Vol.63: e20180522, 2020
http://dx.doi.org/10.1590/1678-4324-2020180522
ISSN 1678-4324 Online Edition

Article - Engineering, Technology and Techniques

Artificial Neural Networks for Filling Missing Streamflow Data in Rio do Carmo Basin, Minas Gerais, Brazil

Gabriela Rezende de Souza[1]
https://orcid.org/0000-0001-5915-7529

Italoema Pinheiro Bello[1*]
https://orcid.org/0000-0001-6891-240X

Flávia Vilela Corrêa[1]
https://orcid.org/0000-0001-5179-5577

Luiz Fernando Coutinho de Oliveira[1]
https://orcid.org/0000-0001-5260-3258

[1]University of Lavras, Department of Water Resources and Sanitation, Lavras, Minas Gerais, Brazil.

Received: 2018.09.25; Accepted: 2020.03.17.

*Correspondence: italoemapb@hotmail.com; Tel.: +55-37-991180636 (I.P.B.)

HIGHLIGHTS

- Artificial Neural Networks applied for filling missing flow data.
- A simple approach of ANN using an open source software.

Abstract: Adequate availability of data directly influences the quality of hydrological studies. In this sense, procedures for filling gaps of observations are often applied in order to improve the length of hydrological series. One technique that can be used is the Artificial Neural Network (ANN), which process information from input data creating an output. This study aims to evaluate the application of ANN to fill missing data from monthly average streamflow series at Rio do Carmo Basin in the state of Minas Gerais, Brazil. A 26-years series (from 1989 to 2012) was used for ANN modelling while the two proceeding years, 2013 and 2014, were used to simulate failures pursuant to evaluating the performance of the ANN. The ANN construction was performed by the software WEKA that uses the multilayer perceptron model with sigmoidal activation functions. Four types of ANN were generated: five attributes and two (MLP1) or five (MLP2) neurons; and with three attributes and one (MLP3) or three (MLP4) neurons. The best-fit model to ANN was the MLP1, verified by Pearson correlation coefficients (0.9824), and coefficient of determination r^2 (0.9646). The model used five attributes, four input data (year, month, streamflow data from Acaiaca and Fazenda Paraíso stations) and one output data (streamflow from Fazenda Oriente station), that considered the temporal variation of streamflow. Hence, the utilization of the ANN generated by the WEKA was adequate and can be considered a simple approach, not requiring great computational programming knowledge.

Keywords: data consistency; data mining; hydrologic estimation.

INTRODUCTION

Hydrological studies quality is directly related to temporal series availability, its quality, spatial distribution, climatological phenomena dynamics and the model quality used for filling missing data. According to Oliveira and coauthors [1], missing values in hydrological data may be due to carelessness of the observers, problems in registration mechanisms, mislay of transcripts of notes or records by operators and shut down of the gauge stations. There are different techniques for filling missing data in temporal series. However, choosing one in expense of another must be verified through statistical evaluation [2]. Various studies have been applying Artificial Neural Networks (ANN) as an alternative technique to fill missing hydrological data, mainly because of its capacity to simulate non-linear processes [3,4].

ANN is a tool that process information starting from one or more inputs and creates an output [5]. The ANN model's development involves the ANN configuration, data collecting, training and generalization check. The intermediate layers number and the neurons number in each ANN are chosen empirically according to the data complexity and the predominance of non-linear characteristics [6]. In the training or learning stage, the ANN uses a data set (input layer), which describes a situation, to extract the provided data characteristics (intermediate layer), being able to reproduce appropriate answers (output layers) to any other situation [6,7]. The amplitude of output values, i.e. the given answer to input signal, is defined by the activation function, responsible for introducing the non-linearity characteristic. In water resources, the sigmoid function is the most used non-linear transfer function [8].

In the verification stage, the ANN efficiency is assessed concerning the generalization capacity in the training phase. Once the obtained output data are statistically equivalent to their references, the ANN is considered trained and able to estimate the established model to any data set [8]. According to Correia and coauthors [2], generally 70% of the data is applied in the learning stage, which must be meaningful and cover largely the problem domain to train the ANN, and 30% to validate the neural network.

