RBRH

*Revista Brasileira de Recursos Hídricos*
*Brazilian Journal of Water Resources*
*Versão On-line ISSN 2318-0331*
*RBRH*, Porto Alegre, v. 29, e12, 2024
*Scientific/Technical Article*

*https://doi.org/10.1590/2318-0331.292420230057*

# Application of data prediction models in a real water supply network: comparison between arima and artificial neural networks

## *Aplicação de modelos de predição de dados em uma rede real de abastecimento de água: comparação entre arima e redes neurais artificiais*

**André Carlos da Silva[1], Fernando das Graças Braga da Silva[1], Victor Eduardo de Mello Valério[1], Alex Takeo Yasumura Lima Silva[1], Sara Maria Marques[1] & José Antonio Tosta dos Reis[2]**

[1]Universidade Federal de Itajubá, Itajubá, MG, Brasil
[2]Universidade Federal do Espirito Santo, Vitória, ES, Brasil

E-mails: andrecsilva58@gmail.com (ACS), fernandobraga@unifei.edu.br (FGBS), victor.dmv@gmail.com (VEMV), alex.takeo@uol.com.br (ATYLS), d2021102060@unifei.edu.br (SMM), jatreis@gmail.com (JATR)

Received: May 19, 2023 - Revised: March 06, 2024 - Accepted: March 12, 2024

## Abstract

Research around the world has focused on developing ways to predict hydraulic parameters in water distribution systems. The application of these forecasts can contribute to the decision-making of water distribution systems managers, aiming to ensure that the demand is met, and even to reduce water losses. The present work sought, among two data prediction models (ARIMA and Multi-Layer Perceptron Artificial Neural Networks), to assess which one can perform best predictions of pressure and discharge rate data. To reach the stipulated goal, real data were obtained from a water supply network provided by NUMMARH - Nucleus of Modeling and Simulation in Environment and Water Resources and Systems of the Federal University of Itajubá, Brazil. These data initially underwent an adjustment so that it was possible to develop a computer program. The results showed that the best prediction model for the data in question was ARIMA, presenting a mean absolute percentage error (MAPE) of 8.54%. Thus, it is concluded that ARIMA models are easy to build and apply, being an advantageous tool to predict such hydraulic parameters.

**Keywords:** Artificial neural networks; ARIMA; Hydraulic parameters; Prediction.

## Resumo

Pesquisas ao redor do mundo vem se concentrando em desenvolver maneiras de prever parâmetros hidráulicos em sistemas de distribuição de água. A aplicação destas previsões pode contribuir para a tomada de decisão dos gestores dos sistemas de distribuição de água, visando a garantia do atendimento da demanda, e até mesmo a redução de perdas de água. O presente trabalho buscou, dentre os modelos ARIMA e Redes Neurais Artificiais do tipo "Perceptron" de múltiplas camadas, identificar qual é capaz de realizar a melhor previsão de dados de pressão e vazão. Primeiro foram obtidos dados reais de uma rede de abastecimento de água em região de topografia irregular, que foram coletados pelo grupo de pesquisa NUMMARH - Núcleo de Modelagem e Simulação em Meio Ambiente e Recursos e Sistemas Hídricos, e fornecidos para execução deste trabalho. Os resultados mostraram que o melhor modelo de previsão para os dados em questão foi o ARIMA, apresentando erro médio percentual absoluto (MAPE) de 8,54%. Como conclusão identificou-se que os modelos ARIMA são de fácil construção e aplicação, sendo assim uma vantagem para se prever tais parâmetros hidráulicos.

**Palavras-chave:** Redes neurais artificiais; ARIMA; Parâmetros hidráulicos; Previsão.

## INTRODUCTION

In water resources, data prediction models are applied in researchs to make different predictions in supply systems in numerous cities around the world, being the basis for optimizing the operation, planning and management of water distribution networks (Zubaidi et al., 2019). There are a vast number of stochastic (linear) prediction models as well as models based on artificial intelligence (non-linear) that can have diverse applications in water distribution networks according to Jetmarova et al. (2017), and neural network models of the Multi-layer Perceptron (MLP) type and integrated regressive models such as the autoregressive integrated moving average (ARIMA), objects of study in this paper.

