Ambiente & Água - An Interdisciplinary Journal of Applied Science
ISSN 1980-993X – doi:10.4136/1980-993X
www.ambi-agua.net
E-mail: ambi.agua@gmail.com

Filling and validating rainfall data based on statistical techniques and artificial intelligence

ARTICLES doi:10.4136/ambi-agua.2767

Received: 05 Jul. 2021; Accepted: 10 Nov. 2021

Camila Bermond Ruezzene[1*]; Renato Billia de Miranda[2];
Talyson de Melo Bolleli[1]; Frederico Fábio Mauad[1]

[1]Escola de Engenharia de São Carlos. Departamento de Hidráulica e Saneamento. Universidade de São Paulo (USP), Avenida Trabalhador são-carlense, n° 400, CEP: 13566-590, São Carlos, SP, Brazil.
E-mail: bolleli@usp.br, mauadffm@sc.usp.br
[2]Gerência de Cursos e Matrizes. Anhanguera Educacional, Alameda Maria Tereza, n° 4266,
CEP: 13278-181, Valinhos, SP, Brazil. E-mail: eng.renato.miranda@gmail.com
[*]Corresponding author. E-mail: camila.ruezzene@gmail.com

ABSTRACT

The study of the hydric regime of rainfall helps in management analysis and decision-making in hydrographic basins, but a fundamental condition is the need for continuous time series of data. Therefore, this study compared gap filling methods in precipitation data and validated them using robust statistical techniques. Precipitation data from the municipality of Itirapina, which has four monitoring stations, were used. Four gap filling techniques were used, namely: normal ratio method, inverse distance weighting, multiple regression and artificial neural networks, in the period from 1979 to 1989. For validation and performance evaluation, the coefficient of determination (R^2), mean absolute error (MAE), mean squared error (RMSE), Nash-Sutcliffe coefficient (Nash), agreement index (D), confidence index were used (C) and through non-parametric techniques with Mann-Witney and Kruskal-Wallis test. Excellent performances of real data were verified in comparison with estimated data, with values above 0.8 of the coefficient of determination (R^2) and of Nash. Kruskal-Wallis and Mann-Whitney tests were not significant in Stations C1 and C2, demonstrating that there is a difference between real and estimated data and between the proposed methods. It was concluded that the multiple regression and neural network methods showed the best performance. From this study, efficient tools were found to fill the gap, thus promoting better management and operation of water resources.

Keywords: artificial neural networks, inverse distance weighting, multiple regression, normal ratio method.

Preenchimento e validação em dados de precipitação através de técnicas estatísticas e de inteligência artificial

RESUMO

O estudo do regime hídrico das chuvas auxilia nas análises de gestão e em tomadas de decisão nas bacias hidrográficas, mas uma condição fundamental é a necessidade de séries temporais contínuas de dados. Diante disso, o objetivo do presente estudo foi realizar a comparação entre os métodos de preenchimento de falha em dados de precipitação e validá-los através de técnicas estatísticas robustas. Foram utilizados dados de precipitação localizados no município de Itirapina que conta com quatro estações de monitoramento. Foi empregado quatro técnicas de preenchimento de falhas, sendo: método razão normal, ponderação distância inversa, regressão múltipla e redes neurais artificiais, no período de 1979 a 1989. Para validação e avaliação do desempenho utilizou-se o coeficiente de determinação (R^2), erro absoluto médio (MAE), erro quadrático médio (RMSE), coeficiente de Nash-Sutcliffe (Nash), índice de concordância (D), índice de confiança (C) e através de técnicas não paramétricas com teste de Mann-Whitney e Kruskal-Wallis. Foram verificados ótimos desempenhos dos dados reais em comparação aos dados estimados, com valores acima de 0,8 do coeficiente de determinação (R^2) e de Nash. Para os testes de Kruskal-Wallis e Mann-Whitney não foram significativos nas estações C1 e C2, demonstrando que existe diferença entre os dados reais e estimados e entre os métodos propostos. Pôde-se concluir que os métodos de regressão múltipla e redes neurais apresentaram os melhores desempenhos. A partir desse estudo verificou-se ferramentas eficientes para o preenchimento de falha promovendo assim, uma melhor gestão e operação dos recursos hídricos.

Palavras-chave: método razão normal, ponderação distância inversa, redes neurais artificiais, regressão múltipla.

1. INTRODUCTION

Rainfall is one of the variables with the greatest influence on society, environment and economy, as it has direct implications for agriculture, climate, hydrology, disaster management, among others. Therefore, evaluating its behavior allows it to assist in the analysis of water availability and decision-making in hydrographic basins. However, for such verifications to be carried out, a fundamental condition is the need for continuous time series of data in order to obtain consistent and reliable results (Correia *et al.* 2016).

A constant problem in developing countries is the absence of continuous data on meteorological variables and rainfall stations in different regions, which reinforces the importance of making the most of existing data. Since gaps in databases can influence data analysis and inferences, missing data-filling methods, such as multiple regression method (MR), inverse distance weighting (IDW), normal ratio method (NRM), and neural networks (NN), stand out in the scientific environment due to their excellent performance in calculating estimates (Depiné *et al.*, 2014; Wanderley *et al.*, 2014; Khosravi *et al.*, 2015; Correia *et al.*, 2016; Bier and Ferraz, 2017; Coutinho *et al.* 2018).

Several studies have highlighted the multiple linear regression method as an efficient tool used to fill gaps in time series, such as rainfall, temperature and humidity data, among others. According to Oliveira *et al.* (2010), multiple linear regression and regional weighting perform better in filling in gaps than regional vector and regional weighting methods based on linear regressions. However, the aforementioned authors have emphasized that these methodologies should not be used without prior regional analysis of their performance.

