Faults	PI	AM	M	SPLINE	STINE	KALMAN	SLR	MLR	RW	M1	ML
5%	BIAS	0.0169	0.0098	0.0001	0.0004	0.0001	0.0187	0.0214	0.0169	0.0001	<0.0001
	RMSE	0.0084	0.0077	0.0045	0.0039	0.0039	0.0049	0.0057	0.0043	0.0020	<0.0001
	MAPE	0.3870	0.3725	0.1939	0.1649	0.1684	0.2877	0.3305	0.2598	0.0999	0.0737
	R	0.9668	0.9719	0.9905	0.9927	0.9927	0.9889	0.9850	0.9913	0.9982	0.9990
	$d_2$	0.9829	0.9856	0.9952	0.9963	0.9963	0.9941	0.9920	0.9954	0.9991	0.9995
10%	BIAS	0.0667	0.0221	0.0001	0.0007	0.0001	0.0703	0.0762	0.0671	0.0002	<0.0001
	RMSE	0.0154	0.0111	0.0066	0.0057	0.0057	0.0124	0.0138	0.0117	0.0028	<0.0001
	MAPE	1.0273	0.7502	0.4083	0.3397	0.3473	1.0846	1.1748	1.0350	0.2010	0.1489
	R	0.8974	0.9425	0.9800	0.9851	0.9850	0.9404	0.9273	0.9469	0.9963	0.9979
	$d_2$	0.9452	0.9698	0.9898	0.9924	0.9923	0.9653	0.9577	0.9693	0.9982	0.9989
15%	BIAS	0.1504	0.0413	0.0002	0.0014	0.0003	0.1540	0.1623	0.1498	0.0003	<0.0001
	RMSE	0.0246	0.0142	0.0085	0.0072	0.0072	0.0220	0.0236	0.0211	0.0035	<0.0001
	MAPE	2.2230	1.1391	0.6391	0.5228	0.5347	2.3717	2.4979	2.3092	0.3054	0.2247
	R	0.7811	0.9080	0.9675	0.9764	0.9763	0.8469	0.8268	0.8575	0.9944	0.9968
	$d_2$	0.8754	0.9505	0.9834	0.9879	0.9878	0.9022	0.8892	0.9087	0.9972	0.9984
25%	BIAS	0.4099	0.0929	0.0004	0.0020	<0.0001	0.4152	0.4284	0.4101	0.0005	<0.0001
	RMSE	0.0475	0.0196	0.0120	0.0099	0.0100	0.0456	0.0477	0.0448	0.0047	<0.0001
	MAPE	6.1715	1.9624	1.1655	0.9337	0.9602	6.4132	6.6123	6.3317	0.5232	0.3852
	R	0.5245	0.8285	0.9357	0.9548	0.9542	0.6210	0.5986	0.6290	0.9900	0.9944
	$d_2$	0.6966	0.9033	0.9668	0.9764	0.9759	0.7204	0.7050	0.7254	0.9949	0.9971
40%	BIAS	1.0423	0.2412	0.0001	0.0032	0.0003	1.0485	1.0652	1.0423	0.0008	0.0002
	RMSE	0.0921	0.0297	0.0174	0.0140	0.0141	0.0909	0.0931	0.0901	0.0070	<0.0001
	MAPE	15.8970	3.8042	2.1929	1.6855	1.7512	16.1982	16.4453	16.098	0.9293	0.6654
	R	0.2729	0.6580	0.8655	0.9087	0.9062	0.3686	0.3539	0.3723	0.9781	0.9885
	$d_2$	0.5526	0.7902	0.9289	0.9511	0.9489	0.5617	0.5557	0.5635	0.9888	0.9941

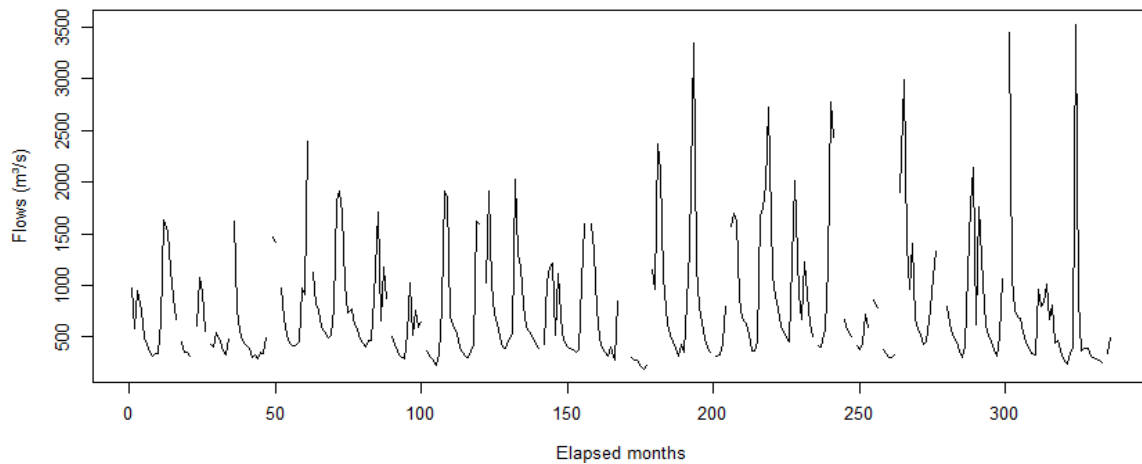
In the hydrology area, the idea of using a single imputation method to recover the missing value (mainly AM and M) is widely used by several authors, as can be seen Ben Aissia *et al.* (2017); Gao *et al.* (2018); Kabir *et al.* (2019); Norliyana *et al.* (2017) and Rahman *et al.* (2017) works. However, despite the method's simplicity the researchers agree that reconstructing the missing value using the same "number" does not reflect the variation that would likely occur if the variables were observed. The real numbers possibly differ from those imputed. Thus, the variation of those same variables is underestimated.

SLR methodology presents better results in relation to MLR according to all performance indicators and regardless of the number of failures. MLR involves more independent variables to predict an adjustment model and, theoretically, its complexity results in a higher efficiency than the corresponding SLR. However, for the SLR methodology, an algorithm was developed in which the independent station was taken as a function of the correlation coefficient  $r$  higher value and, thus, alternate stations were candidates according to the replication progress. For the imputation methodology via MLR, based on the correlation coefficient  $r$  (which assumed values greater than 0.924 between the station pairs), all the other stations (E1, E2, E3, E4 and E5) were considered as adjusted model independent variables, for the dependent variable E6. This justifies the better performance of SLR over MLR in this study. The imputation methodology via RW, which is widely used in hydrological variables series imputation studies, showed superior efficiency than the corresponding to other SI (AM, M, SLR and MLR) methods only for the 5% failure rate and presented less efficiency than those corresponding to SI (SPLINE, STINE and KALMAN) in addition to MI and ML methods, independently of the missing data percentage.