Perceptron Multilayer Neural Networks (MLP) are tools with great potential for use in water distribution networks. Jang & Choi (2017) compares the use of MLP with multiple regression analysis to estimate non-revenue water in Incheon, selecting 173 measurement and control districts (DMA) to apply the techniques, in addition to six parameters, two of which operational (energy demand rate and number of leaks) and four physical (average diameter of pipes, length of pipes, water sent to the network and index of deteriorated pipes), obtaining with the MLPs an average absolute error of 6.2, being better than multiple regression model with mean absolute error of 10.0, recommending its use instead of multiple regression.

Ghosal et al. (2019) adopt MLPs as the basis of a new multivariable prediction platform for modeling water distribution networks, while Kamiński et al. (2017) carried out an evaluation of a water distribution network through neural networks in their work, while Awad & Zaid-Alkelani (2019) used MLP and statistical models to predict urban water demand, since forecasts allow better planning of network operation and maintenance, reducing operational costs and shortages due to failures.

Lorente-Leyva et al. (2019), also presented the application of an artificial neural network model to forecast water demand with time series. This method provided a forecast without the need to include factors such as number of consumers and meteorological indices. This model was compared with traditional forecasting methods such as ARIMA and proved to be better for this type of application as it approaches real consumption behavior, since ARIMA tends to adjust and attenuate forecasts, making them more constant.

Lopez Farias et al. (2018) list a set of models, such as the autoregressive integrated moving average model (ARIMA), MLP, among others, used to predict demand in water distribution networks, while Gharabaghi et al. (2019) used the ARIMA model in the city of El Paso, Texas, United States, to predict the city's monthly water consumption. Using climatic and economic variables and water disposal rate as input data into the model, they were able to identify that climatic and economic issues have more influence on water consumption than the hydraulic variables themselves. However, for the MLP approach, the need to improve the quality and accuracy of the models was identified.

Another application of the ARIMA model is presented at Guarnaccia et al. (2020), which they used in the case study of a distribution network reservoir in the Benevento area, in the Campania region, in Italy. The study consisted of applying the ARIMA model to data collected from May 2018 to January 2019, finding the best adequacy of the data to meet the demand and seasonality of the system. For the authors, the model applied in this case study proved to be efficient, as the residual data analyzed by its simulation are in accordance with those observed in the system.

The use of ARIMA in conjunction with other models has already been developed, as seen in Xu et al. (2019), which brought a new water level prediction model, based on the interaction between the ARIMA model and the RNN (Recurrent Neural Network) model. The experiment was applied to Lake Taihu, in China, where water level data from a period of 30 days and environmental vectors were used to develop this method. The application of this interaction occurs with ARIMA obtaining the predicted residual value of the water level in the study area, and the RNN adjusts these residual sequences. According to the authors, this model is better used than traditional models, as it allows working with linear and non-linear data.

Still in the study brought by Du et al. (2020), they presented the use of the autoregressive moving average model (ARIMA) to estimate daily consumption data, modified by the Markov chain, called ARIMA-M. This joint use makes it possible to correct the forecast error, reducing the overlap of continuous errors, in addition to improving the estimation of future daily consumption data. This model works by connecting historical data based on the ARIMA model to the Markov model to predict the trend of future data. Thus, ARIMA is modified, improving prediction error and predictability. To confirm the developed method, data from 2016 to 2017 of daily water demand in Guangdong province, China, were used.

Zubaidi et al. (2018) used a combination of simple spectrum analysis and MLP as a new approach to predict monthly water demand using climate factors, proving to be a reliable and efficient model with $R = 0.972$ for seasonal data. Adamowski et al. (2012) compared several data forecasting models to predict daily demand in the city of Montreal, Canada. Thus, they used the following prediction models: Multiple Linear Regression Model (MRLM), Multiple Nonlinear Regression Model (MRNLM), Autoregressive and Integrated Moving Average Model (ARIMA), Artificial Neural Network Model (MRNA) and finally the Coupled Wavelet Model. with Artificial Neural Networks (WA-MLP), the latter being the most promising.

It is observed that many papers use techniques involving different methods, including MLP (and its variants) or ARIMA, individually or together, to predict demand in networks (Bo et al., 2021; Porto et al., 2021; Lopez Farias et al., 2018; Pandey et al., 2021; Shirkoohi et al., 2021). However, for pressure prediction, the main concern of this work is with the occurrence of pressure peaks and the detection of leaks, as this is an operational emergency, as can be seen in Alizadeh et al. (2019), Wu & Liu (2017) and Zhou et al. (2019).