On the other hand, the inverse distance weighting method is one of the techniques mostly used to estimate missing data in hydrology and geographic sciences, since it presents satisfactory results in filling in gaps about rainfall data (Teegavarapu and Chandramouli, 2005; Shepard, 1968).

Junqueira *et al.* (2018) have compared different rainfall missing data-filling methodologies and found that methods such as regional weighting, arithmetic mean and regional weighting based on regression have overestimated rainfall rates in the Mortes River Basin (MG), whereas linear regression, multiple regression and inverse distance weighting methods have underestimated them.

According to Bier and Ferraz (2017), regional weighting was the most suitable method used for missing rainfall data-filling purposes, but it did not significantly stand out in comparison to methods such as multiple linear regression, regional weighting, inverse distance weighting, normal ratio method, United Kingdom traditional method and simple arithmetic mean. Normal ratio method has shown satisfactory results and low mean absolute errors (18.1%) in estimates.

The aforementioned authors have also pointed out that these estimates have represented monthly rainfall variations in a reasonable way. They were capable of detecting monthly rainfall peaks between original and estimated series, and it has evidenced the possibility of generating good estimates for monthly and annual rainfall data (Bier and Ferraz, 2017).

Coutinho *et al.* (2018) used a Multilayer Perceptron-type neural network fault filling tool comparing monthly meteorological variables with Multiple Regression models in four stations in the state of Rio de Janeiro from 2002 to 2014. The neural network model presented a high linear correlation (r) with the recorded data of maximum air temperature (r from 0.94 to 0.98), obtaining a mean percentage error (EMP) between 1.05% and 2 .32%. Regarding the relative air humidity, the r coefficient remained between 0.77 and 0.94, and the use of RNA to estimate this variable resulted in an EMP between 1.85% and 2.41%. They concluded that the neural network method is an effective tool for filling and reliably estimating meteorological variables, as the estimated data were close to the real data.

Given the range of procedures and statistical techniques used to fill in missing data in climate series, it is of paramount importance to select the most appropriate methodology capable of meeting the needs of the study by taking into consideration the climatic and geographical reality the meteorological stations are inserted in, as well as of statistically proving the veracity of estimated data (Fante and Sant'Anna Neto, 2016).

Assessing and validating the performance of hydrological models is a crucial process to justify their continued use, as well as verifying their limitations. In this way, it is necessary to apply statistical criteria to quantify the quality and accuracy of the adjustment, in relation to the measured and estimated data, which enables a comprehensive evaluation of the models, enabling their proper reproduction, the prediction of future behavior and the identification of improvements in modeling.

The most commonly used criteria for evaluating models are commonly divided between performance indices, which include: coefficient of determination (R^2), agreement (d) and confidence (c) index, and Nash-Sutcliffe efficiency (Nash) and error checking measures such as mean absolute error (MAE) and mean squared error (RMSE). There are no standard procedures in the literature to assess the performance of models, but it is preferable that different statistical techniques are applied, as different metrics quantify various aspects of model fit and accuracy, providing a quantitative and objective assessment of the agreement between simulated and estimated data (Jackson, 2019).

Thus, the present study compared gap filling methods in precipitation data and validated them using robust statistical techniques.

2. MATERIALS AND METHODS

2.1. Study site featuring

The study site is located in Itirapina County-SP (Figure 1), approximately 218 km away from the capital; its population is estimated as 18,157 inhabitants and its territorial area covers 564.60 km² comprising *Cerrado* and Atlantic Forest biomes (IBGE, 2019).

The county has two important conservation units, namely: Experimental Station and Ecological Station. They are managed by the Forestry Institute, which manages an area of 5,512 hectares and aims at environmental preservation, research and education. The water network in these units, together with the other water bodies, are of paramount importance because they play a key role in the overall balance of the region (Silva *et al.*, 2006).

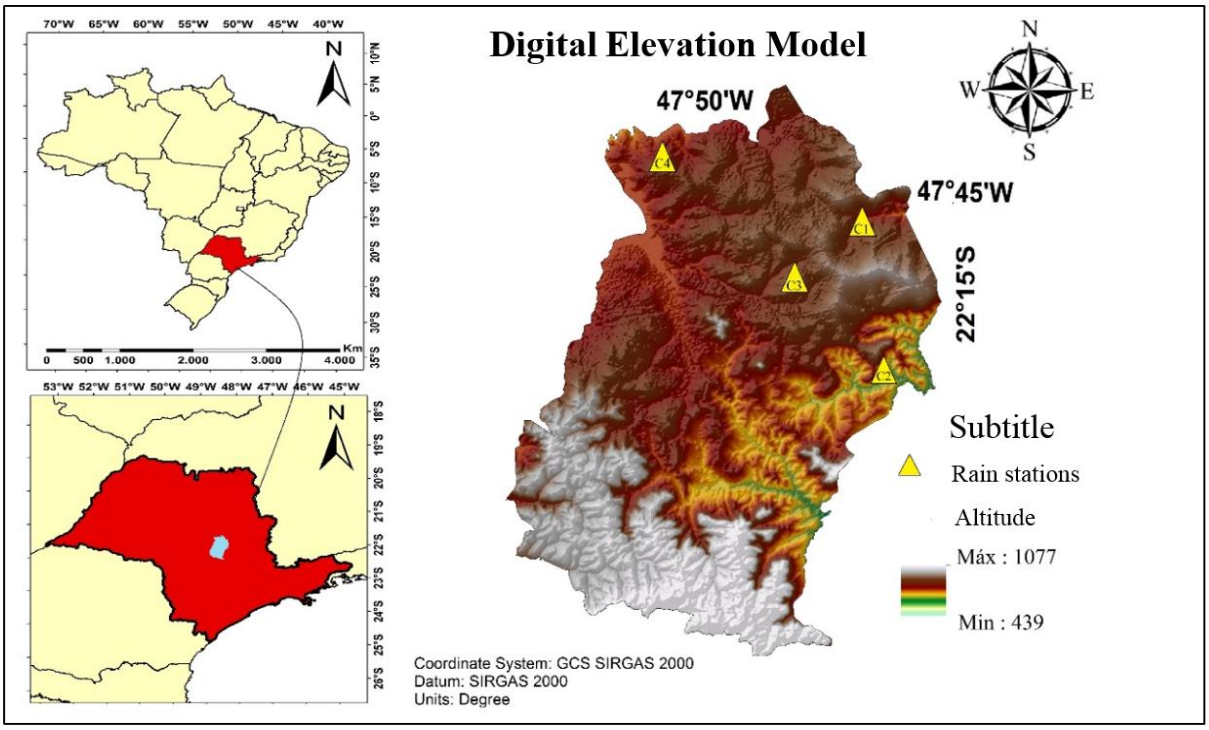
70°W 65°W 60°W 55°W 50°W 45°W 40°W
N
10°N 5°N 0° 5°S 10°S 15°S 20°S 25°S 30°S 35°S
0 500 1.000 2.000 3.000 4.000 Km
53°W 52°W 51°W 50°W 49°W 48°W 47°W 46°W 45°W
18°S 19°S 20°S 21°S 22°S 23°S 24°S 25°S 26°S
Digital Elevation Model
47°50'W
47°45'W
22°15'S
N
W
E
S
C1
C2
C3
C4
Subtitle
Rain stations
Altitude
Máx : 1077
Min : 439
Coordinate System: GCS SIRGAS 2000
Datum: SIRGAS 2000
Units: Degree

