Revista Brasileira de Climatologia
Brazilian Journal of Climatology

OPEN ACCESS

DOI: 10.5380/abclima

FILLING OF FAULTS IN CLIMATOLOGICAL AIR TEMPERATURE SERIES IN BRAZILIAN STATE CAPITALS FROM 1980 TO 2017

PREENCHIMENTO DE FALHAS NAS SÉRIES CLIMATOLÓGICAS DA TEMPERATURA DO AR NAS CAPITAIS BRASILEIRAS DE 1980 ATÉ 2017

RELLENO DE FALLAS EN LAS SERIES CLIMATOLÓGICAS DE LA TEMPERATURA DEL AIRE EN LAS CAPITALES DE LOS ESTADOS BRASILEÑOS DESDE 1980 HASTA 2017

Marcele de Jesus Correa
Universidade Federal do Rio Grande do Norte
marcelejc.marinho@gmail.com

Kellen Carla Lima
Universidade Federal do Rio Grande do Norte
Kellen.lima@ufrn.br

Jonathan Mota da Silva
Universidade Federal do Rio Grande do Norte
jmotasilva@gmail.com

Gilvandro César de Medeiros
Universidade Federal do Rio Grande do Norte
gilvandrocesar@ufrn.edu.br

Abstract: Air temperature is a key variable used to assess climate change. It is essential for many applications and impact studies in science. However, missing values in observed air temperature time series are quite common, which jeopardize its use for climate studies. In order to fill in missing maximum and minimum temperatures data from 21 stations from 1980 to 2017 in the main Brazilian capitals, we used three models: multiple linear regression (MRL), artificial neural network (ANN), and the autoregressive integrated moving average (ARIMA). In the annual averages, the ANN and MLR models presented a better performance in filling the missing data as compared to the ARIMA model,

especially for maximum temperature. Seasonally, ANN overestimated the maximum and minimum temperatures, but it and the MLR model presented the best results (R^2> 0.7) for all seasons, except winter. The ANN was the most suitable model to fill the missing data of maximum and minimum temperatures, even though it could be improved with the increase of the training on its networks. This study contributes to the understanding of essential methodologies for the use of climatic time series.
Keywords: Climatological Series. Neural networks. Arima. Multiple Regression.

Resumo: A temperatura do ar é uma variável chave usada para avaliar as mudanças climáticas. **Esta** variável é essencial para muitas aplicações e estudos de impacto na ciência. No entanto, falhas nos registros de séries temporais de temperatura do ar observadas são bastante comuns, o que prejudica **o** seu uso para estudos climáticos. A fim de preencher a ausência dos dados de temperaturas máximas e mínimas de 21 estações de 1980 a 2017 das principais capitais Brasileiras, nós utilizamos três técnicas: regressão linear múltipla (MRL), rede neural artificial (RNA) e a média móvel integrada autorregressiva (ARIMA). Nas médias anuais os modelos ANN e MLR apresentaram uma melhor destreza no preenchimento de falhas do que o ARIMA, especialmente para temperatura máxima. Sazonalmente, **a** RNA superestimou as temperaturas máximas e mínimas, contudo, este modelo e o MLR apresentaram os melhores resultados (R^2 > 0,7) para todas as estações, exceto o inverno. O RNA foi o modelo mais indicado para preencher os dados faltantes de temperaturas máximas e mínimas, embora este modelo precise ser aperfeiçoado com o aumento do treinamento das suas redes. Este estudo contribui para o entendimento de metodologias essenciais para o uso de séries temporais climáticas.
Palavras-chave: Séries Climatológicas. Redes Neurais. Arima. Regressão Múltipla.

Resumen: La temperatura del aire es una variable clave para evaluar el cambio climático. Es esencial para muchas aplicaciones y estudios de impacto en la ciencia. Sin embargo, las lagunas en los registros de las series temporales de temperatura del aire observadas son bastante comunes, lo que dificulta su uso para los estudios climáticos. Para suplir la ausencia de datos de temperatura máxima y mínima de 21 estaciones de 1980 a 2017 de las principales capitales brasileñas, utilizamos tres técnicas: regresión lineal múltiple (MLR), red neuronal artificial (ANN) y media móvil integrada autorregresiva (ARIMA). En los promedios anuales, los modelos RNA y MLR mostraron una mayor destreza a la hora de rellenar los huecos que el ARIMA, especialmente para la temperatura máxima. Estacionalmente, la RNA sobrestimó las temperaturas máximas y mínimas, sin embargo, este modelo y la RML presentaron los mejores resultados (R^2 > 0,7) para todas las estaciones excepto el invierno. La RNA fue el modelo más adecuado para rellenar los datos que faltaban de las temperaturas máximas y mínimas, aunque este modelo debe mejorarse aumentando el entrenamiento de sus redes. Este estudio contribuye a la comprensión de las metodologías esenciales para el uso de las series temporales climáticas.
Palavras-chave: Series Climatológicas. Redes Neurais. Arima. Regresión múltiple.

Submetido em: 10/09/2020
Aceito para publicação em: 15/05/2021
Publicado em: 22/09/2021

INTRODUCTION

Climatological data and the information obtained from these have great relevance for different anthropogenic activities, because they contribute to decision-making in areas such as water resources, agrometeorology, urban climate, among others. Thus, working with complete climatological series, that is, without failures, allows the researcher to establish non-contradictory or erroneous conclusions. However, complete climatological series is not the case in Brazil, due to some technical/operational problems such as the observer's absence, instrumental failures, a break in the line of communication or geographical location for example, which can lead to incorrect interpretations if not corrected. That is, the presence of many faults in the climatological series interfere with the results found, generating problems of interpretation in the data. (KASHANI and DINPASHOH, 2012; DANTAS, SANTOS and OLINDA, 2016; BIER and FERRAZ, 2017). Thus, it is important to use methodologies that are able to estimate values that correspond to the other values present in the meteorological series of interest.

Different fault filling techniques, as well as the validation of their consistency, are used to estimate the missing data, being statistical methods the most used ones with the multiple linear regression being highlighted (KEMP et al., 1983; TABONY, 1983; XIA et al., 1999; OLIVEIRA et al., 2010, KASHANI and DINPASHOH, 2012; LEE and KANG, 2015; DANTAS, SANTOS and OLINDA, 2016; BIER and FERRAZ, 2017). This is considered one of the simplest methods because it allows the correlation of the station with failures with neighboring stations, in which the weight associated with each observed data will be determined by means of partial correlation, which is the basis of linear regression; or total, which is capable of reducing random but not systematic errors and requires the use of large numbers of neighboring stations (OLIVEIRA et al., 2010; BIER and FERRAZ, 2017).