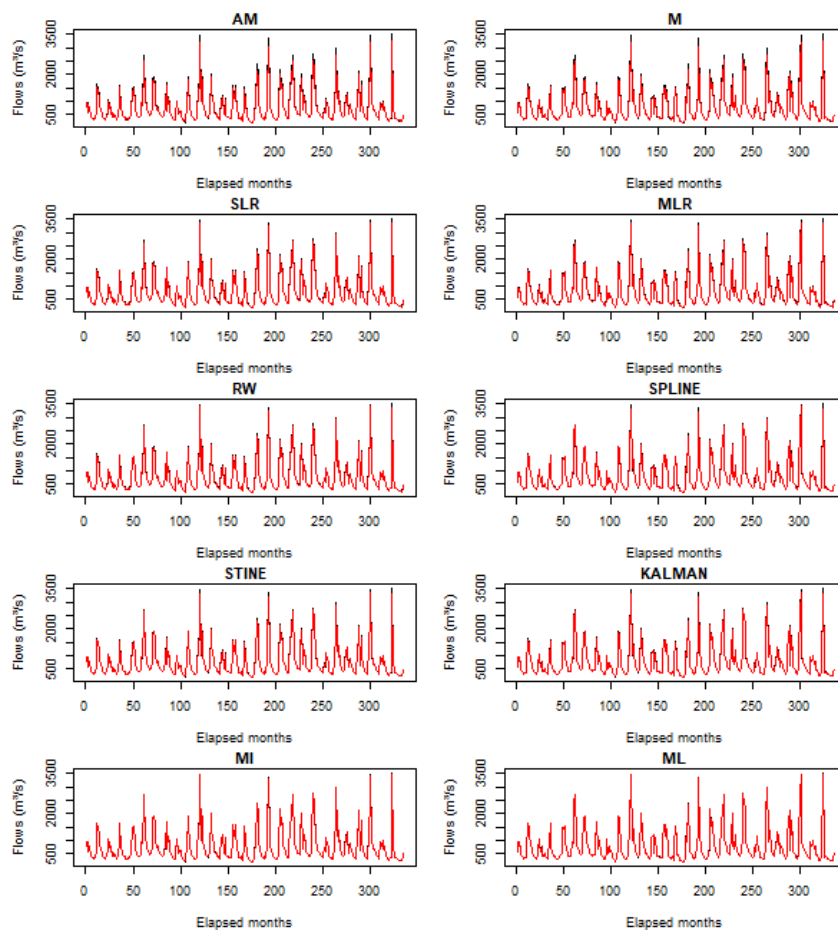
Oliveira *et al.* (2010) concluded that the MLR method showed, according to the performance indexes BIAS, MAE and  $r$ , the best results in annual rainfall faults imputation for six stations located in the state of Goiás, Brazil, followed by vector regional combined with multiple potential regression (MPR), RW, regional vector combined with multiple linear regression, regional weighting based on linear regression, regional vector combined with regional weighting and, lastly, regional vector method. This result corroborates those found by (i) Ventura *et al.* (2016), that, considering BIAS, MAE and  $r$ , verified the superiority of the MLR method for temperature, relative humidity and dew point meteorological variables, considering Manaus, Rio de Janeiro and Porto Alegre Brazilian cities stations series; (ii) those found by Mello *et al.* (2017) that, according to the MAE, for precipitation series corresponding to eight rainfall stations located in Santa Catarina state, Brazil, the MLR method was the one that presented the best performance, followed by RW, regional weighting based on linear regressions and, finally, SLR; and (iii) those found by Bier and Ferraz (2017) who concluded that, for average compensated temperature for meteorological stations located in Rio Grande do Sul state, Brazil, the MLR and the RW methods were the most adequate for missing data estimates, whereas for precipitation there was no method that could be considered best. According to the same authors, this was due to the fact that neighboring stations' precipitation data was less correlated if compared to average temperature compensated data, generating estimates less related to the observed series and presenting larger estimative errors.

Low BIAS, RMSE and MAPE values and high  $R^2$  and  $d_2$  values confirm univariate single imputation methods (SPLINE, STINE and KALMAN) good performance, however, inferior to MI and ML multiple multivariate imputation methods. Additionally, it is possible to verify a slight superiority of imputation quality via ML methodology in relation to MI, regardless of missing data proportion. Figures 3 and 5 show the simulated results for 5 and 40% missing data proportions and Figures 4 and 6 show a visual comparison between observed and imputed data. This comparison shows good performance by all methodologies for the 5% missing data proportion. Therefore, for this smaller data failure percentage, any imputation methodology could be assumed without causing significant changes in the time series characteristics. Thus,

caution is suggested in the use of AM, even for small failure proportions. However, for the 40% proportion, the analysis shows SI methods (SPLINE, STINE and KALMAN) good performance and superiority of MI and ML methods, while for other SI methodologies (AM, M, SLR, MLR and RW) imputed data variability underestimation is notable. Therefore, it can be concluded that the use of such methodologies is not recommended to carry out imputation with high percentages of missing data, as the time series characteristics can be extremely altered.