Figure 1. Location of the study area.

Itirapina County is located in the Cuestas Basálticas region, on São Carlos Plateau. The climate in the region is classified as Cwa, based on Köppen's classification. Summer is hot and rainy, whereas winter is dry. Seasonality observed in the region comprises dry semester from April to September, as well as rainy semester from October to March. Mean annual rainfall reaches 1,450 mm per year and mean annual temperature is approximately 20.8°C (CEPAGRI, 2009; Santos *et al.*, 2018b).

According to Santos *et al.* (2018a), land use and cover classification in Itirapina County corresponds to native vegetation (30.68%), agricultural crops such as sugarcane (19.93%), forestry (18.37%), citrus culture (1.55%), exposed soil (11.97%) and pasture (14.55%). The aforementioned authors have emphasized that the region presents low environmental vulnerability level due to its flat terrain and to the incidence of Oxisol in the county.

2.2. Rainfall data

Monthly historical data about rainfall recorded in Itirapina County (SP) from 1979 to 1989 were selected from stations available at the HIDROWEB portal, on the National Water Agency (ANA, 2020) platform. Table 1 shows the stations and their respective codes, altitude, geographic coordinates, as well as the time (in years) of each analyzed series - these data were consolidated before they were made available.

Table 1. Rain stations located in Itirapina County.

Adopted nomenclature	Code	Altitude	Latitude	Longitude
C1	2247180	760	22°08'54"	47°47'42"
C2	2247184	610	22°18'01"	47°44'38"
C3	2247196	732	22°10'12"	47°53'56.04"
C4	2247198	690	22°10'00"	47°54'00"

Source: Adapted from ANA (2020).

The period from 1979 to 1989 was adopted due to the need to work with a continuous series of data, allowing better representation of the characteristics present in each station and the comparison of real and estimated data for each proposed method, using all stations available for that municipality. The climatological station (C4) located at the Center for Water Resources and Environmental Studies (CRHEA/USP) was included in the group of meteorological stations because it makes it possible to work with primary data. Considering all available data from this period of study, 11.37% of the data were removed and thus obtained in a homogeneous series and with the same number of data for all stations according to the methodology of Coutinho et al. (2018).

2.3. Missing data-filling

Criteria adopted to select the tested methods took into consideration methodologies already consolidated in the field. Thus, rather than being not limited to a single methodology, the current study used several of them in order to compare and validate them, based on different statistical techniques that will be addressed throughout the methodology and results. The following techniques stood out among the main missing data-filling methods and they were used in the current study: multiple regression (MR); inverse distance weighting (IDW), normal ratio method (NRM) and artificial neural networks (ANNs).

2.3.1. Multiple regression (MR)

Rainfall information about the behavior of a dependent variable *Y* in multiple regression depends on two, or more, independent variables *Xj*, *j = 1, ..., p* (Naghettini and Pinto, 2007). Therefore, a model likely to evaluate this association is enabled by Equation 1.

$$Yi = \beta 0 + \beta 1Xi1 + \beta 2Xi2 + \ldots + \beta pXip + ei\,,\ i = 1, \ldots, n \quad (1)$$

Where in: *n* is the number of observations, Y_i is the observation of the dependent variable for the *i-th* individual, $X_i = (X_{i1}, X_{i2}, \ldots, X_{ip})$ is a vector of observations of independent variables for the *i-th* individual, $\beta = (\beta_0, \beta_1, \beta_2, \ldots, \beta_p)$ is a vector of regression coefficients (parameters) and e_i is a random error component.It is presumed that these errors are independent and follow normal distribution with mean equal to zero and unknown variance σ^2.

2.3.2. Inverse Distance Weighting (IDW)

The inverse distance weighting method is applied through the linear combination of observations within a given research radius, whose influence decreases as distance increases. According to Hubbard (1994), the IDW method for missing data-filling is calculated based on Equation 2.

$$D_x = \frac{\sum_{i=1}^{n}(D_i/d_i)}{\sum_{i=1}^{n}(1/d_i)} \quad (2)$$

Where in: *Dx* is the missing monthly data to be filled in the test station, D_i corresponds to data deriving from the neighboring station of order “i” in the month when the failure in the test station takes place, and d_i is the distance between the test station and the neighboring station of order “i”.