Thus, Lima et al. (2018) show a model for near real-time estimation of node pressures for water distribution networks using MLP fed by real-time pressure monitoring data. MLPs have shown the advantage of, in case of possible data failures, reducing the chance of estimation errors, as they have a large set of input data for training that allows the prediction to be carried out. Xu et al. (2020) address pressure prediction based on deep learning through LSTM (Long-Short Term Memory) Neural Networks, obtaining

more accurate predictions and capable of detecting abnormal events in the networks, but at the cost of greater model complexity, with higher computational cost to obtain the prediction.

This paper, therefore, applies the comparison between two of the main methods used for historical series forecasts (ARIMA and MLP) to carry out pressure and discharge forecasts in the network under study, whose data were obtained from the NUMMARH-UNIFEI research group in campaigns previous field trips. The novelty of this article is to use only the ARIMA model, without combining it with other models, to carry out pressure forecasts in a real network sector in a mountainous terrain region, as it is usually used for forecasts of historical series, and to carry out a comparison with the results obtained by the MLP. The MLP is the most common method of pressure predictions, considering that in regions of irregular topography, large pressure oscillations are expected to occur, with consequent difficulty in predicting such parameters. Thus, the objective of this work is to identify which of the two models obtained the best performance in a real and unfavorable situation, with a view to future development of software for use in water distribution networks.

## METHODOLOGY

This work was structured based on the steps represented in the discharge rate graph in Figure 1, which will be detailed in the following topics.

### Obtaining raw data from the study site

Pressure head (m) and discharge (l/s) data obtained by the NUMMARH research group at the Federal University of Itajubá were used, in a measurement campaign that took place in 2014, from September 20 to 28, in a city in the south of Minas Gerais, a region whose irregular topography proves challenging for all aspects of the operation of a water distribution network, especially pressure variations in the network due to topographic heights.

According to Vieira (2019), it appears that this city is touristic and had a population of 41,657 inhabitants in 2010, with the most recent projection being 45,448 inhabitants for the year 2018. Furthermore, the population density of the municipality is approximately 718 inhabitant/ $m^2$ and its area is 58,019 $km^2$.

Pressure meters were then installed in the homes of network consumers and discharge meters were installed at the inlet and outlet of the water supply network. Thus, Figure 2 presents a schematic of the network used in this study, as well as the data collection points (defined as network nodes and identified by numbers).

Table 1 shows how the network nodes are distributed across neighborhoods and the number of residences per neighborhood.

The information obtained about the distribution of network nodes by neighborhood, as well as the number of residences, is presented in Table 1. Pressure head data was gathered at 30 min, obtaining 350 readings per node, and flow was at 10 min, obtaining 1152 readings per node.

### Preliminary treatment of data

After carrying out the measurement campaign, the pressure and discharge rate values of the measurement sensors, which collected data every 10 or 15 min, were stored in an electronic spreadsheet format, made available by the NUMMARH research group. Thus, this stage was characterized by obtaining this raw data and processing it so that the following steps could be carried out. The raw data can be found in the supplementary file of this paper for consultation.

In the subsequent steps, all were carried out with the aid of an algorithm developed in the R programming language, from