In hydrology, the ANN has been applied to a diversity of complex problems, demonstrating good results and viability. Among the ANN utilization in hydrological studies are: rainfall-runoff modelling [9]; filling of missing data in temporal series [2]; filling of rainfall missing data [5,8,10]; outliers detection in micrometeorological data [11]; meteorological data spatial interpolation [12,13] and flow forecast [14].

In this sense, the present work aims to evaluate the ANN application for filling missing data in series of monthly average streamflow in Rio do Carmo Basin, in Minas Gerais state, Brazil using an open source software.

MATERIAL AND METHODS

Data of three streamflow gauge stations located at the Rio do Carmo Basin were obtained from the Hidroweb database [15]. Table 1 shows the station name, identification code, geographical coordinates and the registered years of observations.

Table 1. Streamflow gauge stations used in the study.

Station	Code	Latitude	Longitude	Registered period
Acaiaca Jusante	56335001	20° 21' 41.04"	43° 8' 21.84"	1975 - 2014
Fazenda Ocidente	56337000	20° 16' 1.92"	43° 6' 2.88"	1938 - 2014
Fazenda Paraíso	56240000	20° 23' 25.08"	43° 10' 54.84"	1930 - 2014

For this study, monthly average streamflow series were constructed for the period of registration that is common for the three stations. In this way, 26 years of monthly average streamflow, with no missing data, were utilized, from 1989 to 2012, in order to model the ANN. The years 2013 and 2014 were used to simulate missing data, enabling to evaluate the performance of the ANN in estimating data.

Initially, in order to verify the homogeneity of the stations, a consistency analysis was performed for each station, using the double mass curve methodology, as described by Bertoni and Tucci [16]. For a selected station, the cumulative rainfall annual totals were plotted on the ordinate axis and the mean of rainfall annual totals of the other stations on the abscissa axis. The data can be considered consistent if the analyzed station annual totals produces a linear trend in relation to the other stations, by evaluating the 1:1 regression line adequacy [2].

The neural networks construction and solution were performed on WEKA (Waikato Environment for Knowledge Analysis), which is a free software and available on the website of the University of Waikato [17]. In this software, the generated ANN are the multilayer perceptron (MLP) type and it uses sigmoidal activation functions for the interactions. WEKA is constituted by a set of algorithms previously implemented with several

techniques for data mining, which was developed using the Java language. According to Coulibaly and Évora [18], the MLP has being frequently applied to solve a variety of classification and pattern recognition problems, standing out as one of the most effective models in filling gaps in data. The MPL fundamental learning algorithm is the backpropagation, which is based on the descending gradient method.

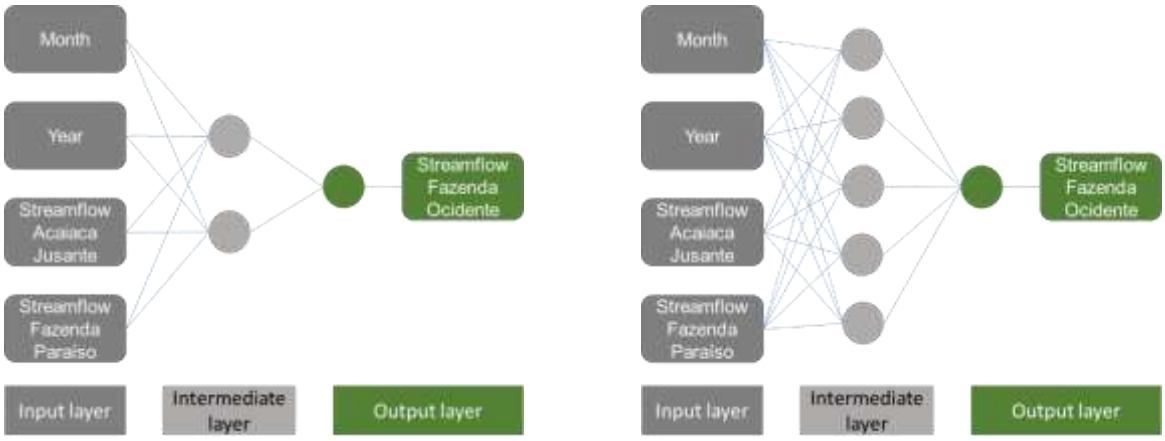
In this study, we simulated gaps in the data of Fazenda Ocidente station, therefore, the data from Acaiaca – Jusante and Fazenda Paraíso stations were inserted as input layers and the output layer was the data from Fazenda Ocidente. To ANN modelling, the software randomly samples 80% of the data to train the model and 20% of data to validate the model. In addition, WEKA is programmed to self-adjust with a default in order to get better results and minimize errors. The supervised training algorithm Multilayer Perceptron was used, in which the maximum period of 500 iterations was fixed, the learning rate was 0.3 and the moment rate was 0.2. The performed steps were as follow:

a) Pre-processing: The streamflow data were imported into the software in “ARFF" file format. Four input attributes were provided (year, month, streamflow values from both Acaica-Jusante and Fazenda Paraíso gauge stations) and one output (Fazenda Ocidente streamflow data);

b) Training, validation and testing: At first, the tests were performed considering all attributes, and then were considered only the streamflow data attributes. Both tests were performed with the default of the program and using modifications. The different modeled ANN were:

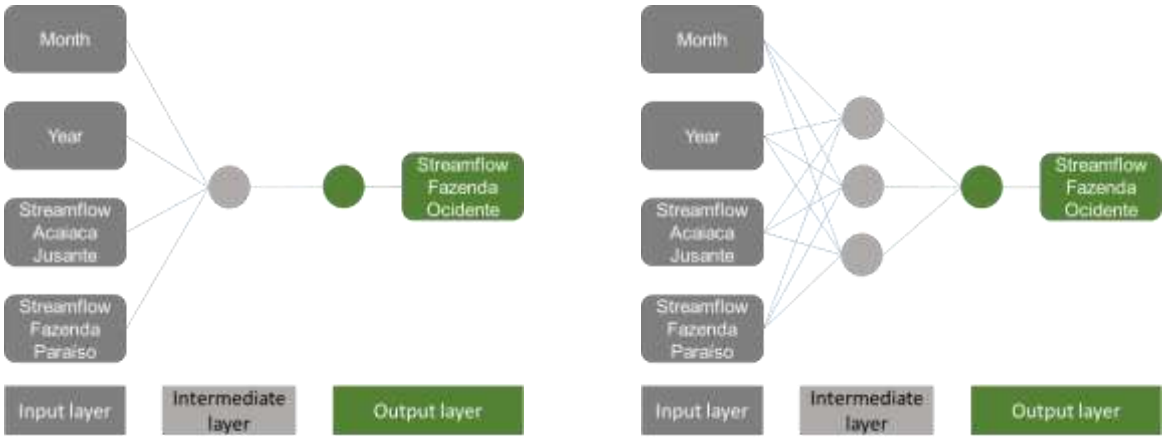
- MLP1: using the five attributes with two neurons (software default) in the activation layer (Figure 1a);
- MLP2: using the five attributes with five neurons (modified) in the activation layer (Figure 1b);
- MLP3: using only the streamflow data with one neuron (software default) in the activation layer (Figure 2a);
- MLP4: using only the streamflow data with three neurons (software default) in the activation layer (Figure 2b);

c) Results of filling missing streamflow: after the tests, the best-adjusted ANN model was chosen and the streamflow for the years of 2013 and 2014 were simulated.



(a) (b)

Figure 1. ANNs' structure for (a) MLP1 with five attributes and two neurons; (b) MLP2 with five attributes and five neurons.



(a) (b)

Figure 2. ANNs' structure for (a) MLP3 with three attributes and one neuron; (b) MLP4 with three attributes and three neurons.

The multiple linear regression method (MLR) was also used to fill the simulated missing data, in order to compare the results obtained by ANN. The method was solved by WEKA and was only adopted by the default program with three attributes (streamflow data from Acaica, Fazenda Paraíso and Fazenda Ocidente).

In order to evaluate the performance of the models, some statistical parameters were calculated. According to Macedo [9], all the equations are described in terms of observed flow (Qo), calculated flow (Qc) and number of observations (N).