2.3.3. Normal Ratio Method (NRM)

According to Young (1992), the normal ratio method lies on weighting data based on records performed at neighboring stations; such a ratio can be calculated through Equation 3.

$$D_x = \frac{\sum_{i=1}^{n} D_i w_i}{\sum_{i=1}^{n} w_i} \quad (3)$$

Where in: *Dx* is the monthly data that needs to be filled in the test station, *Di* corresponds to data deriving from the neighboring station of order “i” in the month when the failure in the test station takes place, and *wi* is the weight assigned to each neighboring station of order “i”, as described in Equation 4.

$$w_i = r^2{}_i\left(\frac{n_i-2}{1-r^2{}_i}\right) \tag{4}$$

Where in: r_i is the correlation between the test station and the neighboring station of order “i”, and n_i is the number of months when data overlapped between the test station and the neighboring station of order “i”. In other words, it is the size of the data series used to calculate the correlation coefficient.

2.3.4. Artificial Neural Networks (ANN)

Neural networks are calculated through mathematical functions; they are naturally prone to store knowledge and make it useful, like the process carried out by the human brain. Nonlinear functions are calculated, which can be appropriate for complex analyses, such as estimating rainfall data (Di Piazza *et al.*, 2011; Depiné *et al.*, 2014; Wanderley *et al.*, 2014; Correia *et al.*, 2016; Coutinho *et al.*, 2018).

The current study has used Multilayer Perceptron (MLP) neural networks due to their greater versatility and applicability in this field. This network type can be used to estimate information and new desired conditions, as well as to find accurate answers to the analyses in question (Wanderley *et al.*, 2014).

Figure 2 shows an example of the neural network architecture corresponding to a station. Networks were trained in the MATLAB software, Version R2015a (https://www.mathworks.com/products/matlab.html) developed by the MathWorks company. It was done by using the Feed-forward backpropagation network type and mean square error as a performance function. The current research has defined that the network architecture should have 3 inputs (corresponding to the stations located in the county), 2 layers, 10 neurons, 1 output and a tan-sigmoid activation function. With respect to network training and validation, 70% of data were used for training; 15%, for testing; and 15%, for validation purposes, as established by the software itself.

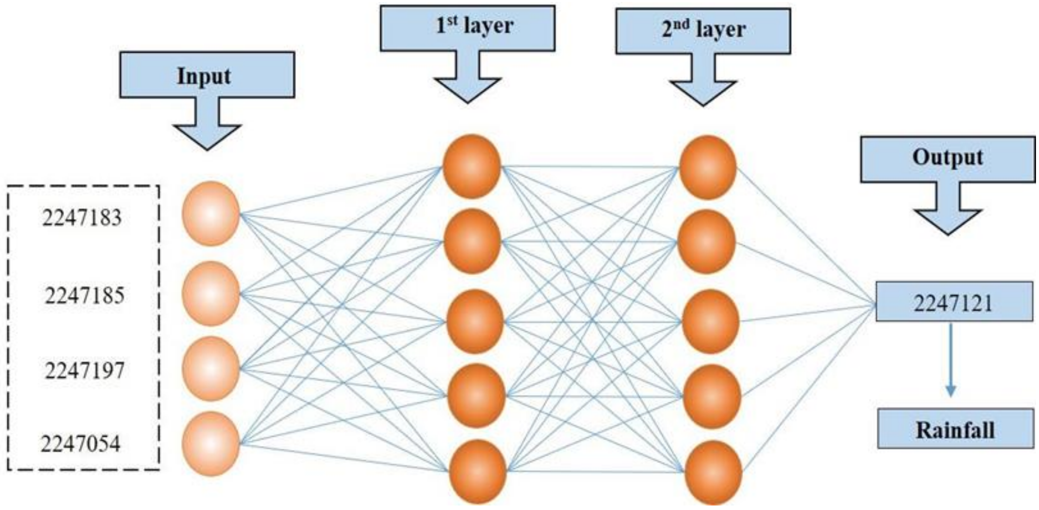

Figure 2. Neural network configuration for Itirapina County, from 1979 to 1989.

As in the example in Figure 2, to fill in the precipitation data for Station 2247180, all available data for that station was removed and three stations were used as input data and one for output, obtaining a total of four networks for this municipality carrying out this entire process for the four seasons. Each trained neural network recognizes and adapts to the characteristic patterns of each station's rainfall data, providing gap-filling results.

2.4. Method-performance validation and evaluation

The coefficient of determination (R²) was calculated to verify the relationship between estimated and measured data. To assess the performance and errors of the failure filling methods, parameters such as mean absolute error (MAE), mean square error (RMSE), index of agreement (D), Pearson's correlation coefficient (r), confidence index (C) and Nash-Sutcliffe efficiency coefficient (Nash), which are applied in several hydrological studies (Goyal, 2014; Pereira *et al.*, 2014; Wanderley *et al.*, 2014; Bier and Ferraz, 2017; Coutinho *et al.*, 2018; Junqueira *et al.*, 2018).

2.4.1. Coefficient of determination (R²)

The coefficient of determination (R²) (Equation 5) assesses the quality of model fit and indicates the extent to which it was capable of explaining the reference data - the higher the recorded value, the better it fits the model.

$$R^2 = \frac{\sum_{i-1}^{n}(\hat{Y}_i - \bar{Y})^2}{\sum_{i-1}^{n}(Y_i - \bar{Y})^2} = 1 - \frac{\sum_{i-1}^{n} e_i^2/(n-1)}{\sum_{i-1}^{n}(Y_i - \bar{Y})^2/(n-1)} \tag{5}$$

Where in: R² is the coefficient of determination measured in (%), Y_i is the observed value of the dependent variable, $\hat{Y}_i$ is the estimated value of the dependent variable, and $\underline{Y}$ is the mean recorded for the dependent variable.