Another approach that has been widely used for the reconstruction of incomplete time series is the use of computational intelligence known as artificial neural networks (ANN), which is a technique inspired by the neural structure of intelligent organisms that recognizes patterns and generalizes information (COUTINHO et al., 2018). That is, it is a technique that seeks to simulate the functioning between the human brain and the complex biological neural system (CORREIA et al., 2016). However, when working with ANN there is the difficulty in finding the best structure of the network that will satisfy the model for estimating the missing

values, which usually consists of investigating a whole space of possible states (VENTURA et al., 2013). Due to certain limitations such as physical relationships based on the experience and preferences of those who use, study and train networks have made statistical methods based on linear relationships more used (SHARMA, RAI and DEV, 2012; LEE and KANG, 2015).

In recent years, one of the most popular approaches for time series modeling is the Autoregressive Integrated Moving Averages (ARIMA), which goal is to carefully and rigorously investigate past observations of a time series to develop appropriate models able to predict future values for the series with missing data (EL-MALLAH and ELSHARKAWY, 2016; MURAT et al., 2018). In this way, this method only needs the previous data from a time series to perform the prediction, which causes the model to improve the accuracy of the prediction while reducing the number of parameters to a minimum. However, the disadvantage of using this technique is to choose the most appropriate models to identify which will be the most feasible one to estimate the missing data from time series. Another disadvantage is that this is not a good forecast model for long-term series (WADI et al, 2018).

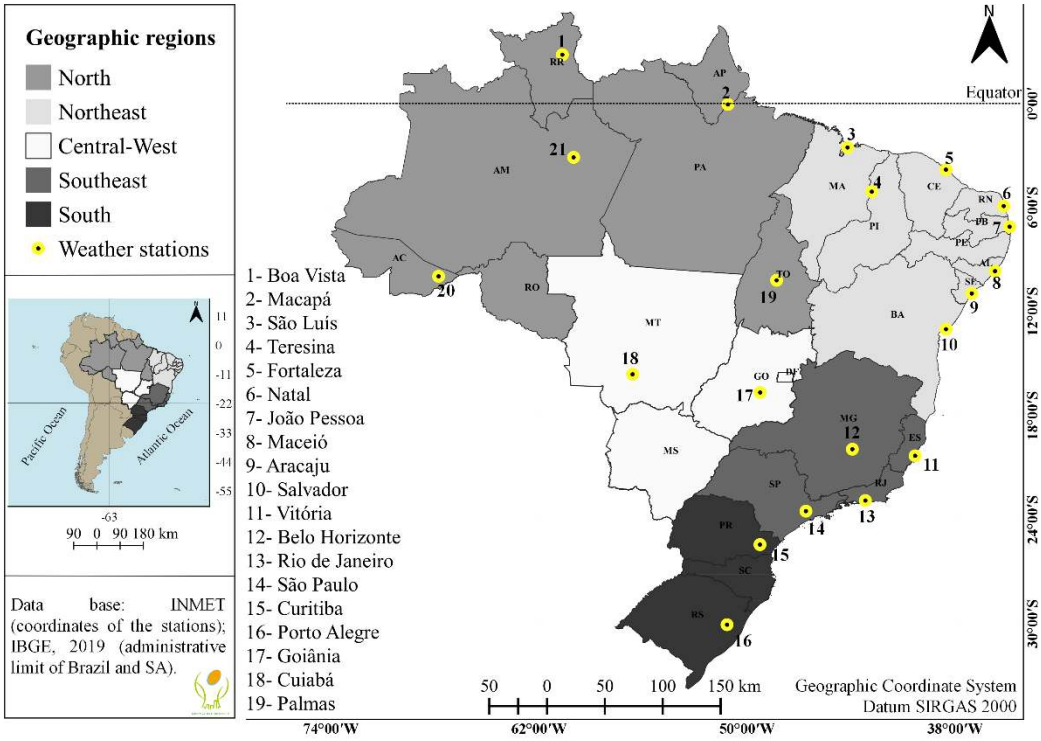
Therefore, this work aimed to assess the accuracy of statistical models in the reconstruction of climatological series with a large number of failures. To this, we evaluated three techniques of distinct approaches applied to climatological series of maximum and minimum monthly air temperature in 17 meteorological stations located in state capitals of Brazil for a period of 37 years (1980-2017).

MATERIALS AND METHODS

Study Area

Data from climatological series of 21 meteorological stations located in different geographic regions of Brazil (Figure 1) were used in order to verify the percentage of failures and fill them out from different statistical models. With a territorial area greater than 8.5 million km², Brazil has different climatic types when considering its territorial area in the north-south direction, where latitudes range from +5° to -33°, respectively (IBGE, 2019).

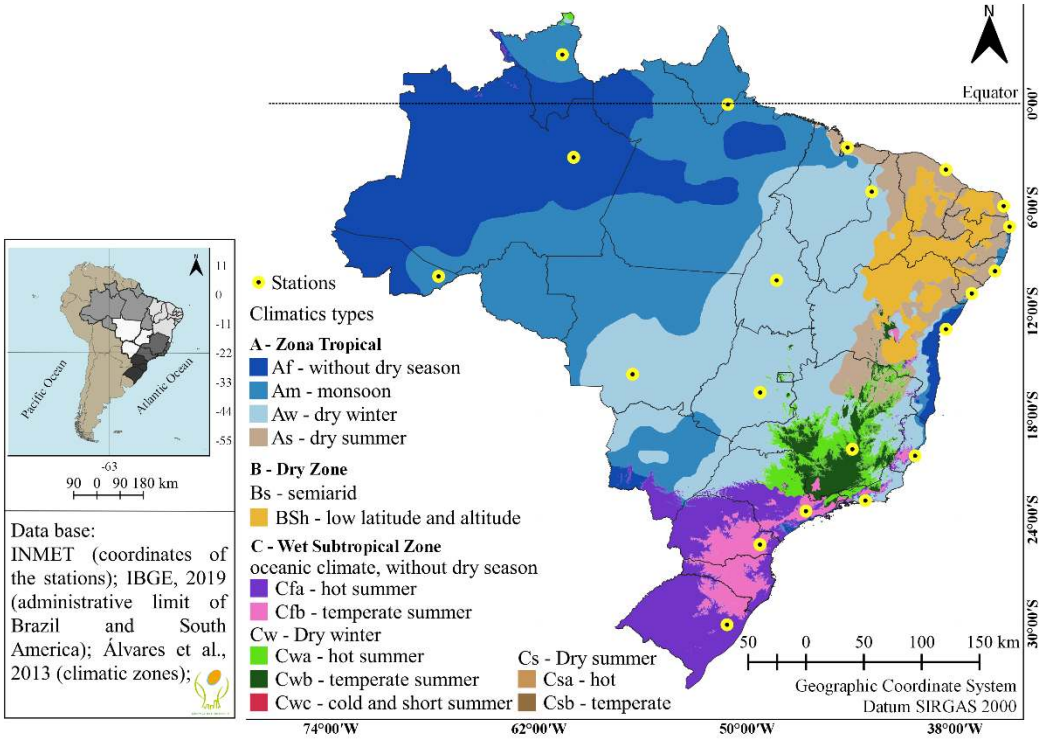
Figure 1 - Location of the weather stations in relation to the geographical regions of Brazil.