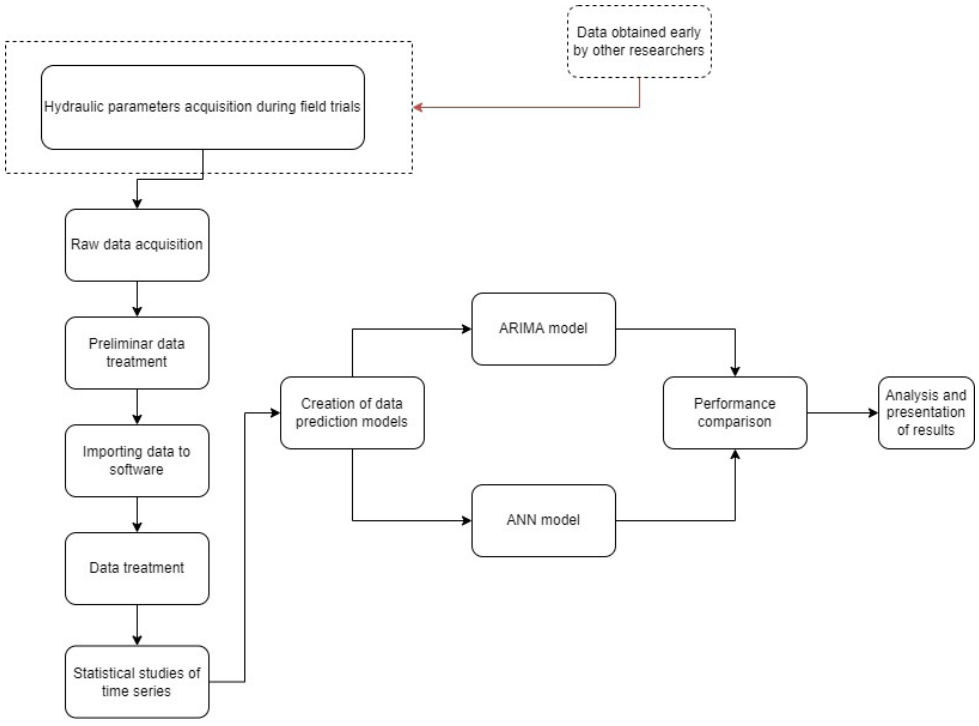


**Figure 1.** Methodology followed in this research.

**Table 1.** Network nodes and the neighborhoods to which they belong.

| Neighborhood | Moradas da Serra | | | | | Vale dos Pinheiros | | | | | | | | Palmela | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| No. of residences | 73 | | | | | 218 | | | | | | | | 201 | |
| Nodes | 62 | 85 | 108 | 114 | 115 | 51 | 103 | 11 | 112 | 113 | 109 | 116 | 117 | 14 | 105 |

**Source:** Vieira (2019).

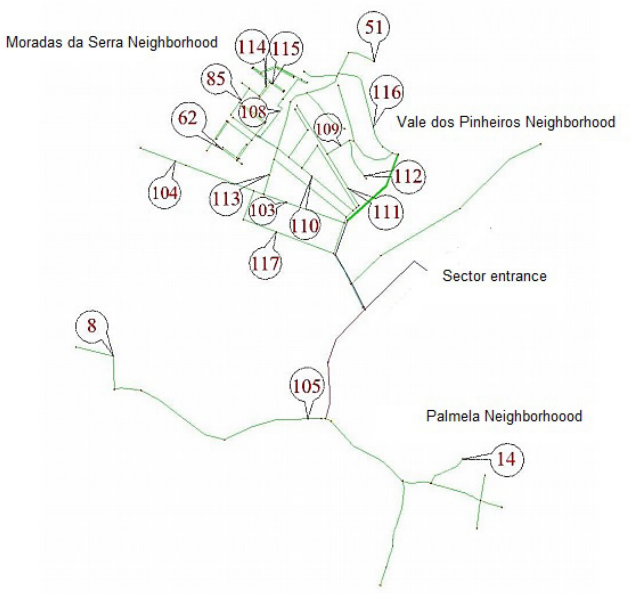


**Figure 2.** Location of measurement points.
**Source:** Vieira (2019).

organizing the dataset into dataframes (two dataframes, one for data measured every 10 min and the other for data measured every 15 min), which is a large data matrix, where the measurement times were organized in rows, and the network nodes in columns, and then the pressure or discharge rate values were filled into each cell of the dataframes.

## Multi-Layer Perceptron Artificial Neural Networks (MLP)

The Artificial Neural Network used in this article is of the feed-forward Multi-Layer Perceptron (MLP) type, with input layers, a simple hidden layer and an output layer, as shown in Figure 3. The neurons in the input layer (17) distribute the signals for the neurons of the hidden layer (9), being connected through weights, and the hidden and output layers have the activation function. 1000 epochs were used for the MLP in question.