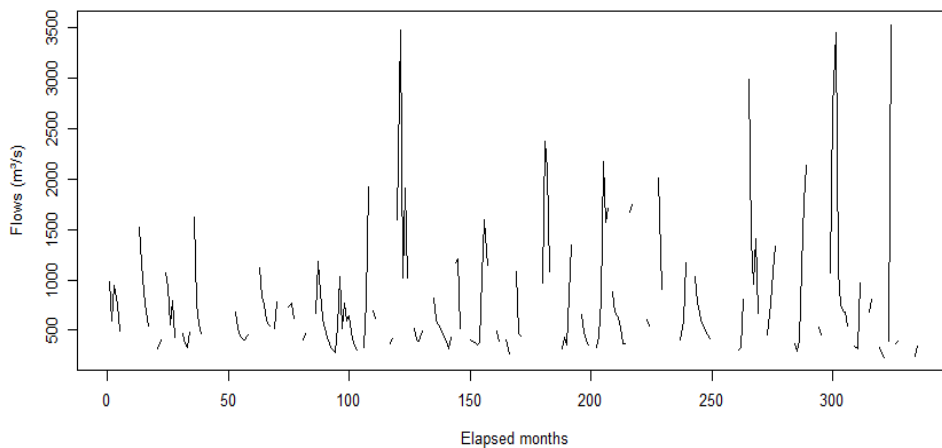


**Figure 3.** Average monthly flow time series for the E6 station, considering 5% missing data.

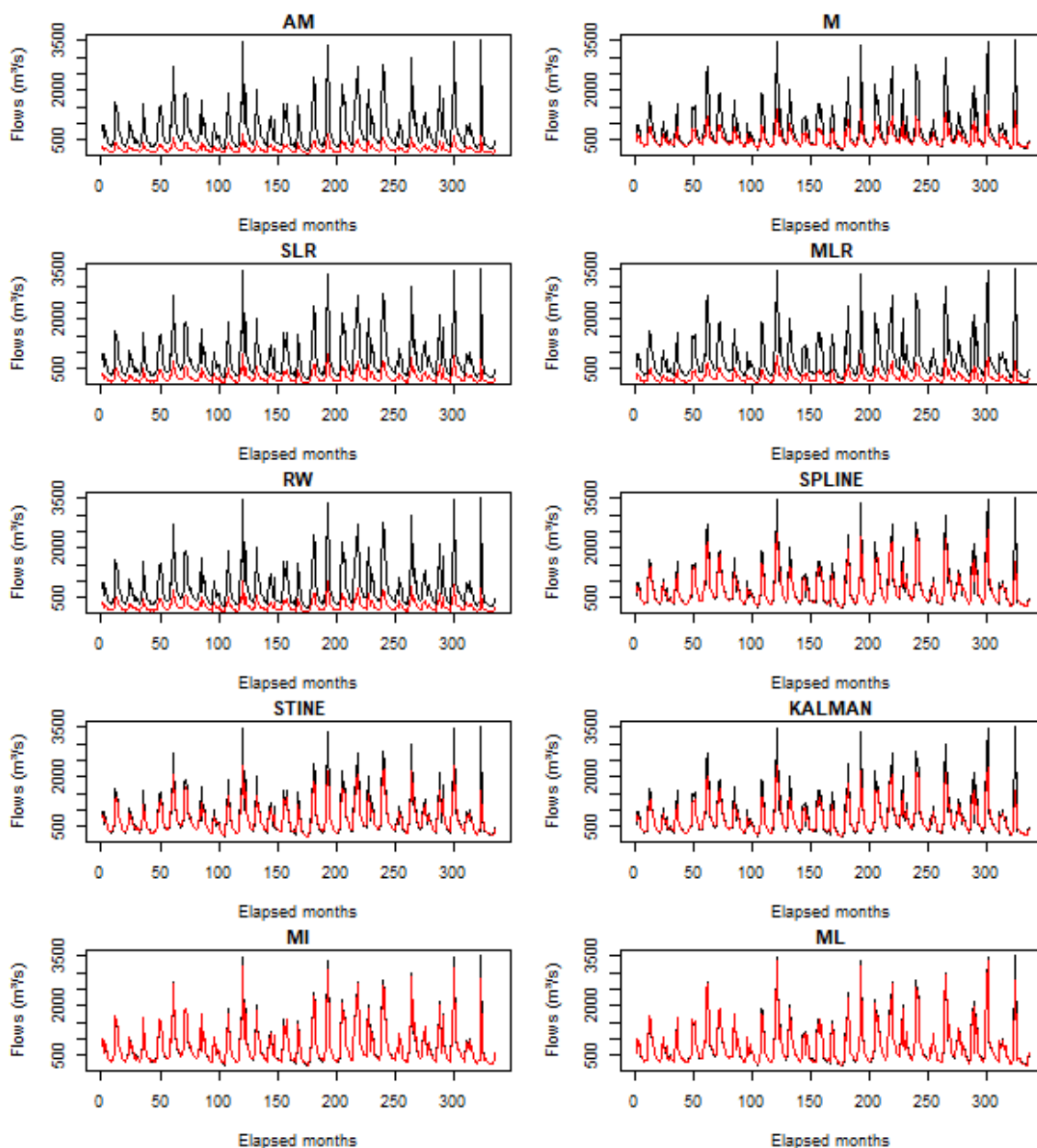


**Figure 4.** Observed (black line) and imputed (red line) average monthly flows time series for the E6 station, for each imputation methodology, considering 5% missing data.





**Figure 5.** Average monthly flow time series for the E6 station, considering 40% missing data.



**Figure 6.** Observed (black line) and imputed (red line) average monthly flows time series for the E6 station, for each imputation methodology, considering 40% missing data.

Table 5 shows relative difference results between the quality indicators used in imputation methods evaluation in this study for Doce River flow series in a scenario where missing data percentage increased from 5% to 40%. As can be seen, all methodologies tested in this study showed a loss of quality with the missing data proportion increase and, based on that, the table was organized in order to order the methodologies compared in this study from the one with greatest to the least loss of quality according to indicators used. From this, we can conclude that, in cases where there are support stations for imputation, the best methodologies for imputing missing data in flow data are the multivariate ones MI and ML, while, in cases where there are no support stations for imputation, the best options of imputation methodologies to be used are the univariate ones Kalman Smoothing and Stineman interpolation.

**Table 5.** Relative difference in the imputation method quality when failures increased from 5% to 40%.

Imputation Methodology	BIAS	RMSE	MAPE	R <sup>2</sup>	d <sub>2</sub>
AM	-98.38%	-90.88%	-97.57%	-71.77%	-43.78%
MLR	-97.99%	-93.88%	-97.99%	-64.07%	-43.98%
SLR	-98.22%	-94.61%	-98.22%	-62.73%	-43.50%
RW	-98.38%	-95.23%	-98.39%	-62.44%	-43.39%
M	-95.94%	-74.07%	-90.21%	-32.30%	-19.83%
SPLINE	0.00%	-74.14%	-91.16%	-12.62%	-6.66%
KALMAN	-66.67%	-72.34%	-90.38%	-8.71%	-4.76%
STINE	-87.50%	-72.14%	-90.22%	-8.46%	-4.54%
MI	-87.50%	-71.43%	-89.25%	-2.01%	-1.03%
ML	-50.00%	0.00%	-88.92%	-1.05%	-0.54%

Based on these results, it can be said that, multivariate MI and ML methodologies application in river flow rates missing data imputation, confirmed their good performance, already proven for imputations considering other variables (Junger, 2008; Pinto, 2013; Nunes *et al.*, 2010; Vinha and Laros, 2018), as occurred in cases of application univariate methodologies KALMAN and STINE whose good performance had also been confirmed to handle missing data in exchange rate data (Burger *et al.*, 2018). Therefore, it is suggested that both methodologies should be taken as quality references for fluviometric variables missing data imputation, especially the ML. Lastly, Table 6 shows the main advantages and disadvantages for each methodology evaluated in the present study.