Source: Elaborated by the authors (2020)

The Figure 2 show that the stations are localized in different climatic zones, according to the Köppen climate classification, and that Brazil has three climatic zones: Tropical, Dry and Subtropical Humid and 12 more types of climates; i.e, the Tropical zone occurs in a large part of the country because in these areas there are no limiting factors of altitude, precipitation and temperature to impose other climatic factors. The dry zone, represented by the semi-arid climate, is notably the typical climate of northeastern Brazil, occurring basically in landscapes where annual rainfall is less than 800 mm. The subtropical zone is typical of the southern region, covering 13.7% of the Brazilian territory, well represented by plateaus and mountains (ÁLVARES et al., 2013).

Figure 2 - Location of the weather stations in relation to weather zones.



Source: Elaborated by the authors (2020)

Time series

Data of the monthly averages for maximum and minimum air temperature corresponding to the period of 38 years (01/01/1980 to 31/12/2017) for each of the 21 stations were used. More information on the characteristics of the stations is found in Table 1. The time series were obtained from the Meteorological Database for Teaching and Research of the National Institute of Meteorology (BDMEP/INMET), referring to the meteorological stations located in the state capitals of Brazil.

The filling of gaps in the INMET time series by the statistical techniques of Multiple Liner Regression and Neural Networks was performed using interpolated data in a grid point with spatial resolution of 0.25 x 0.25 by Xavier et al. (2016), such as the monthly averages of minimum and maximum air temperature (°C), relative air humidity (%), radiation ($MJ.m^{-2}$) and wind speed ($m.s^{-1}$), which served as independent variables (or input) for the training of statistical models in order to estimate the values for the gaps present in the investigated time series. It is highlight that the grid of a point refers to a geographical representation of the world as a matrix of cells organized in rows and columns. Each cell in the grid is referenced by its geographical location X and Y, for example the air temperature variable.

Table 1 - Identification of the 21 weather stations according to the location in the five geographical regions

Região Geográfica	Município	Latitude (°)	Longitude (°)	Altitude (m)
Norte	Boa Vista	2,82	-60,66	83
	Macapá	-0,05	-51,11	14,46
	Manaus	-3,1	-60,01	61,25
	Palmas	-10,19	-48,3	280
	Rio Branco	-9,96	-67,8	160
Nordeste	Aracaju	-10,95	-37,04	4,72
	Fortaleza	-3,81	-38,53	26,45
	João Pessoa	-7,1	-34,86	7,43
	Maceió	-9,66	-35,7	64,5
	Natal	-5,91	-35,2	48,6
	Salvador	-13,01	-38,53	51,41
	São Luís	-2,53	-44,21	50,86
	Teresina	-5,08	-42,81	74,36
Centro-Oeste	Cuiabá	-15,61	-56,61	145
	Goiânia	-16,66	-49,25	741,48
Sudeste	Belo Horizonte	-19,93	-43,93	915
	Rio de Janeiro	-22,89	-43,18	11,1
	São Paulo	-23,5	-46,61	792,06
	Vitória	-20,31	-40,31	36,2
Sul	Curitiba	-25,43	-49,26	923,5
	Porto Alegre	-30,05	-51,16	46,97

Source: Elaborated by the authors (2020)

Methodology

Data analysis

The time series of the monthly averages of the air temperature, maximum and minimum, were analyzed for to quantify the number of faults existing in each of the 21 stations. For this, the "mstats" function of the "mtsdi" package present in version 4.0.2 of software R was used. This function calculates the proportion of missing observations in a given data set by rows and columns (R Core Team, 2020). After their identification, these flaws were filled with the use of mathematical models: multiple linear regression, artificial neural networks and the Box-Jenkins approach, also known as the Autoregressive Integrated Moving Averages model (ARIMA).

Multiple Linear Regression (MLR)

The model is an extension of a linear model that determines an association between a dependent variable and two or more independent variables according to Equation 1:

$$\hat{Y} = b_0 + b_1X_1 + b_2X_2 + \cdots + b_kX_k + e_1, \qquad (1)$$

Where X_1, X_2, X_k are the predictor / independent variables; b_0 and b_k are the linearity and regression coefficients, respectively; and e_i error.

The regression coefficients of the model were obtained from the least squares method and its validation followed the assumptions of a linear regression (LEECH et al., 2003; KURTNER et al., 2004; WILKS, 2006). MLR is usually used to estimate rainfall and air temperature not recorded over a climatological series in order to fill in the gaps that have occurred (TABONY, 1983; KASHANI and DINPASHOH, 2012; FANTE and NETO, 2013; BIER and FERRAZ, 2017). In this work, the MLR used the database developed by Xavier et al. (2014) due to the absence of weather stations close to INMET stations, as explained in section 2.2.

Artificial Neural Networks (ANN)

Artificial neural networks (ANN) aim to involve adaptive mechanisms that allow computers to learn and replicate a certain pattern from examples present in a dataset (NEGNEVITSKY, 2011). Thus, ANN are also formed by neurons (basic processing units) interconnected and propagating signals between them, allowing the transmission of information depending on the stimuli (input variables and signals) received. One of the simplest ANN architectures is perceptron, based on a mathematical formulation developed in 1957 by Frank Rosenblatt (GÉRON, 2017).

The input layer has 7 neurons due to the 7 input variables (data from INMET weather stations, in which the amount of hidden layers as well as the amount of neurons in each of these layers was arbitrated and with one of hyperbolic tangent activation, Equation 2, similar to the one performed by Coutinho et al., 2018.

$$f_{(u_i)} = tgh\left(\frac{u_i}{2}\right) = \frac{1 - \exp(-u_i)}{1 + \exp(-u_i)} \qquad 2$$

Neural Networks were applied (modeling and training) with an MLP algorithm implemented in the library *TensorFlow in* Python 3, and with the use of the mean square error as a cost function, for the daily data of maximum and minimum temperatures (01/01/1980 à 31/12/2017) for all cities evaluated. For the same ANN training period, seven more variables were used (heat stroke, tar evaporation, average compensated temperature, average humidity and relative humidity, average wind speed and precipitation) of the respective INMET maximum and minimum temperature weather stations. These additional variables were normalized in such a way that they started to vary from -1 to 1, with zero being attributed to nonexistent data. In general, the architecture of the neural network was structured with a seven-dimensional input layer, consisting of these variables, and with two hidden layers.

The activation function employed in these layers was the hyperbolic tangent with seven neurons per layer, resulting in an ANN with an output layer and with a single neuron. For training, 48 neural networks were used, divided equally between the maximum and minimum temperatures for all cities. In this process, 200 interactions / epochs were carried out for each ANN, with the same architecture being preserved in all networks. Thus, a single response was generated, which corresponded to the maximum and minimum temperatures estimated based on the combination of the seven additional variables.