The importance of the activation function comes from its ability to control the level of activation and the strength of the output signal of an artificial neuron. In general, we seek to use non-linear activation functions, as this characteristic allows greater competence in approximating functions (Amaral, 2020). The activation function used is sigmoidal, σ (z), which can be seen in Equation 1, where z is the input signal.

$$\sigma(z)=\frac{1}{\left(1+e^{-z}\right)} \quad (1)$$

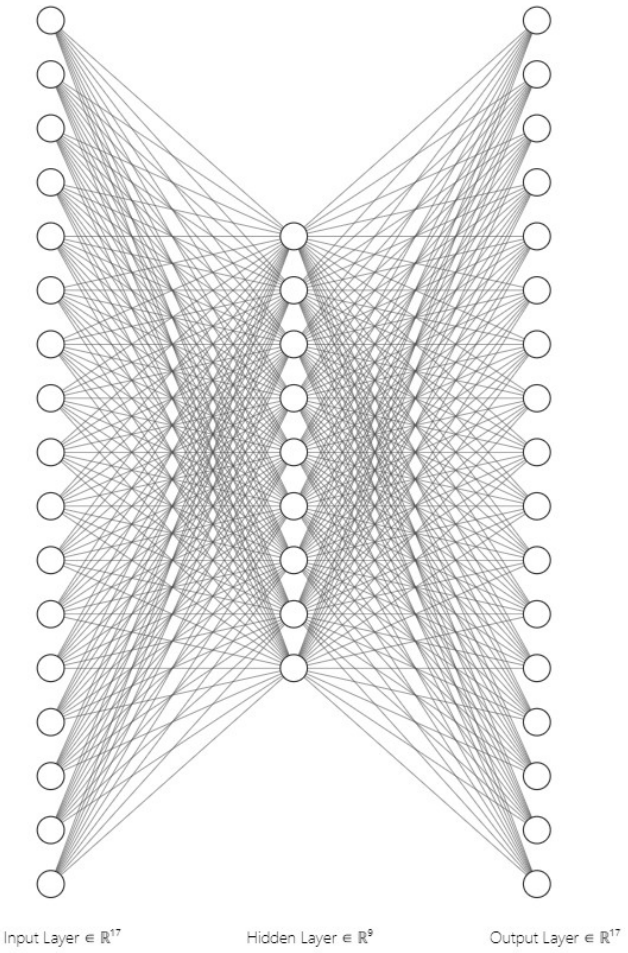


**Figure 3.** Structure of a multilayer perceptron (MLP).

## ARIMA model

The functioning of the ARIMA model is explained by Equation 2, where the moving average (MA) process is presented (Bueno, 2018).

$$y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1} \quad (2)$$

where $\varepsilon_t$ is a white noise that represents the stochastic process error, $\mu$ is the parameters of autoregressive portion of model and $\theta$ is parameters of moving average. As the process $y_t$ is a function of the contemporary error $\varepsilon_t$ at time t, and the immediately preceding error $\varepsilon_{t-1}$, it is said that this is a process of moving averages of order one, called MA (1), where q = 1. If this same process depended on the error of two previous steps $\varepsilon_{t-2}$, it would be of order two, MA (2), where q = 2.

Still a process of moving averages (MA) of order q, that is, its generalized form for q lags is represented by Equation 3 below (Bueno, 2018).

$$y_t = \mu + \sum_{j=0}^{q} \theta_j \varepsilon_{t-j}, \theta_0 = 1 \quad (3)$$

Evolving in the processes that constitute the prediction model, we have the autoregressive processes (AR), which are represented by Equation 4, where $\Phi$ is, like $\mu$, parameters of autoregressive portion of model and $\varepsilon_t$ is the contemporary error (Bueno, 2018).

$$y_t = c + \Phi y_{t-1} + \varepsilon_t \quad (4)$$

Thus, it is white noise that represents the error, but as this stochastic process depends on previous observations of its time series, it is called autoregressive of order one (p = 1), AR(1). If this process depended on two previous observations this would be an AR(2), where p = 2, and so on. Finally, an autoregressive process of order p is given by Equation 5.

$$y_t = c + \Phi_1 y_{t-1} + \Phi_2 y_{t-2} + \ldots + \Phi_p y_{t-p} + \varepsilon_t = c + \sum_{j=1}^{p} \Phi_j y_{t-j} + \varepsilon_t \quad (5)$$

By combining the two types of processes, AR(p) and MA(q), we have the ARIMA (p,q) process, the so-called autoregressive moving averages, according to Equation 6.

$$y_t = c + \sum_{j=1}^{p} \Phi_i y_{t-i} + \sum_{j=0}^{q} \theta_j \varepsilon_{t-j} \quad (6)$$

In this way, a process is obtained in which it is possible to know the variable of interest at time "t" based on errors from previous times and based on its own observed values, that is, from its own historical series (Bueno, 2018).