- Mean Squared Error (MSE)

$$MSE = \frac{\sum_1^N (Q_0 - Q_c)^2}{N}, \quad (1)$$

- Average Absolute Error (AAE)

$$AAE = \frac{\sum_1^N (Q_0 - Q_c)}{N}, \quad (2)$$

- Efficiency Coefficient (EC)

$$EC = 1 - \frac{\sum_1^N (Q_0 - Q_c)^2}{\sum_1^N (Q_0 - Q_{0m})^2}, \quad (3)$$

- Pearson's Correlation Coefficient (r)

$$r = \frac{N \sum (Q_o Q_c) - \sum Q_0 \sum Q_c}{\sqrt{\left[N \sum Q_0^2 - \left(\sum Q_0^2\right)\right] \left[N \sum Q_c^2 - \left(\sum Q_c^2\right)\right]}}, \quad (4)$$

- Determination Coefficient (R)

$$R = r^2, \quad (5)$$

RESULTS

Figures 3, 4 and 5 show the double mass curves of the three stations utilized in this study, plotting one accumulated annual total rainfall in relation to the other two average.

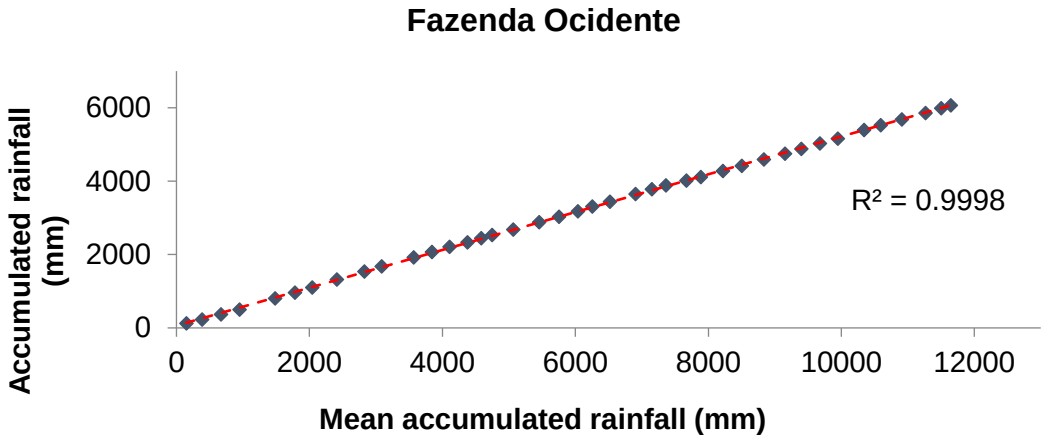


Figure 3. Double mass curve for Fazenda Ocidente station.

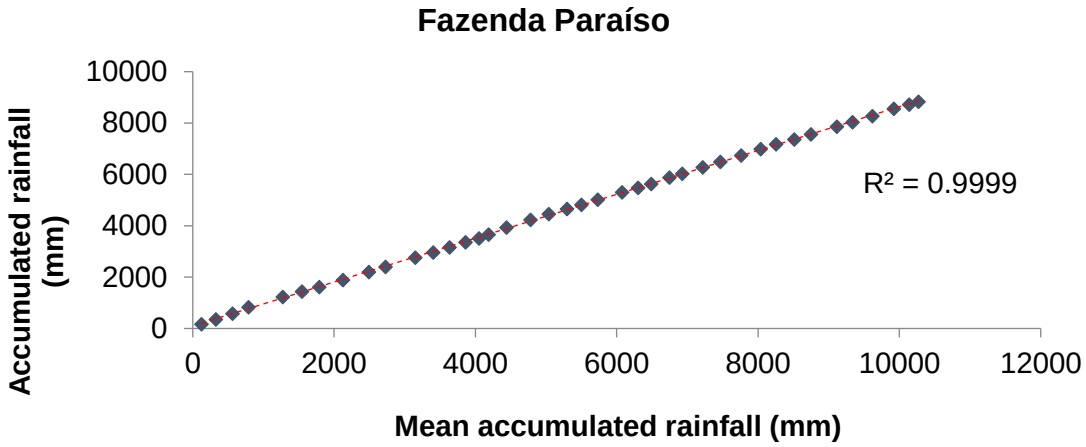


Figure 4. Double mass curve for Fazenda Paraíso station.