2.4.2. Mean absolute error (MAE)

According to Alves *et al.* (2012), MAE refers to the mean absolute deviation of interpolated values in comparison to the observed ones. It is considered an accurate and robust measure to check numerical models; ideally, its values should be as close or equal to zero as possible (Equation 6).

$$MAE = \frac{\sum_{j=1}^{n}|O_j - x_j|}{n} \tag{6}$$

Where in: MAE is the mean absolute error (mm), O_j concerns values observed in the measurement stations, x_j corresponds to values estimated through the missing data-filling method, and "n" is the number of observations.

2.4.3. Root mean square error (*RMSE)*

RMSE enables checking the mean magnitude of estimated errors. The obtained value is always positive; the closer to zero, the better the estimated values. This parameter can be calculated through Equation 7.

$$RMSE = \sqrt{\frac{\sum_{j=1}^{n}(O_j - x_j)^2}{n}} \tag{7}$$

Where in: RMSE is the mean square error (mm), O_j concerns values observed in the measurement stations, x_j corresponds to values estimated by the missing data-filling method, and "n" is the number of observations.

2.4.4. Confidence (c) and agreement (d) indices

The confidence index enables checking the precision and accuracy of results. The index of agreement is used in different simulations of a single phenomenon. Values recorded for this index range from 0 (lack of agreement) to 1 (excellent agreement). Table 2 shows the criteria used to assess performance. These parameters can be calculated through Equations 8, 9 and 10.

$$D = 1 - \frac{\sum_{j=1}^{n}(O_j - x_j)^2}{\sum_{j=1}^{n}(|x_j - \bar{O}| + |O_j - \bar{O}\ |)^2} \tag{8}$$

$$r = \frac{\frac{\sum_{j=I}^{N}(x_j - \bar{x}) \times (O_j - \bar{O})}{N}}{\frac{\sqrt{\sum_{j=1}^{N}(x_j - \bar{x})^2}}{N} \times \frac{\sqrt{\sum_{j=1}^{N}(O_j - \bar{O})^2}}{N}} \tag{9}$$

$$C = (r \times D) \tag{10}$$

Where in: D refers to the agreement index (dimensionless), r Pearson's correlation coefficient (dimensionless), C confidence index (dimensionless); O_j are the values observed at the measurement stations, $\underline{O}$ mean of observed values, $\underline{x}$ mean estimated values, x_j correspond to values estimated by the filling method and n to the number of observations.

Table 2. Confidence index (C) values used to evaluate and analyze models' performance.

C value	Performance
> 0.85	Excellent
0.76 a 0.85	Very good
0.66 a 0.75	Good
0.61 a 0.65	Intermediate
0.51 a 0.60	Tolerable
0.41 a 0.50	Poor
≤ 0.40	Terrible

Source: Coutinho *et al.* (2018).

2.4.5. Nash–Sutcliffe efficiency coefficient (Nash)

The Nash-Sutcliffe efficiency coefficient (Equation 11) is one of the most important and usual statistical methods applied in hydrology to assess the performance of hydrological models, as described by Pereira *et al.* (2014). This coefficient can range from -∞ to 1; the value corresponding to 1 represents the ideal adjustment of estimated data.

$$\mathrm{Nash} = 1 - \frac{\sum_{i=1}^{n}(X_{obs,i} - X_{sim})^2}{\sum_{i=1}^{n}(X_{obs,i} - \bar{X}_{obs})^2} \tag{11}$$

Where in: *Nash* is the Nash-Sutcliffe efficiency coefficient (dimensionless), X_{obs} are the observed rainfall data, X_{sim} are the rainfall data simulated by the model, $\underline{X}_{obs}$ is the mean recorded for data observed during the simulation period, and n is the number of events.

Model classification was herein adopted based on Silva *et al.* (2006): models presenting coefficient value higher than 0.75 were classified as adequate and good, those whose coefficient value ranged from 0.36 to 0.75 were classified as acceptable, whereas the ones presenting coefficient value lower than 0.36 were considered unacceptable.

2.4.6. Descriptive and inferential analysis and data normality tests

The assumption of normality was checked through the Anderson Darling test, due to its sensitivity in giving more weight to distribution tail points (Espinosa *et al.*, 2004). Shapiro-Wilk test was also applied, since it is one of the most efficient tests used to identify non-normal data (Shapiro and Wilk, 1965).

Both normality tests adopted a significance level of 0.05. H_0 - which corresponds to data presenting normal distribution - was rejected whenever p-values were lower than 0.05. Thus, whenever such data did not show normality, they were subjected to non-parametric analysis techniques, as well as to data median analysis.

The following hypotheses were used in the Anderson Darling and Shapiro-Wilk tests:

$H_{0:}$ data follow normal distribution.

$H_{1:}$ data do not follow normal distribution.

2.4.7. Mann-Whitney (MW) and Kruskal-Wallis (KW) non-parametric tests

Mann-Whitney non-parametric test was carried out at a significance level of 0.05 in order to test whether there were significant differences between the value estimated through the methods and the real rainfall reference data. H_0 was accepted and H_1 was rejected whenever the p-value was higher than the significance level - the two samples presented the same distribution (Triola, 2008).

Hypothesis testing:

H_0: data derived from equal samples.

H_1: data derived from different samples.