Box-Jenkins Model (ARIMA)

The time series modeling through ARIMA aims to carefully analyze and rigorously process past observations of the climatological series in order to develop an appropriate model that can describe the structure inherent to the series of interest for the prediction of missing data, since the ARIMA model is a statistical method used to decompose and predict time series data, modeling the correlations in the data (MURAT et al., 2018; WADI, ALMASARWEH and ALSARAIREH, 2018).

The ARIMA methodology takes into account the seasonality of the time series, however, most series are non-stationary and it is necessary to apply a certain number of differences (d) between the data. Thus, the Self-Regressive (AR), Integrated (I) models, which is the number of differentiations, and the Moving Averages (MA) are processes that represent the ARIMA model from the p, d and q orders, respectively. For the seasonal ARIMA model, this can be represented from the insertion of the seasonal operator to the model and

represented by SARIMA (p, d, q)/(P, D, Q), according to Yodah et al. (2013) and Camelo et al. (2017). In order to predict the missing data, the "forecast" package (R Core Team, 2017) was used.

The structure of the ARIMA model was based on the data from the INMET station time series, considering the iterative cycle of the stages below, according to Morettin and Toloi (2006):

(i) A general class of models is considered for analysis and specification, for example, transforming the non-stationary series (when applicable) into stationary by differentiation. The Dickey-Fuller and Philips-Perron tests were used to assess the stationarity of the climatological series (GUJARATI and PORTER, 2009), in which the hypothesis decision-making was based on the p-value for the 5% significance level;

Dickey-Fuller and Philips-Perron tests

$H0$: the series is non-stationary (it has at least one unit root).

$H1$: the series is stationary (does not have a unit root).

Rejection rule: if the p-value is less than α, that is, p <0.05, (rejects $H0$).

The model was also identified based on the analysis of the autocorrelation (FAC) and partial autocorrelation (FACP) functions and other criteria, such as those of Akaike (AIC), English Akaike Information Criterion and Bayesian (BIC), Bayesian Information Criterion, both (Equation 3 and 4) consider the lowest value among the compared models, for:

$$AIC = log\ \hat{\sigma}_k^2 + \frac{n+2k}{n}, \qquad (3)$$

$$AIC = log\ \hat{\sigma}_k^2 + \frac{k\ log\ n}{n}, \qquad (4)$$

Where $\hat{\sigma}_k^2$ is obtained by $\frac{SSE\ (k)}{n}$, for $SSE\ (k)$ equal to the sum of the squared residuals of the model with the number of regression coefficients k and, n the sample size. (ii) Then, the estimation of the model parameters based on the values of the AIC and BIC criteria of the above equations, respectively; (iii) The verification or diagnosis of the adjusted model was done by means of an analysis of the residues, in order to verify the adequacy for the prediction of the missing data in the climatological series using the statistical tests of Shapiro-Wilk, Durbin-Watson and Breusch-Pagan for normal , independence and homoscedasticity of the data, respectively, in which the decision-making took into account the p-value of each test

performed, according to Camelo et al.

Model adjustment measures

The performance of the models was evaluated based on statistical metrics to identify the accuracy of the models proposed in filling out failures. The following statistical error metrics were used: Mean Error (ME), or Bias, which measures the tendency of the model to overestimate or underestimate the simulated temperature in relation to observational values (Equation 5) according to Hallak and Pereira Filho (2011). Mean Absolute Error (MAE), which measures the average error value between the observed and simulated series. (Equation 6), is less affected by extreme values or outliers (DÉQUÉ, 2003; BIER and FERRAS, 2017; CAMEL, 2017). While the Coefficient of Determination (R^2) is the square of the sample correlation coefficient (r) and is considered one of the ways to evaluate the adjustment of the model (Equation 7), according to Martins (2018) and Nogueira et al. (2020).

Mean Error (ME)

$$ME = BIAS = \bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i \quad (5)$$

Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |V_{est_i} - V_{obs_i}| \quad (6)$$

$$R^2 = \frac{SQR}{SQT} = 1 - \frac{SQE}{SQT} = \frac{\hat{\beta}_1 \sum_{i=1}^{n}(x_i - \bar{x})\, Y_i}{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}, \quad (7)$$

being *n* is the number of years observed, which in this case corresponds to the period from 1980 to 2017; V_{est_i} the value of the estimate in each month by the proposed models and V_{obs_i} is the observed value (INMET). SQT measures the variation of Y, independent of X, and SQE measures the variation of Y, considering the variable X in the regression model; x_i is the predictor variable and Y_i is the predicted variable. The *ME* and *MAE* errors have their values in °C, while R^2 is $0 \leq R^2 \leq 1$.

Spatial interpolation of model adjustment measures

The spatialization maps of the metrics were made by interpolating the values obtained from the adjustment measures of each of the 21 stations. Thus, the interpolation through the Inverse of Distance Weighted (IDW) was used, in which the samples of the points are weighed during the process according to the influence of one point relative to another with the distance, from an unknown point that one wants to create, that is, being considered one of the most used interpolation techniques for spatially distributed points and by assigning greater weight to the nearest point, decreasing this weight with the increase of distance and as a function of the coefficient power α (FILHO et al., 2019; MURARA, 2019).

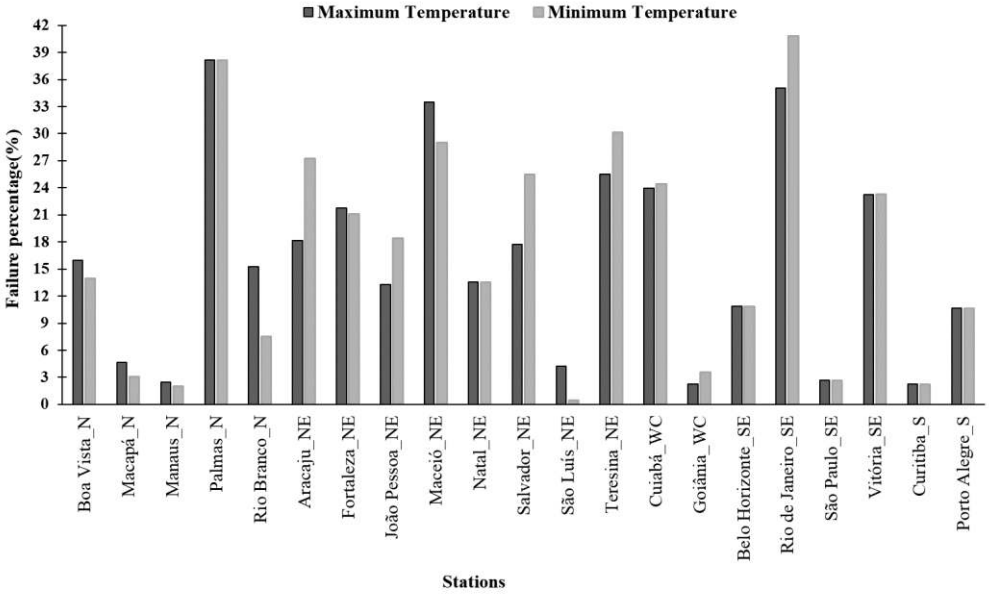
The maps of the measurements adjustment for the maximum and minimum air temperatures for the annual and seasonal series of meteorological stations in the 21 state capitals of Brazil consisted of geostatistical analysis generated from the software Qgis v. 3.10.7, using the Geographic Coordinate System and datum SIRGAS 2000. The input data were the location coordinates of the stations with the values of the adjustment measures.