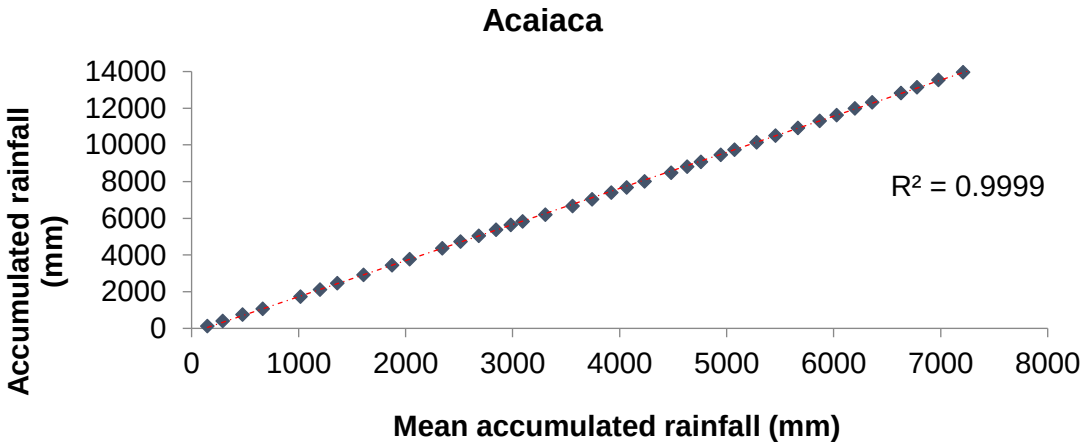


Figure 5. Double mass curve for Acaiaca station.

The data series used for ANN training and validation presented a length of 312 data values, where 58 data values were used in the validation phase. Table 2 shows the errors for each statistical model. It was observed that for the ANN using 5 attributes, the software default model with 2 neurons in the intermediate layer (MLP1) was the one that best suited the data series provided. For the ANN generated using 3 attributes, the one with 3 neurons (MLP4) was the most adequate. Comparing all the ANN, the MLP1 presented smaller errors and values very similar to the regression model adjusted (REGLIN), which the equation was as follows.

$$Q_{Fazenda\ Ocidente} = 0{,}5053*Q_{Acaiaca} - 0{,}0972*Q_{Fazenda\ Paraíso} - 1{,}2403 \quad (6)$$

Table 2. Validation statistics of the adjusted models.

Adjustment model	MLP1	MLP2	MLP3	MLP4	REGLIN
Attributes number	5	5	3	3	-
Neurons number	2	5	1	3	-
Data used in validation	58	58	58	58	58
r^2	0.9511	0.9402	0.9457	0.9502	0.9516
Correlation coefficient (Pearson)	0.9752	0.9696	0.9725	0.9748	0.9755
Average Absolute Error (AAE)	1.3135	1.9796	2.0638	1.8337	1.1505
Root Mean Square Error (RMSE)	1.697	2.3371	2.4057	2.1982	1.5612
Absolute Relative Error	23.70%	35.72%	37.24%	33.09%	20.76%
Root Relative Mean Square Error	23.21%	31.97%	32.91%	30.07%	21.36%

Thus, the MLP1, MLP4 and REGLIN models were applied to simulate the streamflow for the Fazenda Ocidente gauge station. The observed data in the years 2013 and 2014 and the calculated streamflow by the models were plotted on a scatterplot in order to observe the fitting of the models, by means of the regression coefficient r^2 (Figure 6). In addition, Table 3 shows the error statistics for the performance evaluation.