In this way, fluviometric monitoring and historical series data consistency as well as the adequate missing data treatment are fundamental for water-resource studies and projects, turning possible adequate water-resource use planning, river basin management, flow forecasting, industrial and agricultural public supply, navigation, basic sanitation, water concession and granting, academic studies and water use conflicts resolution.

## 4. CONCLUSIONS

Reliable flow series are essential for studies related to water resources. However, incomplete series can result even from the operation of large monitoring networks with data quality control. This fact indicates the importance of techniques that allow imputation of missing data. However, these techniques have been little used in hydrological studies. In this study, it was verified that, for the Doce River flow series, the multiple imputation and maximum likelihood methodologies, considered as references in the imputation of missing data for series

involving several environmental and social variables, performed better than those most commonly utilized. It was also concluded that the improvement in the results of the imputations in relation to the others increases as the flow series present greater proportions of the missing data.

**Table 6.** Advantages and disadvantages for each methodology evaluated in the present study.

Methodology	Advantages	Disadvantages
AM	Easy to implement	Decreases data variability
M	Easy to implement	Decreases data variability
SLR	Requires additional variables with an acceptable correlation	In situations under and overestimation conditions, the variability of the data decreases
MLR	Requires additional variables with an acceptable correlation	In situations under and overestimation conditions, the variability of the data decreases
RW	Requires additional variables with an acceptable correlation	Requires additional variables that have acceptable correlations. Tends to underestimate data variability
SPLINE	It's a nonlinear approach. Provides a "smooth" interpolant. Doesn't usually get "wiggly" like higher-order polynomial interpolation can	Has a limited ability to predict oscillations from univariate data. Requires a bit more work than linear interpolation to implement
STINE	Produces a imputation known to be robust against sporadic outliers and performs better than spline interpolations, where abrupt changes are observed; Solves the non-monotonic problem of linear and spline interpolation	Is not as smooth method as the linear and spline methods
KALMAN	It avoids the influence of possible structural breaks during the missing values estimation	Assumes a large knowledge of probabilistic theory, specifically Gaussian conditional properties of random variables, which may limit its study and application scope
MI	Incorporates the uncertainty in each database generated, reducing the standard error of the final imputed values	Requires statistical knowledge and computational sophistication
ML	Because is based on available data, extracts information from its behavior	Requires statistical knowledge and computational sophistication

Simulations showed that, for 5% missing data, all the imputation methods present good performances. For this small missing data proportion, statistical efficiency is not compromised. Even so, for this missing data amount, the arithmetic mean methodology, which presents the worse results, should be avoided. The imputation method quality begins to become different for proportions above 10%. The arithmetic mean, simple and multiple linear regression and regional weighting imputation methodologies may be considered limited, even though the last three methods are multivariate and under the condition that the stations showed high

homogeneity among them. Therefore, multivariate methods that consider the single imputation principle were not efficient. On the contrary, univariate methodologies Spline interpolation, Stine interpolation and Kalman Smoothing being single methods showed good results in imputation at high proportions of missing data. Also, multiple imputation and maximum likelihood multivariate methods were shown to be more robust and accurate for missing data treatment of the variable under study, being recommended for failure imputation procedures. For this, it is fundamental that the variables' data present homogeneity among them. Therefore, such methods are recommended for the proper treatment of missing data, which allows us to guarantee quality in imputed time series, thus being able to subsidize the planning and control of water resources.

The method of imputation by multiple linear regression can be improved, since the adjustment of  $n_i < 5$  neighboring stations were based on the correlation coefficient and, according to the replication progress of replication, it is possible that one or more stations do not present a qualified correlation coefficient. As for multiple imputation, it is concluded that its efficiency can be increased as an ideal  $m$  imputation value is established for the variable under study, considering different proportions of missing data. Therefore, it is recommended that future studies be developed because the procedure efficiency is related to the number of  $m$  databases generated and, in general, the higher the percentage of missing values, the greater is  $m$ . It is also recommended that studies be developed regarding the imputation of missing data for river flows in daily and monthly series.

## 5. REFERENCES

- ABDELGAWAD, H.; ABDULAZIM, T.; ABDULHAI, B.; HADAYEGHI, A.; HARRETT, W. Data Imputation and Nested Seasonality Time Series Modelling for Permanent Data Collection Stations: Methodology and Application to Ontario. **Canadian Journal of Civil Engineering**, v. 42, n. 5, p. 287–302, 2015. <https://doi.org/10.1139/cjce-2014-0087>
- ABU ROMMAN, Z.; AL-BAKRI, J.; AL KUISI, M. Comparison of methods for filling in gaps in monthly rainfall series in arid regions. **International Journal of Climatology**, v. 41, n. 15, p. 6674–6689, 2021. <https://doi.org/10.1002/joc.7219>.
- AFRIFA-YAMOA, E.; MUELLER, U. A.; TAYLOR, S. M.; FISHER, A. J. Missing data imputation of high-resolution temporal climate time series data. **Meteorological Applications**, v. 27, n. 1, p. e1873, 2020. <https://doi.org/10.1002/met.1873>.
- AGÊNCIA NACIONAL DE ÁGUAS (Brasil). **HIDROWEB: Sistema de Informações Hidrológicas**. Brasília, 2018. Available at: <http://hidroweb.ana.gov.br/default.asp>. Access in: Apr. 27, 2018.
- ALLISON, P. D. **Missing data**. Thousand Oaks: Sage publications, 2001. (Quantitative Applications in the Social Sciences, 136).
- BARBOSA, M. T. M.; LIMA, A.; KUEHNE, B. T.; BATISTA, B. G.; FILHO, D.M.L.; PEIXOTO, M.L. M. Imputação de dados faltantes no monitoramento de consumo energético residencial em Smart Grids. In: WORKSHOP DE COMPUTAÇÃO URBANA (COURB), 2., 2018, Campos do Jordão. **Anais[...]**. Porto Alegre: Sociedade Brasileira de Computação, 2018.
- BARNETCHE, D.; KOBIYAMA, M. Aplicação do hycymodel no preenchimento de falhas de monitoramento de vazões. **Geosul**, v. 21, n. 42, p.185-194, 2006.