RESULTS AND DISCUSSION

Data failure

Figure 3 shows the percentage of the existing faults in the studied air temperature series of each station by geographic region. These failures were verified using the *mstats* function of the package *mtsdi* package from R software's (R Core Team, 2017). It was observed that the stations with the highest percentage of failures in the maximum temperature series were Cuiabá, Vitória, Teresina, Maceió, Rio de Janeiro and Palmas, with values ranging between 22% and 39. For the minimum temperature, the stations of Vitória, Cuiabá, Salvador, Aracaju, Maceió, Teresina, Palmas and Rio de Janeiro presented a percentage of failures ranging from 21% to 41%. The high percentage of these failures can be explained due to the collection made in a conventional way, performed through the annotation of the data by an individual collector, which at a given moment can be deficient, as pointed out by Diaz et al. (2018) and Machado and Assis (2018). The stations that presented a percentage of failures below 5% for both maximum and minimum temperatures were Macapá, Manaus, São Luís, Goiânia, São Paulo and Curitiba.

Figure 3 - Distribution (%) failures of the 21 weather stations



Source: Elaborated by the authors (2020)

It is noted that the geographics regions that concentrate the highest percentage of failures for both maximum and minimum air temperature are the Northeast and Southeast regions, being represented by the Maceió and Rio de Janeiro stations, respectively.

Spatial analysis of model adjustment measures

Annual time series from 1980 to 2017

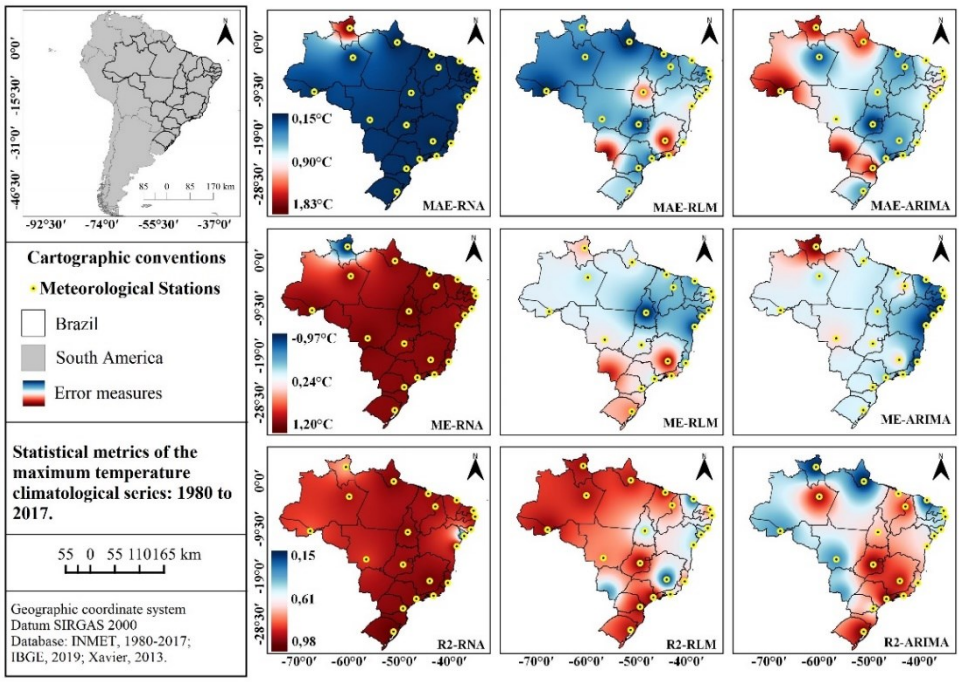
The Figure 4 shows the spatialization of statistical metrics between 1980 to 2017 (mean error - ME, mean absolute error - MAE and coefficient of determination - R^2) applied to evaluate the performance of the methods of Artificial Neural Network (ANN), Multiple Linear Regression (MLR) and Autoregressive Integrated Moving Averages (ARIMA) in filling the faults observed in the maximum air temperature for the 21 stations. The MLR model correlates the corresponding observations with the grid point data. The RNA model, on the other hand, makes an alternative computational approach inspired by studies of the brain and nervous system (HAYKIN, 2008), that is, it is a method that has the storage and processing structures of the biological nervous system as a basis for estimating the missing data (LEAHY et al., 2008). While the estimate made by ARIMA decomposes and predicts the gaps in time series by modeling the correlations in the series data itself (WADI et al., 2018). One can observe that for the absolute mean error, the artificial neural networks method presented an

error below 0.5°C in almost all seasons studied. The exception was Boa Vista in the extreme end of the North region, where absolute mean error was approximately 1.83°C. Regarding the Multiple Regression technique, highest errors were in Palmas, Belo Horizonte and Rio de Janeiro, where errors ranged between 0.9°C and 1.83°C. While in the ARIMA model, the smallest errors were concentrated in the Northeast, Central-West and Southeast regions, ranging from 0.15°C to 0.9°C.

For the mean error, the performance of the technique by neural networks was low, since only the Boa Vista station presented the lowest error when compared to the other stations. For Multiple Regression, the greatest errors were concentrated in the South and Southeast regions of Brazil; while ARIMA presented the greatest error at Boa Vista station. Regarding the coefficient of determination (R^2), the ANN technique presented an adjustment greater than 0.61 (61%) in almost all stations evaluated, i.e., an adjustment of 98% (0.98) between the simulated data and the observed time series. For the MLR, the best adjustment was in the North and South regions of the south of the country (0.98), while in the other seasons the coefficient varied between 0.15 and 0.61 (15% and 61%), respectively.

Unlike ARIMA, where model fits best to the sample in most of the Northeast, Southeast and South regions had values above 0.7 (70%). It is observed that in the east of northeastern Brazil the coefficient was almost 15% (0.15), that is, low adjustment between the model and the sample. Figure 4 shows that ANN was the best model in the three adjustment measures, followed by MLR and ARIMA. A similar result was found by Afrifa-Yamoah et al. (2020) when filling time series of temperature, humidity and wind speed using MLR and ARIMA techniques for 4 stations in Western Australia. The authors verified that the MLR model presented better performance in filling out failures than ARIMA.