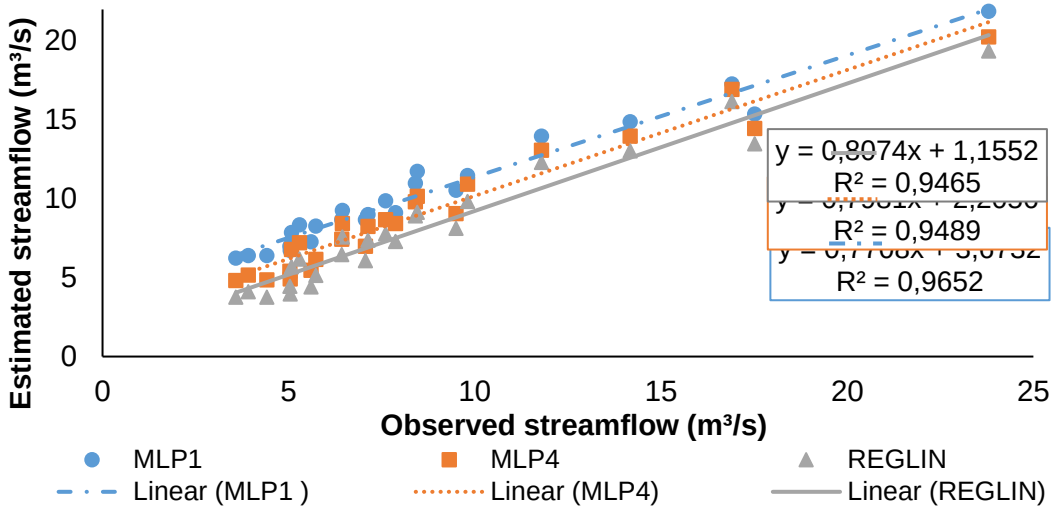


Figure 6. Observed and estimated flow for Fazenda Ocidente streamflow gauge station.

Table 3. Performance statistics of the simulated missing values in Fazenda Ocidente streamflow in the years of 2013 and 2014.

Model	AAE	MSE	RMSE	r (Pearson)	EC (Nash-Sutcliffe)	r^2
MLP1	2.0393	4.6489	2.1561	0.9824	0.8040	0.9652
MLP4	1.0960	1.9999	1.4142	0.9741	0.9157	0.9489
REGLIN	0.9419	2.0051	1.4160	0.9729	0.9154	0.9465

DISCUSSION

The double mass curve analysis shows that the three stations presented high r^2 values, 0.9998 for Fazenda Ocidente, 0.9999 for Fazenda Paraíso and 0.9999 for Acaiaca. This analysis demonstrates the data

consistency and homogeneity. Therefore, it is appropriate to use hydrologic data from Acaiaca and Fazenda Paraíso gauge stations, to fill the existing gaps in observations at Fazenda Ocidente gauge station.

The MLP1 model presented the highest mean error values and lower Nash-Sutcliffe efficiency coefficient (0.8040). However, this method was the one that best fitted for estimating the missing values by Pearson correlation coefficients (0.9824), and determination coefficient r^2 (0.9646).

Depiné and coauthors [5], who applied ANN to estimate hourly streamflow missing data, found Nash-Sutcliffe coefficient minimum values of 0.91 in the validation and testing phases and minimum of 0.81 in data verification. Wanderley and coauthors [13] obtained r^2 values ranging from 0.72 to 0.99, demonstrating an acceptable fit between the estimated ANN precipitation and the observed values for the study months. The results above 0.80 of Nash-Sutcliffe's efficiency coefficients (0.8040) and r^2 determination (0.9646) for the MLP1 characterize a very good estimate. These values reaffirm an adequate fit between the streamflow estimated by the ANN and the values observed, as presented by Correia and coauthors [2].

Another relevant fact is that the MLP1 considered the attributes "year" and "month" for its validation and tests, while the MLP4 disregarded them, proving that seasonality and temporal effects influence ANN. Depiné and coauthors [5] also demonstrated the temporal effects in their study when training ANN. These authors constructed ANN with and without the months and years attributes and observed an improvement in the statistical performance indexes for the second scheme. In addition, the MLP1 adequacy can be explained by the fact that ANN construction used the software default parameters that are previously implemented to make iterations and result in smallest errors.

The multiple linear regression is a technique widely used for filling missing data. The errors from MLP1 were very similar to the REGLIN and in some cases superior. That said, it can be inferred that the ANN technique is a quite efficient and suitable approach for estimating missing streamflow data.

CONCLUSION

The utilization of ANN to fill streamflow monthly average series proved to be effective in Rio do Carmo Basin. The model that best fit was the MLP1 (five attributes, four input and one output), composed by five neurons. In this study, simpler neural networks were built. Therefore, a better ANNs understanding and activation layers improvement, neurons number, training and activation functions, could produce better results. Also, the application of the WEKA software to this study showed an easy approach for ANN generation, not requiring great computational programming knowledge.