- BAYER, D. M.; CASTRO, N. D. R.; BAYER, F. M. Modelagem e previsão de vazões médias mensais do rio Potiribu utilizando modelos de séries temporais. **RBRH**, v. 17, n. 2, p. 229-239, 2012. <https://doi.org/10.21168/rbrh.v17n2.p229-239>.
- BEN AISSIA, M. A.; CHEBANA, F.; OUARDA, T. B. M. J. Multivariate missing data in hydrology – Review and applications. **Advances in Water Resources**, v. 110, p. 299–309, 2017. <https://doi.org/10.1016/j.advwatres.2017.10.002>.
- BERA, A. K.; JARQUE, C. M. Efficient tests for normality, homoscedasticity and serial independence of regression residuals: Monte Carlo evidence. **Economics Letters**, v. 7, n. 4, p. 313-318, 1981. [https://doi.org/10.1016/0165-1765\(81\)90035-5](https://doi.org/10.1016/0165-1765(81)90035-5).
- BIER, A. A.; FERRAZ, S. E. T. Comparação de metodologias de preenchimento de falhas em dados meteorológicos para estações no sul do Brasil. **Revista Brasileira de Meteorologia**, v. 32, n. 2, p. 215-226, 2017. <https://doi.org/10.1590/0102-77863220008>.
- BLEIDORN, M. T.; SCHMIDT, I. M.; KNAAK, J.; LIMA, G. B.; PINTO, W. P.; BRAUM, E. S. Análise comparativa de vazões ecológicas e disponibilidade hídrica mensal e anual para o rio Doce. In: CONGRESSO FLORESTAL LATINO-AMERICANO, 7., 12-15 Jun. 2018, Vitória. **Anais[...]** Vitória: Programa de Pós-graduação em Ciências Florestais; Universidade Federal do Espírito Santo, 2018. Available at: <https://www.even3.com.br/anais/conflat/95094-analise-comparativa-de-vazoes-ecologicas-e-disponibilidade-hidrica-mensal-e-anual-para-o-rio-doce/>. Access in Oct. 10, 2018.
- BOX, G. E.; COX, D. R. An analysis of transformations. **Journal of the Royal Statistical Society: Series B (Methodological)**, v. 26, n. 2, p. 211-243, 1964. <https://doi.org/10.1111/j.2517-6161.1964.tb00553.x>.
- BRASIL. Presidência da República. Lei n. 9.433, de 9 de janeiro de 1997. Institui a Política Nacional de Recursos Hídricos, cria o Sistema Nacional de Gerenciamento de Recursos Hídricos, regulamenta o inciso XIX do art. 21 da Constituição Federal, e altera o art. 1º da Lei nº 8.001, de 13 de março de 1990, que modificou a Lei nº 7.990, de 28 de dezembro de 1989. **Diário Oficial [da] União**: seção 1, Brasília, DF, 09 jan. 1997.
- BURGER, S.; SILVERMAN, S.; van VUUREN, G. Deriving Correlation Matrices for Missing Financial Time-Series Data. **International Journal of Economics and Finance**, v. 10, n. 10, p. 105-120, 2018. <https://doi.org/10.5539/ijef.v10n10p105>.
- BUUREN, S. V.; GROOTHUIS-OUDSHOORN, K. mice: Multivariate imputation by chained equations in R. **Journal of statistical software**, v. 45, n. 3, p. 1-68, 2010. <https://doi.org/10.18637/jss.v045.i03>.
- CAMARGOS, V. P.; CÉSAR, C. C.; CAIAFFA, W. T.; XAVIER, C. C.; PROIETTI, F. A. Imputação múltipla e análise de casos completos em modelos de regressão logística: uma avaliação prática do impacto das perdas em covariáveis. **Cadernos de Saúde Pública**, v. 27, n. 12, p. 2299-2313, 2011. <https://doi.org/10.1590/S0102-311X2011001200003>.
- CANCHALA-NASTAR, T.; CARVAJAL-ESCOBAR, Y.; ALFONSO-MORALES, W.; CERÓN, W. L.; CAICEDO, E. Estimation of missing data of monthly rainfall in southwestern Colombia using artificial neural networks. **Data in Brief**, v. 26, p. 104517, 2019. <https://doi.org/10.1016/j.dib.2019.104517>.

- CARRERAS, G.; MICCINESI, G.; WILCOCK, A.; PRESTON, N.; NIEBOER, D.; DELIENS, L. *et al.* Missing Not at Random in End of Life Care Studies: Multiple Imputation and Sensitivity Analysis on Data from the ACTION Study. **BMC Medical Research Methodology**, v. 21, n. 13, 2021. <https://doi.org/10.1186/s12874-020-01180-y>.
- CHEN, L.; XU, J.; WANG, G.; SHEN, Z. Comparison of the Multiple Imputation Approaches for Imputing Rainfall Data Series and Their Applications to Watershed Models. **Journal of Hydrology**, v. 572, p. 449–460, 2019. <https://doi.org/10.1016/j.jhydrol.2019.03.025>.
- CHOI, C.; JUNG, H.; CHO, J. An Ensemble Method for Missing Data of Environmental Sensor Considering Univariate and Multivariate Characteristics. **Sensors**, v. 21, n. 22, p. 7595, 2021. <https://doi.org/10.3390/s21227595>.
- COELHO, A. L. N. **Alterações Hidrogeomorfológicas no Médio-Baixo Rio Doce/ES**. 2007. Tese (Doutorado em Geografia) - Instituto de Geociências, Departamento de Geografia, Universidade Federal Fluminense, Niterói, 2007.
- COLLINS, A. J.; FOX, J.; LONG, J. S. Modern Methods of Data Analysis. **The Statistician**, v. 40, n. 4, p. 458, 1991. <https://doi.org/10.2307/2348744>.
- COSTA, R. L.; GOMES, H. B.; PINTO, D. D. C.; ROCHA JÚNIOR, R. L.; SILVA, F. D. S.; GOMES, H. B. *et al.* Gap Filling and Quality Control Applied to Meteorological Variables Measured in the Northeast Region of Brazil. **Atmosphere**, v. 12, n. 10, p. 1278, 2021. <https://doi.org/10.3390/atmos12101278>.
- EKEU-WEI, I. T.; BLACKBURN, G. A.; PEDRUCO, P. Infilling missing data in hydrology: Solutions using satellite radar altimetry and multiple imputation for data-sparse regions. **Water**, v. 10, n. 10, p. 1-22, 2018. <https://doi.org/10.3390/w10101483>.
- ELSHORBAGY, A. A.; PANU, U. S.; SIMONOVIC, S. P. Group-based estimation of missing hydrological data: II. Application to streamflows. **Hydrological Sciences Journal**, v. 45, n. 6, p. 867-880, 2000. <https://doi.org/10.1080/02626660009492389>.
- ELY, D. F.; LIMBERGER, L.; MANGILI, F. B.; GAMERO, P.; SCHMENGLER, M. Análise de métodos para o preenchimento de falhas aplicados em séries de dados pluviométricos do estado do Paraná (Brasil). **Raega - O Espaço Geográfico em Análise**, v. 51, p. 122–142, 2021.
- FERRARI, G. T.; OZAKI, V. Missing data imputation of climate datasets: Implications to modeling extreme drought events. **Revista Brasileira de Meteorologia**, v. 29, n. 1, p. 21-28, 2014. <https://doi.org/10.1590/S0102-77862014000100003>.
- FIOREZE, A. P.; OLIVEIRA, L. F. C. Uses of hydric resources at the Santa Bárbara River hydrographical basin, Goiás state, Brazil. **Pesquisa Agropecuária Tropical**, v. 40, n. 1, p. 28-35, 2010. <https://doi.org/10.5216/pat.v40i1.3869>.
- FONSECA, J. S. D.; MARTINS, G. D. A. **Curso de estatística**. São Paulo: Atlas, 2011.
- GAO, Y. **Dealing with missing data in hydrology - Data analysis of discharge and groundwater timeseries in Northeast Germany**. 2017. Doctoral dissertation (Doctorate degree in Natural Sciences) - Department of Earth Sciences, Freie Universität Berlin, Berlin, 2017.
- GAO, Y.; MERZ, C.; LISCHIED, G.; SCHNEIDER, M. A review on missing hydrological data processing. **Environmental Earth Sciences**, v. 77, n. 2, p. 1-12, 2018. <https://doi.org/10.1007/s12665-018-7228-6>.