Figure 4 - Spatialization of statistical metrics (mean error - ME, mean absolute error - MAE and coefficient of determination - R^2) for the period from 1980 to 2017 of the maximum temperature applied to evaluate the performance of artificial neural network (ANN), multiple linear regression (MLR) and the Autoregressive Integrated Moving Averages (ARIMA).



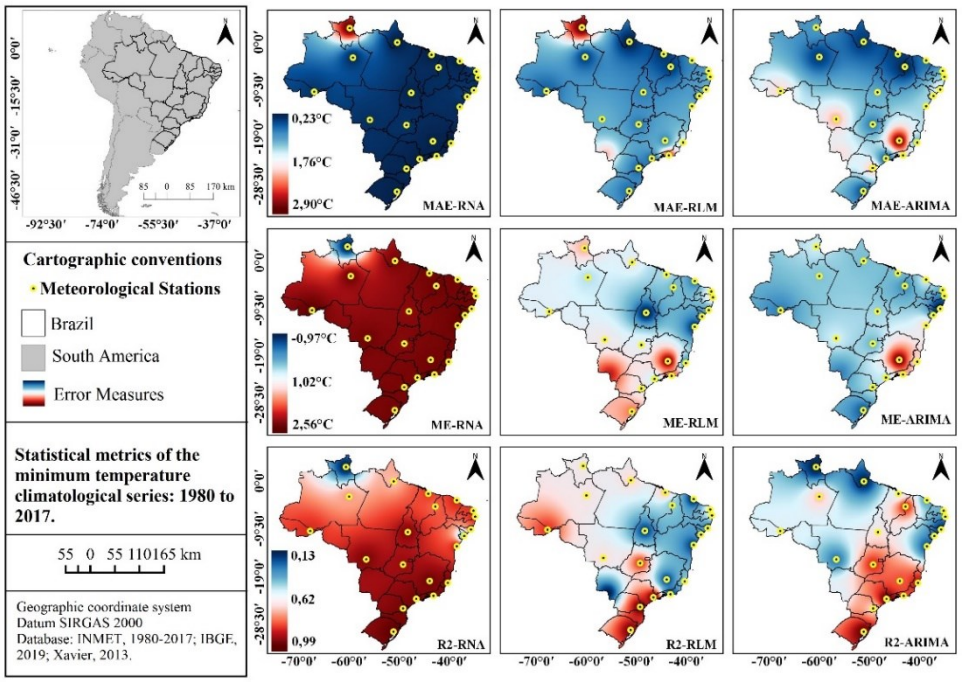
Source: Elaborated by the authors (2020)

Figure 5 shows the spatialization of statistical metrics (mean error - MS, absolute mean error - EMA and coefficient of determination - R^2) for the period from 1980 to 2017 of the minimum temperature. It is observed that the spatialization of minimum temperature errors is similar to that of the maximum temperature, however the values of the mean and average absolute errors are higher for the minimum temperature, which can be explained by the percentage of failures being higher for this temperature in almost all seasons than for the maximum temperature, causing a biased result in the model, that is, overestimated data, as pointed out by Hallak and Pereira Filho, 2011. The coefficient of determination was above 62% (0.62) for the North and Northeast regions and approximately 99% (0.99) for the Central-West, South and Southeast regions. For the MLR technique, the absolute mean error was below 1.76°C for almost all stations, except for Boa Vista, which had an error above 2°C; while the mean error varied between 1.02°C and 2.56°C in the Midwest, South and Southeast regions, mainly.

It was also observed that the coefficient of determination was low in the Northeast and high in the South region. The Box-Jenkins (ARIMA) model showed absolute mean error

ranging from 0.23°C to 1.76°C and mean error below 1.0°C in almost all seasons, and coefficient of determination greater than 62 (0.62) at the stations of São Luís, Teresina, Goiânia, and other stations located in southern and southeastern Brazil.

Figure 5 - Spatialization of statistical metrics (mean error - MS, mean absolute error - MAE and coefficient of determination - R^2) for the period from 1980 to 2017 of the minimum temperature applied to evaluate the performance of artificial neural network (ANN), multiple linear regression (MLR) and the Autoregressive Integrated Moving Averages (in English, ARIMA).

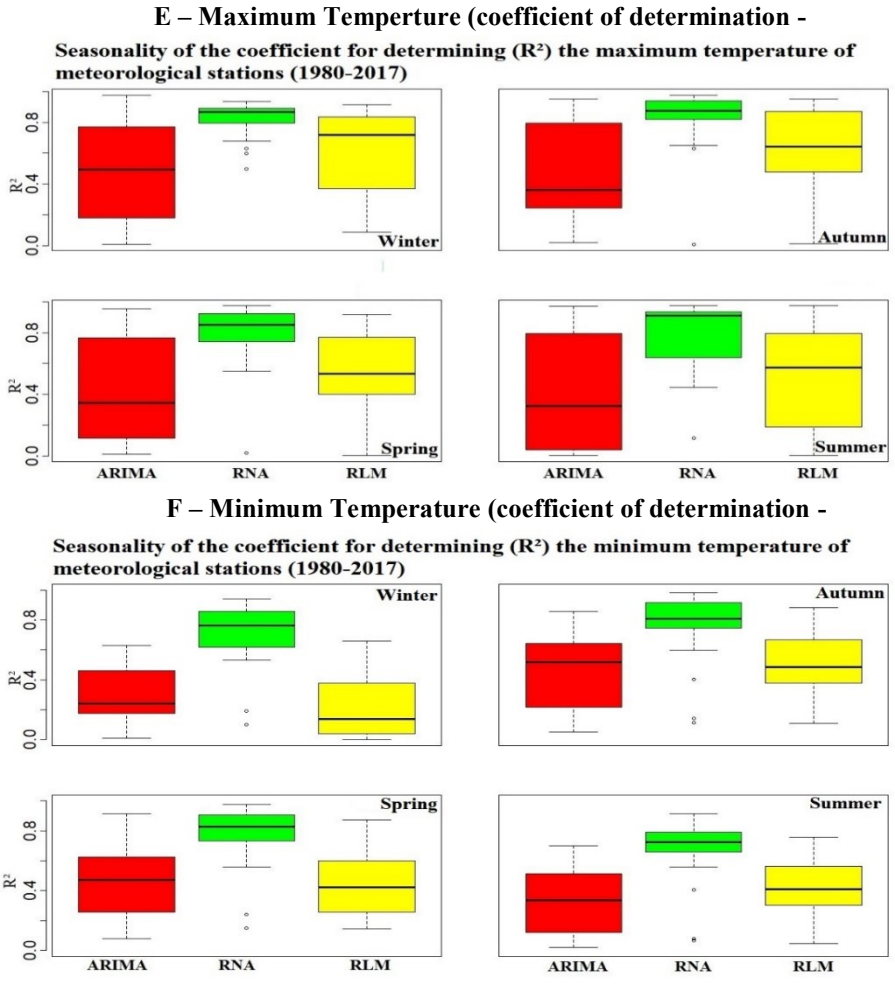


Source: Elaborated by the authors (2020)