Funding: This research received no external funding.
Acknowledgments: The authors acknowledge CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) (Brazil) for the scholarship grants. As well, the Federal University of Lavras due to the work support.
Conflicts of Interest: The authors declare no conflict of interest.

REFERENCES

1. Oliveira LF de, Fioreze AP, Medeiros AM, Silva MAS. Comparação de metodologias de preenchimento de falhas de séries históricas de precipitação pluvial anual. Rev. Bras. Eng. Agrícola e Ambient. 2010,14(11)1186–92.
2. Correia TP, Dohler RE, Dambroz CS, Binoti DHB. Aplicação de Redes Neurais Artificiais no Preenchimento de Falhas de Precipitação Mensal na Região Serrana do Espírito Santo. Geosci. = Geociências. 2016, 35, 560–7.
3. Dias TL, Cataldi M, Ferreira VH. Aplicação de técnicas de redes neurais e modelagem atmosférica para elaboração de previsões de vazão na Bacia do Rio Grande (MG). Eng. Sanit. e Ambient. 2017,22(1),169–78.
4. Özçelik R, Diamantopoulou M, Brooks JR, Wiant Jr H V. Estimating tree bole volume using artificial neural network models for four species in Turkey. J. Environ. Manage. 2010, 91, 742–3.
5. Depine H, Maria N, Castro R, Pedrollo OC. Incertezas no preenchimento de falhas de chuvas horárias com redes neurais artificiais. REA – Rev. Estud. Ambient. 2014, 15, 48–57.
6. Silva FF, Silva FF, Matias I de O, Souza CLM. A Comparison of Artificial Neural Networks and Traditional Time Series Models : a Prediction of the Oil. Interdiscip. Sci. J. 2017, 4, 225–38.
7. Mulero Á, Pierantozzi M, Cachadiña I, Equilibria GDN. An Artificial Neural Network for the surface tension of alcohols. Fluid Phase Equilib. 2017, 449, 28–40.
8. Anesi HD. Influência do preenchimento de falhas de dados horários de precipitação por redes neurais artificiais (RNAs) na simulação hidrológica de base física em uma bacia rural. PhD, Universidade Federal do Rio Grande do Sul, Porto Alegre, October 2014.

9. Macedo MJH. Aplicações de Redes Neurais Artificiais e Satélite TRMM na Modelagem Chuva-Vazão da Bacia Hidrográfica do Rio Paraguaçu/BA. PhD in Meteorology, Universidade Federal de Campina Grande, Campina Grande, 2013.
10. Depiné H, Maria N, Castro R, Pinheiro A, Pedrollo O. Preenchimento de falhas de dados horários de precipitação utilizando redes neurais artificiais. Rev. Bras. Recur. Hídricos 2014, 19, 51–63.
11. Bonfante AG, Ventura TM, de Oliveira AG, Marques HO, Oliveira RS, Martins CA, et al. Uma abordagem computacional para preenchimento de falhas em dados micro meteorológicos. Rev. Bras. Ciências Ambient. 2013, 27, 61–70.
12. Wanderley HS, Amorim RFC de, Carvalho FO de. Variabilidade espacial e preenchimento de falhas de dados pluviométricos para o estado de Alagoas. Rev. Bras. Meteorol. 2012, 27, 347–54.
13. Wanderley HS, Amorim RFC de, Carvalho FO de. Interpolação espacial de dados médios mensais pluviométricos com redes neurais artificiais. Rev. Bras. Meteorol. 2014,, 29, 389–96.
14. Sousa W, Sousa F de. Rede neural artificial aplicada à previsão de vazão da Bacia Hidrográfica do Rio Piancó. Rev. Bras. Eng. Agrícola e Ambient. 2010, 14, 173–80.
15. Sistema Nacional de Informações sobre Recursos Hídiicos. HIDROWEB. Available Online: http://www.snirh.gov.br/hidroweb/serieshistoricas
16. Bertoni JC, Tucci CEM. Precipitação. In: Hidrologia: Ciência e Aplicação. 2007. page 177–241.
17. University of Waikato. Weka 3 - Data Mining with Open Source Machine Learning Software in Java. Available online: https://www.cs.waikato.ac.nz/ml/weka/
18. Coulibaly P, Evora ND. Comparison of neural network methods for infilling missing daily weather records. J. Hydrol. 2007, 341, 27–41.