- GARCÍA-PEÑA, M.; ARCINIEGAS-ALARCÓN, S.; BARBIN, D. Imputação de dados climáticos utilizando a decomposição por valores singulares: uma comparação empírica. **Revista Brasileira de Meteorologia**, v. 29, n. 4, p. 527-536, 2014. <https://doi.org/10.1590/0102-778620130005>.
- GHAZALI, S. M.; SHAADAN, N.; IDRUS, Z. A Comparative Study of Several EOF Based Imputation Methods for Long Gap Missing Values in a Single-Site Temporal Time Dependent (SSTTD) Air Quality (PM10) Dataset. **Pertanika Journal of Science & Technology**, v. 29, n. 4, 2021. <https://doi.org/10.47836/pjst.29.4.21>.
- GREEN, P.; SILVERMAN, B. **Nonparametric Regression and Generalized Linear Models**. New York: Chapman and Hall/CRC, 1993.
- GYAU-BOAKYE, P.; SCHULTZ, G. A. Filling gaps in runoff time series in West Africa. **Hydrological Sciences Journal**, v. 39, n. 6, p. 621-636, 1994. <https://doi.org/10.1080/02626669409492784>.
- HAMZAH, F. B.; MOHAMAD HAMZAH, F.; MOHD RAZALI, S. F.; EL-SHAFIE, A. Multiple Imputations by Chained Equations for Recovering Missing Daily Streamflow Observations: A Case Study of Langat River Basin in Malaysia. **Hydrological Sciences Journal**, v. 67, n. 1, p. 137-149, 2022. <https://doi.org/10.1080/02626667.2021.2001471>.
- HAMZAH, F. B.; MOHD HAMZAH, F.; MOHD RAZALI, S. F.; JAAFAR, O.; ABDUL JAMIL, N. Imputation methods for recovering streamflow observation: A methodological review. **Cogent Environmental Science**, v. 6, n. 1, p. 1-21, 2020. <https://doi.org/10.1080/23311843.2020.1745133>.
- HERAS, D.; MATOVELLE, C. Machine-learning methods for hydrological imputation data: analysis of the goodness of fit of the model in hydrographic systems of the Pacific-Ecuador. **Revista Ambiente & Água**, v. 16, 2021. <https://doi.org/10.4136/ambi-agua.2708>.
- JIANG, L.; ZHAO, T.; FENG, C.; ZHANG, W. Improvement of Random Forest by Multiple Imputation Applied to Tower Crane Accident Prediction with Missing Data. **Engineering, Construction and Architectural Management**, 2021. <https://doi.org/10.1108/ECAM-07-2021-0606>.
- JIAO, S.; TIEZZI, F.; HUANG, Y.; GRAY, K.A.; MALTECCA, C. The Use of Multiple Imputation for the Accurate Measurements of Individual Feed Intake by Electronic Feeders. **Journal of Animal Science**, v. 94, n. 2, p. 824-832, 2016. <https://doi.org/10.2527/jas.2015-9667>.
- JOHNSTON, C. A. **Development and evaluation of infilling methods for missing hydrologic and chemical watershed monitoring data**. 1999. Thesis (Master of Science degree in Environmental Engineering) - Faculty of the Virginia Polytechnic Institute and State University, Virginia, 1999.
- JUNGER, W. L. **Análise, imputação de dados e interfaces computacionais em estudos de séries temporais epidemiológicas**. 2008. Tese (Doutorado em Saúde Coletiva) – Instituto de Medicina Social, Universidade do Estado do Rio de Janeiro, Rio de Janeiro, 2008.
- JUNGER, W. L.; LEON, A. P. Imputation of missing data in time series for air pollutants. **Atmospheric Environment**, v. 102, p. 96-104, 2015. <https://doi.org/10.1016/j.atmosenv.2014.11.049>.