The annual time series of maximum and minimum temperatures showed spatialization of similar metrics for the ANN model, in which, in general, the model presented minor errors between the simulated and observed data, however, tended to overestimate the results when the mean error was verified, a different situation when the MLR and ARIMA models were taken. The result shows that this metric may have been affected by some outlier present in the series, since it is a long period. As highlighted by Hallak and Pereira-Filho (2011), the average error cannot be considered as a measure of accuracy because it is sensitive to this type of event Another factor to be considered is the adjustment of weights of the RNA model, with the objective of stimulating the network to identify patterns among the explanatory variables used in the training of the network, with the objective of estimating the missing data (RUSSELL and NORVIG, 2010). Figure 6 shows the boxplots of the coefficient of determination

(R^2) used to evaluate the distribution of the model data used in the completion of the 21 meteorological stations throughout the Brazilian territory. Once again, we can observe that the ANN model with greatest explanatory power between the predicted and observed data, this is due to the greater number of training of the model so that each series obtained the best result, for example (GÉRON, 2017). However, the greatest variability of this coefficient was found in the ARIMA model in the four seasons of the maximum temperature and in autumn and summer for the minimum temperature. The MLR model was the second to present lower variability of the determination coefficient, which may be related to the low collinearity between the predictor variables used to estimate the data Unlike the results found by Oliveira et al., 2010, who used the MLR model and found satisfactory results when associating this technique with other models.

Figura 6 - Box-plot of adjustment measures (coefficient of determination - R^2) used in the evaluation of the performance of the filling models of the climatological series of maximum and minimum air temperature.



Source: Elaborated by the authors (2020)

CONCLUSION

The main objective of this work was to compare and evaluate the performance of three statistical models for filling failures in time series: artificial neural networks (ANN), multiple linear regression (MLR) and Autoregressive Integrated Moving Averages (ARIMA) or Box-Jenkins method. The evaluation was based on statistical metrics, mean error or bias, absolute mean error and coefficient of determination applied in time series of maximum and minimum mean air temperature to fill faults of 21 stations located in different geographic regions of Brazil.

The results showed that the application of ANN and MLR models tended to perform better than ARIMA, especially for maximum temperature. However, the mean error overestimated the ANN model while the MLR and Arima models overestimated the Belo Horizonte station in the Southeast of the country for the minimum temperature. A similar pattern was found in the seasonal period of overestimation in the prediction of data for the ANN model. However, it can be considered that there was good performance of ANN and MLR models in filling faults for the maximum and minimum temperature variable, since the best results were above 0.7 (70%) for the coefficient of determination, except for the minimum temperature in winter, whose value was up to 0.66 in almost all of Brazil using the model.

Although the results obtained with the application of the ANN model presented best performance in relation to the other two models for the prediction of data with failures for maximum and minimum temperature, it is necessary a greater number of training of the networks in order to simulate data with values closer to the observed data.

ACKNOWLEDGMENTS

The authors acknowledge the Coordination for the Improvement of Higher Education Personnel (CAPES) for granting a scholarship to master's research with the Graduate Program in Climate Sciences (PPGCC) of the Federal University of Rio Grande do Norte. Professor Alexandre C. Xavier, from the Department of Rural Engineering of the University of Espírito Santo, for the availability of meteorological variables at the grid point for the period from 1980 to 2017.

REFERENCE

AFRIFA-YAMOAH, E. et al. Missing data imputation of high-resolution temporal climate time series data. **Meteorological Applications**, v. 27, n. 1, 2020. doi:10.1002/met.1873

ALVARES, C. et al. Köppen's climate classification map for Brazil. **Meteorologische Zeitschrift**, v. 22, n. 6, p. 711–728, 2013.

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, [S.l.],** v. 32, n. 2, 215-226, 2017.

BOX, G.E.P.; JENKINS, G.M. **Time series analysis forecasting and control,** 2nd ed., San Francisco, Holden-Day, 1976.

BUUREN, S. V. **Flexible imputation of missing data**. [S.l.]: CRC press, 2012.

BUUREN, S. V.; OUDSHOORN, C. Multivariate imputation by chained equations. **MICE V1.0 user's manual**. Leiden: TNO Preventie en Gezondheid, 2000. 18, 31.

CAMELO, H. N. et al. Métodos de Previsão de Séries Temporais e Modelagem Híbrida ambos Aplicados em Médias Mensais de Velocidade do Vento para Regiões do Nordeste do Brasil. **Revista Brasileira de Meteorologia**, v. 32, n. 4, 565-574, 2017.

CHHABRA, C.; VASHISHT, V.; RANJAN, J. A. Comparison of Multiple Imputation Methods for Data with Missing Values. **Indian Journal of Science and Technology**, v. 10, n. 19, 2017.

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.

DALGAARD, P. **Introductory Statistics with R.** 2nd ed., Springer Verlag, 2008.

DANTAS, L. G.; SANTOS, C. A. C. dos.; OLINDA, R. A. de. Reamostragem de séries pluviométricas no estado da Paraíba. **Revista Brasileira de Geografia Física**, v. 09, n. 04, 997-1006, 2016.

DAS, K. R.; IMON, A. H. M. R. A Brief Review of Tests for Normality. **American Journal of Theoretical and Applied Statistics**, v. 5, n. 1, 5-12, 2016.

DEPINÉ, H.; CASTRO, N. M. R.; PEDROLLO, O. C. Incertezas no Preenchimento de Falhas de Chuvas Horárias com Redes Neurais Artificiais. **Estudos Ambientais**, v. 15, n. 2, p. 48-57, 2013.

DURBIN, J.; WATSON, G. S. Testing for serial correlation in Least squares Regression. III. **Biometrika**, v. 58, n. 11-19, 1971.

EL-MALLAH E.S.; ELSHARKAWY S.G. Time-series modeling and short term prediction of annual temperature trend on Coast Libya using the box-Jenkins ARIMA Model. **Advances in Research**, v. 6, n. 5, p. 1-11, 2016.

FILHO, D. F. F. et al. Aplicação de técnicas de interpolação para espacialização de chuvas da rede hidrográfica: estudo de caso Calha Norte – PA. **Revista Brasileira de Climatologia**, v. 24, 2019.

GÉRON, A. **Hands-On Machine Learning with Scikit-Learn and TensorFlow**. 1. ed. USA: O'Reilly Media, 2017.

GUJARATI, D.N.; PORTER, D.C. **Basic Econometrics**. Fourth Edition. McGraw-Hill, 922 p, 2009.

HALLAK, R.; PEREIRA FILHO, A. J. Metodologia para análise de desempenho de simulações de sistemas convectivos na região metropolitana de São Paulo com o modelo ARPS: sensibilidade a variações com os esquemas de advecção e assimilação de dados. **Revista Brasileira de Meteorologia**, v. 26, n. 4, p. 591–608, 2011.