Non-parametric KW test initially suggested by Kruskal and Wallis (1952) was also applied at significance level of 0.05 in order to determine whether the medians among all four (4) missing data-filling methods used in the current study have significantly differed from each other. Thus, the null hypothesis was rejected whenever p-value was lower than, or equal to, the significance level, and it led to the conclusion that not all medians were equal to each other.

3. RESULTS AND DISCUSSION

Figure 3 shows the time-based rainfall variability among all four stations located in Itirapina County. They followed similar behavior, i.e., they recorded the highest rainfall values in January 1983 and 1981 (554 mm and 441.2 mm, respectively), which corresponded to the C4 station of CRHEA-USP.

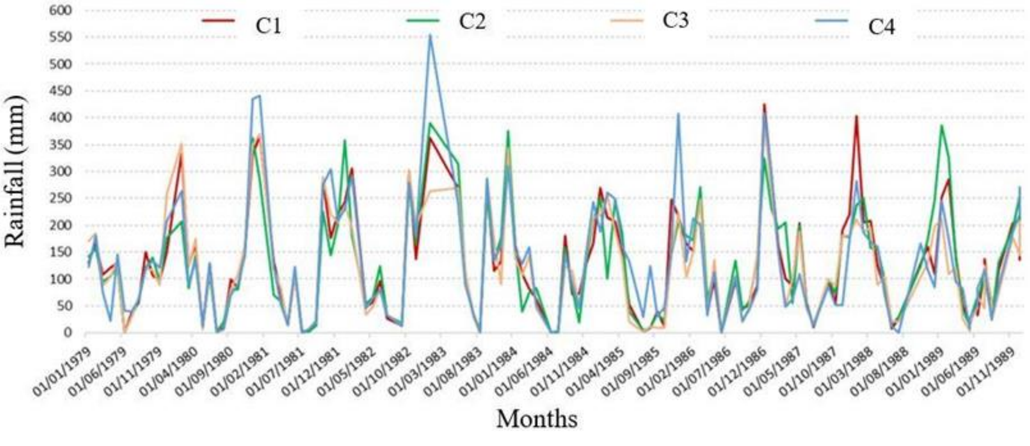

Figure 3. Time-based rainfall variation in all four investigated stations, from 1979 to 1989.

Souza and Galvani (2017) have analyzed rainfall events recorded by 16 stations located in the Jacaré Guaçu River Basin, Itirapina County, from 1968 to 1998. Results have shown that this region presents great spatial and temporal variability in rainfall events, as seen in tropical areas. The annual rainfall amplitude among the analyzed stations reached 366.6 mm, whereas the seasonal amplitude reached 70.8 mm in the lesser rainy period and 291.4 mm in the rainy period.

The full analysis of all series played a key role in helping to find likely explanations for temporal-spatial rainfall dynamics and for its main factors acting in the investigated river basin. Such representative features of each station have directly influenced results recorded for each method.

After the preliminary analysis was over, Anderson Darling and Shapiro-Wilk normality tests were applied to all real and estimated data; results of both tests have shown p-value lower than the significance level of 0.05, and it enabled concluding that the analyzed data did not follow normal distribution. Lima *et al.* (2008) have found similar results and indicated that high monthly rainfall variability is a decisive factor for such a behavior. Thus, non-parametric statistics should be used for other analyses.

3.1. Evaluating the performance of the missing data-filling method applied in Itirapina County

Tables 3, 4, 5 and 6 show the performance analysis applied to the normal ratio method, inverse distance weighting, multiple regression and neural networks, respectively, based on coefficient of determination (R^2), mean absolute error (MAE), root mean square error (RMSE), Nash-Sutcliffe efficiency coefficient (Nash), index of agreement (D), confidence index (C) and performance.

Table 3. Performance of the normal ratio method based on R^2, MAE, RMSE, Nash, D, C and performance generated for all stations, from 1979 to 1989.

Station	R^2 (%)	MAE (mm)	RMSE (mm)	Nash	D	C	Performance
C1	0.899	21.030	32.090	0.898	0.972	0.922	Excellent
C2	0.820	27.242	42.955	0.807	1.000	0.903	Excellent
C3	0.844	25.939	39.426	0.824	1.000	0.919	Excellent
C4	0.815	28.714	46.123	0.814	1.000	0.903	Excellent

Table 4. Performance of the inverse distance weighting method based on R^2, MAE, RMSE, Nash, D, C and performance generated for all stations, from 1979 to 1989.

Station	R^2 (%)	MAE (mm)	RMSE (mm)	Nash	D	C	Performance
C1	0.895	36.253	53.353	0.721	0.906	0.857	Excellent
C2	0.810	51.676	71.211	0.470	0.816	0.735	Good
C1	0.829	28.504	41.144	0.810	0.944	0.859	Excellent
C4	0.844	36.860	59.860	0.690	0.934	0.858	Excellent

Table 5. Performance of the multiple regression method based on R^2, MAE, RMSE, Nash, D, C and performance generated for all stations, from 1979 to 1989.

Station	R^2 (%)	MAE (mm)	RMSE (mm)	Nash	D	C	Performance
C1	0.906	20.271	30.987	0.896	0.975	0.928	Excellent
C2	0.843	25.296	38.768	0.813	0.956	0.878	Excellent
C1	0.869	24.390	34.169	0.849	0.964	0.899	Excellent
C4	0.931	29.558	44.369	0.795	0.946	0.913	Excellent

Table 6. Performance of the neural network method based on R², MAE, RMSE, Nash, D, C and performance generated for all stations, from 1979 to 1989.