- KABIR, G.; TEFAMARIAM, S.; HEMSING, J.; SADIQ, R. Handling incomplete and missing data in water network database using imputation methods. **Sustainable and Resilient Infrastructure**, v. 5, n. 6, p. 365-377, 2019. <https://doi.org/10.1080/23789689.2019.1600960>.
- KAMWAGA, S.; MULUNGU, D. M.; VALIMBA, P. Assessment of empirical and regression methods for infilling missing streamflow data in Little Ruaha catchment Tanzania. **Physics and Chemistry of the Earth, Parts A/B/C**, v. 106, p. 17-28, 2018. <https://doi.org/10.1016/j.pce.2018.05.008>.
- KHALIFELOO, M. H.; MOHAMMAD, M.; HEYDARI, M. Multiple imputation for hydrological missing data by using a regression method (Klang river basin). **International Journal of Research Engineering and Technology**, v. 4, n. 6, p. 519-524, 2015. <https://doi.org/10.15623/ijret.2015.0406090>.
- KHAN, N. A.; TORRALBA, K. D.; ASLAM, F. Missing data in randomised controlled trials of rheumatoid arthritis drug therapy are substantial and handled inappropriately. **RMD open**, v. 7, n. 2, p. e001708, 2021. <https://doi.org/10.1136/rmdopen-2021-001708>.
- KIM, M.; BAEK, S.; LIGARAY, M.; PYO, J.; PARK, M.; CHO, K. H. Comparative studies of different imputation methods for recovering streamflow observation. **Water**, v. 7, n. 12, p. 6847-6860, 2015. <https://doi.org/10.3390/w7126663>.
- LITTLE, R. J. A.; RUBIN, D. B. **Statistical analysis with missing data**. Hoboken: John Wiley & Sons, 2019. <https://doi.org/10.1002/9781119013563>.
- LONGHI, E. H.; FORMIGA, K. T. M. Metodologias para determinar vazão ecológica em rios. **Revista Brasileira de Ciências Ambientais (Online)**, v. 20, p. 33-48, 2011.
- McKNIGHT, P. E.; McKNIGHT, K. M.; SIDANI, S.; FIGUEREDO, A. J. **Missing data: a gentle introduction**. New York: Guilford Press, 2007.
- MELLO, Y. R.; KOHLS, W.; OLIVEIRA, T. M. N. Uso de diferentes métodos para o preenchimento de falhas em estações pluviométricas. **Boletim de Geografia**, v. 35, n. 1, p. 112-121, 2017. <https://doi.org/10.4025/bolgeogr.v35i1.30893>.
- MOREIRA, M. C. **Gestão de recursos hídricos: sistema integrado para otimização da outorga de uso da água**. 2006. Dissertação (Mestrado em Engenharia Agrícola) - Universidade Federal de Viçosa, Viçosa, 2006.
- NISHINA, K.; ITO, A.; HANASAKI, N.; HAYASHI, S. Reconstruction of spatially detailed global map of  $\text{NH}_4^+$  and  $\text{NO}_3^-$  application in synthetic nitrogen fertilizer. **Earth System Science Data**, v. 9, n. 1, p. 149-162, 2017. <https://doi.org/10.5194/essd-9-149-2017>.
- NKIAKA, E.; NAWAZ, N. R.; LOVETT, J. C. Using self-organizing maps to infill missing data in hydro-meteorological time series from the Logone catchment, Lake Chad basin. **Environmental Monitoring and Assessment**, v. 188, n. 7, p. 1-12, 2016. <https://doi.org/10.1007/s10661-016-5385-1>.
- NORLIYANA, W.; ISMAIL, W.; ZAWIAH, W.; ZIN, W.; IBRAHIM, W. Estimation of rainfall and stream flow missing data for Terengganu, Malaysia by using interpolation technique methods. **Malaysian Journal of Fundamental & Applied Sciences**, v. 13, n. 3, p. 213-217, 2017. <https://doi.org/10.11113/mjfas.v13n3.578>.



- NUNES, L. N. **Métodos de imputação de dados aplicados na área de saúde**. 2007. Tese (Doutorado em Epidemiologia) - Faculdade de medicina, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2007.
- NUNES, L. N.; KLÜCK, M. M.; FACHEL, J. M. G. Uso da imputação múltipla de dados faltantes: uma simulação utilizando dados epidemiológicos. **Cadernos de Saúde Pública**, v. 25, n. 2, p. 268-278, 2009. <https://doi.org/10.1590/S0102-311X2009000200005>.
- NUNES, L. N.; KLÜCK, M. M.; FACHEL, J. M. G. Comparação de métodos de imputação única e múltipla usando como exemplo um modelo de risco para mortalidade cirúrgica. **Revista Brasileira de Epidemiologia**, v. 13, n. 4, p. 596-606, 2010.
- OLIVEIRA, L. F.; FIOREZE, A. P.; MEDEIROS, A. M.; SILVA, M. A. Comparação de metodologias de preenchimento de falhas de séries históricas de precipitação pluvial anual. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v. 14, n. 11, p. 1186-1192, 2010. <https://doi.org/10.1590/S1415-43662010001100008>.
- PAYROVNAZIRI, S. N.; XING, A.; SALMAN, S.; LIU, X.; BIAN, J.; HE, Z. Assessing the impact of imputation on the interpretations of prediction models: A case study on mortality prediction for patients with acute myocardial infarction. **AMIA Joint Summits on Translational Science**, v. 2021, p. 465-474, 2021.
- PELISSON, A. A. **Aprendizado de máquina para previsão de geração de energia fotovoltaica em dados de estações solarimétricas**. 2021. Dissertação (mestrado) - Universidade Federal de Santa Catarina. Joinville, 2021.
- PEÑA-ANGULO, D.; NADAL-ROMERO, E.; GONZÁLEZ-HIDALGO, J. C.; ALBALADEJO, J.; ANDREU, V.; BAGARELLO, V. *et al.* Spatial variability of the relationships of runoff and sediment yield with weather types throughout the Mediterranean basin. **Journal of Hydrology**, v. 571, p. 390-405, 2019. <https://doi.org/10.1016/j.jhydrol.2019.01.059>.
- PIGOTT, T. D. A review of methods for missing data. **Educational Research and Evaluation**, v. 7, n. 4, p. 353-383, 2001. <https://doi.org/10.1076/edre.7.4.353.8937>.
- PINTO, W. P. **O uso da metodologia de dados faltantes em séries temporais com aplicações a dados de concentração (PM<sub>10</sub>) observados na Região da Grande Vitória**. 2013. Dissertação (Mestrado em Engenharia Ambiental) - Universidade Federal do Espírito Santo, Vitória, 2013.
- PINTO, W. P.; LIMA, G. B.; ZANETTI, J. B. Análise comparativa de modelos de séries temporais para modelagem e previsão de regimes de vazões médias mensais do Rio Doce, Colatina-Espírito Santo. **Ciência e Natura**, v. 37, n. 3, p. 1-11, 2015. <http://dx.doi.org/10.5902/2179460X17143>.
- PRUSKI, F. F.; PEREIRA, S. B.; NOVAES, L. F. D.; SILVA, D. D. D.; RAMOS, M. M. Precipitação média anual e vazão específica média de longa duração, na Bacia do São Francisco. **Revista Brasileira de Engenharia Agrícola e Ambiental**, v. 8, n. 2-3, p. 247-253, 2004. <https://doi.org/10.1590/S1415-43662004000200013>.
- R DEVELOPMENT CORE TEAM. **R: A language and environment for statistical computing**. Vienna, 2018.