HAYKIN, S.O. **Neural Networks and Learning Machines**. 3. ed. Upper Saddle River: Prentice Hall, 889 p. 2008.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Áreas urbanizadas do Brasil**. Série Relatórios Metodológicos. v. 44. Rio de Janeiro: IBGE, 2015.

INSTITUTO BRASILEIRO DE GEOGRAFIA E ESTATÍSTICA (IBGE). **Classificação e caracterização dos espaços rurais e urbanos do Brasil: uma primeira aproximação.** n. 11. Rio de Janeiro: IBGE, 2015.

INSTITUTO DE PESQUISA E ESTRATÉGIA ECONÔMICA DO CEARÁ (IPECE). **Perfil municipal: Jaguaruana***, 2017*. Fortaleza: IPECE, 2018b.

INSTITUTO DE PESQUISA E ESTRATÉGIA ECONÔMICA DO CEARÁ (IPECE). **Perfil geossocioeconômico: um olhar para as macrorregiões de planejamento do Estado do Ceará**. Fortaleza: IPECE, 2014.

INSTITUTO DE PESQUISA E ESTRATÉGIA ECONÔMICA DO CEARÁ (IPECE). **Perfil municipal: Fortaleza***,* 2017. Fortaleza: IPECE, 2018a.

KASHANI, M.H.; DINPASHOH, Y. Evaluation of efficiency of different estimation methods for missing climatological data. **Stochastic Environmental Research and Risk Assessment**, v. 26, n. 1, p. 59-71, 2012.

KEMP, W.P et al. Estimating missing daily maximum and minimum temperatures. **Journal of climate and applied meteorology**, v. 22, n. 9, 1587-1593, 1983.

KOUSKY, V. E. Frontal Influences on Northeast Brazil. **Monthly Weather Review**, v.107, p. 1140-1153, 1979.

LEAHY, P.; KIELY, G.; CORCORAN, G. Structural optimization and input selection of an artificial neural network for river level prediction. **Journal Hydrology**, v. 355, p. 192-201, 2008.

LEE, H.; KANG, K. Interpolation of Missing Precipitation Data Using Kernel Estimations for Hydrologic Modeling. **Advances in Meteorology**, v. 2015, 2015. doi.org/10.1155/2015/935868

MONTOGMERY, D. C.; JENNINGS, C. L.; KULAHCI, M. Introduction to Time Series Analysis and Forecasting. **Wiley-intersciense**, 2008.

MORETTIN, P. A., TOLOI, C. M. C. Análise de Séries Temporais. 2. ed. São Paulo: **Edgard Blucher**, 2006.

MURARA, P. G. Técnicas de interpolação para a pesquisa em climatologia regional e agroclimatologia. Ano 15. Edição Especial. XIII Simpósio Brasileiro de Climatologia Geográfica. **Revista Brasileira de Climatologia**, 2019.

MURAT, M. et al. Forecasting daily meteorological time series using ARIMA and regression models. **International Agrophysics**, v. 32, p. 253-264, 2018.

NEGNEVITSKY, M. **Artificial Intelligence*:*** A Guide to Intelligent Systems. 3. ed. Canada: Pearson Education, 2011.

NOGUEIRA, D. B.; DA SILVA, A. O.; DA SILVA, A. P. N. Comparação entre métodos de interpolação espacial para a estimativa da distribuição de precipitação no Ceará-Brasil. **IRRIGA**, v. 25, n. 1, p.131–142, 2020.

OLIVEIRA, L. F. C. de. 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.

PAPASTATHOUPOULOS, I.; TAWN, J A. A generalised Student's t-distribution. **Statistics and Probability Letters**, v. 83, p. 70-77, 2013.

R CORE TEAM (2017). **R: A language and environment for statistical computing.** R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

R CORE TEAM (2018). **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

REBOITA, M. S. et al. Regimes de precipitação na América do Sul: uma revisão bibliográfica. **Revista Brasileira de Meteorologia***,* v. 25, n. 2, p. 185–204, 2010.

RUSSELL, S.; NORVIG, P. **Artificial Intelligence:** A modern Aproach. 3. ed. USA: Pearson Education, 2010.

SALVIANO, M. F.; GROPPO, J. D.; PELLEGRINO, G. Q. Análise de Tendências em Dados de Precipitação e Temperatura no Brasil. **Revista Brasileira de Meteorologia**, v. 31, n. 1, p. 64-73, 2016.

SANCHES, F.; VERDUM, R.; FISCH, G. **Preenchimento de falhas em séries de dados pluviométricos de Uruguaiana (RS) e análise de tendência.** Disponível em: https://www.researchgate.net/publication/264082524>. Acesso em: 11 mai. 2018.

SAVIN, N. E. E WHITE, K. J. The Durbin-Watson test for serial correlation with extreme sample sizes or many regressors. **Econometrica**, v. 45, n. 8, p. 1989-1996, 1977.

SHAH, A. D. et al. Comparison of Random Forest and Parametric Imputation Models for Imputing Missing Data Using MICE: A CALIBER Study. **American Journal of Epidemiology**, v. 179, n. 6, 2014.

SHAPIRO, S. S.; WILK, M. B. An analysis of variance test for normality (complete samp1es). **Biometrika**, v. 52, n. 3 e 4, p. 591-611, 1965.

SHARMA, V.; RAI, S.; DEV, A. A comprehensive study of artificial neural networks. **International Journal of Advanced Research in Computer Science and Software Engineering**, vol. 2, n. 10, p. 278–284, 2012.

TABONY, R.C. The estimation of missing climatological data. **Journal of Climatology**, v. 3, n. 3, p. 297-314, 1983.

VENTURA, T. M. et al. Uma abordagem computacional para preenchimento de falhas em dados micrometeorológicos. **Revista Brasileira de Ciências Ambientais**, n. 27, 2013.

WADI, S. A.; ALMASARWEH, M.; ALSARAIREH, A. A. Predicting Closed Price Time Series Data Using ARIMA Model. **Modern Applied Science**, v. 12, n. 11; 2018.

WU, WEI; XU, AN-DING; LIU, HONG-BIN. High-resolution spatial databases of monthly climate variables (1961–2010) over a complex terrain region in southwestern China. **Theoretical and Applied Climatology**, v. 119, p. 353-362, 2014.

XAVIER, A. C.; KING, C. W.; SCANLON, B. R. Daily Gridded Meteorological Variables in Brazil (1980–2013). **International Journal of Climatology**, p. 2644–2659, 2016.

XIA, Y.; FABIAN, P.; STOHL, A.; WINTERHALTER, M. Forest climatology: estimation of missing values for Bavaria, Germany. **Agricultural and Forest Meteorology**, v. 96, p. 131-144, 1999.

YODAH WALTER. O.; KIHORO, J. M.; ATHIANY, K.H.O.; KIBUNJA H. W. Imputation of incomplete non- stationary seasonal time series data. **Mathematical Theory and Modeling**, v.3, n. 12, 2013.