Station	R² (%)	MAE (mm)	RMSE (mm)	Nash	D	C	Performance
C1	0.903	17.486	26.784	0.931	0.982	0.933	Excellent
C2	0.884	23.151	33.197	0.867	0.968	0.910	Excellent
C1	0.884	21.605	32.985	0.872	0.968	0.910	Excellent
C4	0.828	25.676	44.828	0.831	0.956	0.870	Excellent

All four methods used in the current research – namely: NRM, IDW, MR and NN - presented excellent performances in all stations in Itirapina County. Values recorded for coefficient of determination (R²) and Nash coefficient were higher than 0.8; except for station C4, which recorded Nash coefficient value of 0.795. The other parameters presented low error values. Based on the MR method, stations C1 and C4 have shown the best performances in R² (0.906 and 0.931, respectively). Based on NN, station C1 presented the best performance in R² and Nash coefficient, as well as the lowest error values in comparison to other methods.

Similar results were reported by Coutinho *et al.* (2018), who analyzed multiple linear regression and neural network models. According to them, linear regression methods have shown satisfactory results, high correlation indices and low mean errors in comparison to real data.

Depiné *et al.* (2014) Correia *et al.* (2016) and Wanderley *et al.* (2014) have also concluded that the neural network method has efficiently reproduced the missing rainfall data-filling process and made it possible to compare the complex inputs and outputs of simulations. However, the aforementioned authors have pointed out that there may be variations in the results due to the procedure adopted for network testing, training and validation processes. They also highlighted that estimated data are more accurate when there is lower spatial variability in rainfall events.

Table 7 shows the p-values of the Mann-Whitney test corresponding to each method and station. It was done to investigate whether there was significant difference between real and estimated data. The p-values of Kruskal-Wallis test are also shown to help identify whether there was significant difference among the proposed methods.

Table 7. p-values of Mann-Whitney (MW) and Kruskal-Wallis (KW) tests.

Station	MW-NRM	MW-IDW	MW-MR	MW-NN	KW
C1	0.024	0.024	0.788	0.633	0.018
C2	0.868	0	0.860	0.934	0.001
C3	0.569	0.459	0.864	0.980	0.380
C4	0.847	0.125	0.776	0.873	0.176

Thus, NRM was not significant for Station C1 since the p-value was lower than 0.05. With respect to the IDW method, two of the four stations did not show significant values, namely: Stations C1 (p-value 0.024) and C2 (p-value 0). The other two stations have reached p-values of 0.459 and 0.125 for Stations C1 and C3, respectively. The MR and NN methods, on the other hand, recorded p-values higher than the significance level, indicating lack of difference between real and estimated data. Based on the NN method, all stations recorded a p-value higher than 0.7, except for Station C1. Thus, it is worth emphasizing the importance of adopting in-depth non-parametric statistical methods to perform this analysis type, since the exclusive use of Nash and determination coefficients may not help in identifying such issues.

The Krukal-Wallis test has evidenced significant differences among the four methods applied to Stations C1 and C2, which recorded p-values of 0.018 and 0.001, respectively. Significant p-values were observed for the other stations, although there was no difference among the four analyzed methods. It is worth emphasizing that the neural networks subjected to KW test recorded the best performance for Station C1, whereas MR, NRM and IDW presented inferior performances; however, any of the investigated methods can be used for missing data-filling purposes.

Non-parametric analyses applied to the two stations (C1 and C2) that did not record significant p-values have indicated significant differences between actual and estimated data. This outcome can be explained by the orographic effect associated with that region, as well as by episodes of rainfall rate fluctuations that were not identified by the NRM and IDW methods, a fact that hindered the estimates.

4. FINAL CONSIDERATIONS

The missing data-filling methods used and validated through the herein-adopted statistical techniques presented excellent performances. Based on analyses applied to all four methods, neural networks and multiple regression were the ones presenting the best results; thus, they are the most suitable methods to be used for missing rainfall data-filling purposes.

Thus, based on the use of efficient missing data-filling tools, it is possible to minimize the social, environmental and economic consequences by promoting better water-resource management and operation by public and private agents, as well as by society as a whole. In addition, these tools can help mitigate issues such as flooding, drought, water supply, electric power generation, among others, enabling the use of continuous data series for these studies.

5. ACKNOWLEDGMENTS

The authors would like to thank the National Council for Scientific and Technological Development (CNPq) to the first and third author.

6. REFERENCES

AGÊNCIA NACIONAL DE ÁGUAS (Brasil). **Website.** Available at: http://www3.ana.gov.br/portal/ANA/panorama-das-aguas/quantidade-da-agua. Access: February 15th, 2020.

ALVES, E. D. L.; BIUDES, M. S.; VECCHIA, F. A. Z. Interpolação espacial na climatologia: análise dos critérios que precedem sua aplicação. **Geonorte**, v. 3, n. 8, p. 606–618, 2012.

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

CEPAGRI. **Clima dos Municípios Paulistas**. Available at: http://www.cpa.unicamp.br/outras-informacoes/clima_muni_272.html. Access: October 20th, 2020.

CORREIA, T. P. *et al.* Aplicação de redes neurais artificiais no preenchimento de falhas de precipitação mensal na região serrana do Espírito Santo. **Geociências**, v. 35, n. 4, p. 560-567, 2016.

COUTINHO, E. R. *et al.* Application of artificial neural networks (ANNs) in the gap filling of meteorological time series. **Revista Brasileira de Meteorologia**, v. 33, n. 2, p. 317-328, 2018. https://doi.org/10.1590/0102-7786332013

DEPINÉ, H. *et al.* Preenchimento de falhas de dados horários de precipitação utilizando redes neurais artificiais. **Revista Brasileira de Recursos Hídricos**, v. 19, n. 1, p. 51-63, 2014.