- RADO, O.; FANAHA, M.A.; TAKTEK, E. Performance analysis of missing values imputation methods using machine learning techniques. *In: ARAI, K.; BHATIA, R.; KAPOOR, S. (eds.). Intelligent Computing-Proceedings of the Computing Conference*. Springer Cham, 2019. p. 738-750. [https://doi.org/10.1007/978-3-030-22871-2\\_51](https://doi.org/10.1007/978-3-030-22871-2_51).
- RAHMAN, N. A.; DENI, S. M.; RAMLI, N. M. Generalized linear model for estimation of missing daily rainfall data. *In: AIP CONFERENCE, 2017. Proceedings[...]* AIP Publishing LLC, 2017. <https://doi.org/10.1063/1.4981003>.
- REISEN, V. A.; MOLINARES, F. A. F.; TEIXEIRA, E. C. Modelagem de séries temporais sazonais na presença de outliers: Estudo de caso da vazão máxima mensal do rio Jucu, ES, Brasil. **RBRH**, v. 13, n. 2, p. 45-53, 2008. <http://dx.doi.org/10.21168/rbrh.v13n2.p45-53>.
- ROTH, P. L.; SWITZER, F. S.; SWITZER, D. M. Missing data in multiple item scales: A Monte Carlo analysis of missing data techniques. **Organizational Research Methods**, v. 2, n. 3, p. 211–232, 1999. <https://doi.org/10.1177/109442819923001>.
- RUBIN, D. B. Procedures with nonignorable nonresponse. *In: RUBIN, D. B. Multiple Imputation for Nonresponse in Surveys*. New York: John Wiley & Sons, 1987. p. 202-240.
- SABINO, C. V. S.; LAGE, L. V.; ALMEIDA, K. C. B. Uso de métodos estatísticos robustos na análise ambiental. **Engenharia Sanitária e Ambiental**, v. 19, n. SPE, p. 87-94, 2014. <https://doi.org/10.1590/S1413-41522014019010000588>.
- SARMENTO, R. **Termo de referência para a elaboração de estudos sobre a vazão ecológica na bacia do rio São Francisco**. Referência: Edital, (05) do ano de 2006, PROJETO 704BRA2041 da Organização das Nações Unidas para a Educação, a Ciência e a Cultura – UNESCO. 2007. Available at: [http://cbhsaofrancisco.org.br/2017/?wpfb\\_dl=1584](http://cbhsaofrancisco.org.br/2017/?wpfb_dl=1584). Access in Apr. 20, 2018.
- SATTARI, M.; REZAZADEH-JOUDI, A.; KUSIAK, A. Assessment of different methods for estimation of missing data in precipitation studies. **Hydrology Research**, v. 48, n. 4, p. 1032-1044, 2017. <https://doi.org/10.2166/nh.2016.364>.
- SCHAFER, J. L. Multiple imputation: a primer. **Statistical methods in medical research**, v. 8, n. 1, p. 3-15, 1999. <https://doi.org/10.1177%2F096228029900800102>.
- SEMIROMI, M. T.; KOCH, M. Reconstruction of groundwater levels to impute missing values using singular and multichannel spectrum analysis: application to the Ardabil Plain, Iran. **Hydrological Sciences Journal**, v. 64, n. 14, p. 1711-1726, 2019. <https://doi.org/10.1080/02626667.2019.1669793>.
- SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, v. 52, n. 3/4, p. 591-611, 1965. <https://doi.org/10.2307/2333709>.
- SILVA, M. J. C. **Imputação múltipla: comparação e eficiência em experimentos multi ambientais**. 2012. Dissertação (Mestrado em Estatística e Experimentação Agronômica) – Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 2012.
- SOUSA, H. T.; PRUSKI, F. F.; BOF, L. H. N.; CECON, P. R.; SOUZA, J. R. C. **SisCAH 1.0: Sistema computacional para análises hidrológicas**. Brasília: ANA; Viçosa, MG: UFV, 2009.

- STARRETT, S. K.; STARRETT, S. K.; HEIER, T.; SU, Y.; TUAN, D.; BANDURRAGA, M. Filling in missing peak flow data using artificial neural networks. **ARPN Journal of Engineering and Applied Sciences**, , v. 5, n. 1, p. 49-55, 2010.
- SWENSON, N. G. Phylogenetic imputation of plant functional trait databases. **Ecography**, v. 37, n. 2, p. 105-110, 2014. <https://doi.org/10.1111/j.1600-0587.2013.00528.x>.
- TENCALIEC, P. **Developments in statistics applied to hydrometeorology: Imputation of streamflow data and semiparametric precipitation modeling**. 2017. Thesis (Doctorate degree in Applied Mathematics) - University Grenoble Alpes, Grenoble, 2017.
- TENCALIEC, P.; FAVRE, A. C.; PRIEUR, C.; MATHEVET, T. Reconstruction of missing daily streamflow data using dynamic regression models. **Water Resources Research**, v. 51, n. 12, p. 9447-9463, 2015. <https://doi.org/10.1002/2015WR017399>.
- TUCCI, C. E. **Hidrologia: Ciência e Aplicação**. Porto Alegre: Editora da UFRGS/Edusp/ABRH, 1997.
- TURICCHI, J.; O'DRISCOLL, R.; FINLAYSON, G.; DUARTE, C.; PALMEIRA, A. L.; LARSEN, S. C. *et al.* Data imputation and body weight variability calculation using linear and nonlinear methods in data collected from digital smart scales: simulation and validation study. **JMIR mHealth and uHealth**, v. 8, n. 9, p. e17977, 2020. <https://doi.org/10.2196/17977>.
- VEGA-GARCIA, C.; DECUYPER, M.; ALCÁZAR, J. Applying cascade-correlation neural networks to in-fill gaps in Mediterranean daily flow data series. **Water**, v. 11, n. 8, p. 1691, 2019. <https://doi.org/10.3390/w11081691>.
- VENTURA, T. M.; SANTANA, L. L. R.; MARTINS, C. A.; FIGUEIREDO, J. M. Análise da aplicabilidade de métodos estatísticos para preenchimento de falhas em dados meteorológicos (Analysis methods of application for statistical data in meteorology). **Revista Brasileira de Climatologia**, v. 19, n. 12, p. 168-177, 2016. <http://dx.doi.org/10.5380/abclima.v19i0.44989>.
- VINHA, L. G. A; LAROS, J. A. Dados ausentes em avaliações educacionais: comparação de métodos de tratamento. **Estudos em Avaliação Educacional**, v. 29. n. 70, p. 156-187, 2018. <https://doi.org/10.18222/ea.v0ix.3916>.
- WIJESEKARA, W.M.L.K.N.; LIYANAGE, L. Comparison of imputation methods for missing values in air pollution data: Case study on Sydney air quality index. *In*: ARAI, K.; BHATIA, R.; KAPOOR, S. (eds.). **Advances in Information and Communication: Proceedings of the 2020 Future of Information and Communication Conference (FICC)**. Springer Cham, 2020. p. 257-269. [https://doi.org/10.1007/978-3-030-39442-4\\_20](https://doi.org/10.1007/978-3-030-39442-4_20).