DI PIAZZA, A. *et al.* Comparative analysis of different techniques for spatial interpolation of rainfall data to create a serially complete monthly time series of precipitation for Sicily, Italy. **International Journal of Applied Earth Observation and Geoinformation**, v. 13, n. 3, p. 396-408, 2011. https://doi.org/10.1016/j.jag.2011.01.005

ESPINOSA, M. M.; CALIL JÚNIOR, C.; LAHR, F. A. R. Métodos paramétricos e não paramétricos para determinar o valor característico em resultados de ensaio de madeira. **Scientia Forestalis**, n. 66, p. 76-83, 2004.

FANTE, K. P.; SANT'ANNA NETO, J. L. Técnicas estatísticas para a homogeneização de dados de temperatura em séries temporais climatológicas. **Revista Brasileira de Climatologia**, v. 18, p. 143-156, 2016. http://dx.doi.org/10.5380/abclima.v18i0.43202

GOYAL, M. K. Monthly rainfall prediction using wavelet regression and neural network: an analysis of 1901–2002 data, Assam, India. **Theoretical and Applied Climatology**, v. 118, n. 1-2, p. 25-34, 2014. https://dx.doi.org/10.1007/s00704-013-1029-3

HUBBARD, K. G. Spatial variability of daily weather variables in the high plains of the USA. **Agricultural and Forest Meteorology**, v. 68, n. 1, p. 29-41, 1994. https://doi.org/10.1016/0168-1923(94)90067-1

IBGE. **Censo 2010**. Available at: https://cidades.ibge.gov.br. Access: September 10th, 2019.

JACKSON, E. K. Introductory overview: Error metrics for hydrologic modelling – A review of common practices and an open source library to facilitate use and adoption. **Environmental Modelling & Software**, v. 119, p. 32-48, 2019.

JUNQUEIRA, R.; AMORIM, J. S.; OLIVEIRA, A. S. Comparação entre diferentes metodologias para preenchimento de falhas em dados pluviométricos. **Sustentare**, v. 2, n. 1, p. 198-210, 2018.

KHOSRAVI, G. *et al.* A Modified distance-weighted approach for filling annual precipitation gaps: application to different climates of Iran. **Theoretical and Applied Climatology**, v. 119, n. 1-2, p. 33-42, 2015. https://dx.doi.org/10.1007/s00704-014-1091-5

KRUSKAL, W. H.; WALLIS, W. A. Use of ranks in on-criterion variance analyses. **Journal of the American Statistical Association**, v. 47, n. 260, p. 583-621, 1952.

LIMA, J. S. de S. *et al.* Variabilidade temporal da precipitação mensal em Alegre – ES. **Revista Ciência Agronômica**, v. 39, n. 02, p. 327-332, 2008.

NAGHETTINI, M.; PINTO, E. J. A. **Hidrologia Estatística**. Belo Horizonte: Serviço Geológico do Brasil-CPRM, 2007. 381p.

OLIVEIRA, L. F. C. *et al.* 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.

PEREIRA, D. R. *et al.* Hydrological simulation using SWAT model in a headwater basin in southeast Brazil. **Engenharia Agrícola**, v. 34, n. 4, p. 789-799, 2014. https://doi.org/10.1590/S0100-69162014000400018

SANTOS, B. F.; TREVISAN, D. P.; MOSCHINI, L. E. Avaliação da vulnerabilidade ambiental do município de Itirapina – SP. **Geo Temas**, v. 8, n. 1, p. 42-59, 2018a. https://doi.org/10.33237/geotemas.v8i1.2822

SANTOS, B. C.; SANCHES, R. G.; SOUZA, P. H. A dinâmica atmosférica no verão 2013 – 2014 no município de Itirapina/SP e sua caracterização pluviométrica utilizando anos padrões. **Caminhos de Geografia**, v. 19, n. 68, p. 1-18, 2018b. https://doi.org/10.14393/RCG196801

SILVA, C. E. F. *et al.* **Plano de manejo integrado das unidades de Itirapina.** 2006. Available at: https://smastr16.blob.core.windows.net/iflorestal/2013/03/Plano_de_Manejo_EEc_Itirapina.pdf. Access: November 14th, 2020.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samples). **Biometrika**, v. 52, p. 591-611, 1965.

SHEPARD, D. A two-dimensional interpolation function for irregularly spaced data. *In*: NATIONAL CONFERENCE OF THE ASSOCIATION FOR COMPUTING MACHINERY, 23., 1968. **Proceedings**… New York: ACM, 1968. p. 517-524.

SOUZA, V.; GALVANI E. Distribuição Espaço Temporal Da Precipitação Pluvial E Sua Interação Com O Relevo Na Bacia Do Rio Jacaré Guaçu (SP). **Ciência e Natura**, v. 39, p. 110-124, 2017. https://dx.doi.org/10.5902/2179460X27334

TEEGAVARAPU, R. S. V.; CHANDRAMOULI, V. Improved weighting methods, deterministic and stochastic data-driven models for estimation of missing precipitation records. **Journal Of Hydrology**, v. 312, n. 1-4, p. 191-206, 2005. https://doi.org/10.1016/j.jhydrol.2005.02.015

TRIOLA, M. **Introdução a estatística**. Rio de Janeiro: LTC, 2008. 722 p.

WANDERLEY, H. S.; AMORIM, R. F. C.; CARVALHO, F. O. Interpolação espacial de dados médios mensais pluviométricos com redes neurais artificiais. **Revista Brasileira de Meteorologia**, v. 29, n. 3, p. 389-396, 2014. https://doi.org/10.1590/0102-778620130639

YOUNG, K. C. A Three-way model for interpolating monthly precipitation values. **Monthly Weather Review**, v. 120, n. 11, p. 2561-2569, 1992. https://doi.org/10.1175/1520-0493(1992)120<2561:ATWMFI>2.0.CO;2