## Development of the program for time series analysis and data forecasting

After preliminary processing of the data (pressure and discharge rate obtained from field tests), the construction of the program began using the R language to perform a statistical analysis (with data prediction bias) and the prediction itself. Figure 4 shows the graph of the discharge rate of the algorithm developed in this work.

The entire program was developed in the RStudio programming environment. First, the data was imported, which was in .xls format, using the read.xlsx tool from the xlsx() library (Dragulescu & Arendt, 2020). The program then processed the data, transforming the data values from each node in the network into a time series using the ts function that belongs to the prediction library (Hyndman et al., 2020).

Still at this stage, the time series were separated into testing periods (September 21st to 25th) and training periods (September 26th and 27th), using the window() function. These datasets were stored in new variables. Subsequently, the entire time series was decomposed for statistical analysis, related to forecasting models. For this decomposition, the decompose function from the prediction library was also used.

Then, using the training data set of all analyzed time series, the ARIMA model was created for each node in the network, using the auto.arima function also from the prediction library. With the models established, the prediction function of the same library was applied to them. This has the function of forecasting the data, based on the model created previously.

Then, using the precision function from the prediction library, statistical performance metrics were calculated, calculating the efficiency of each ARIMA prediction model and observing the difference between the values generated in the prediction and the separate values for the test sets. This same sequence was used to create the MLP models (creation of models using the training sets, and application of the prediction function to predict the data), using the nntar function, and the precision function to calculate the performance metrics of these models .

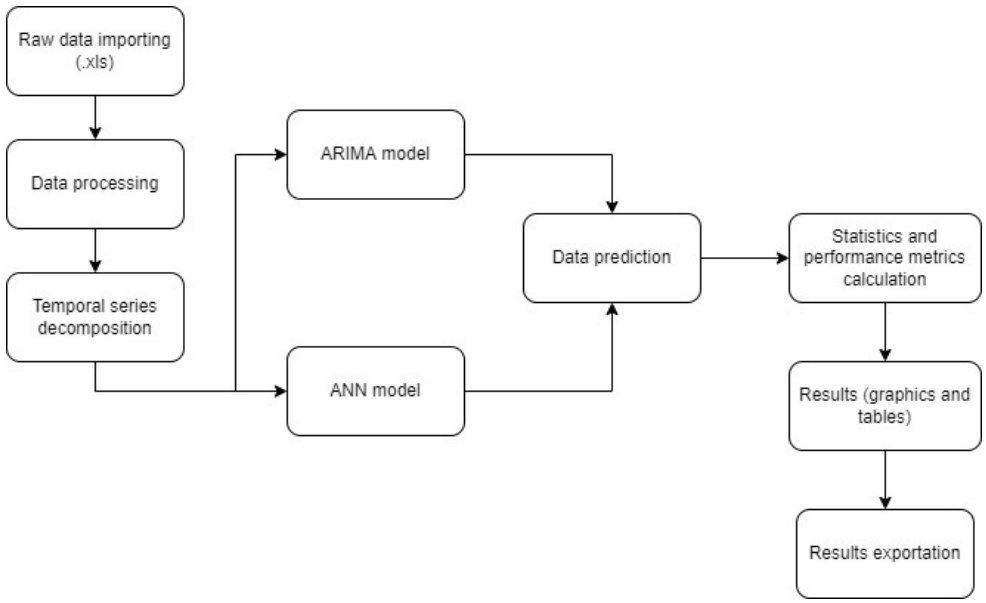


**Figure 4.** Operation diagram of the program developed in this work to make predictions of hydraulic parameters.

Finally, all graphs were plotted, placing the original dataset in contrast to the predicted data, which was exported in figure format. Statistical performance metrics were also exported in the form of tables. To evaluate the efficiency of the models, statistical performance metrics were calculated, which can be observed in Chai & Draxler (2014), namely:

- MAE - Mean Absolute Error: This statistical performance metric calculates the difference between the values presented by the forecast and the actual measured values (in absolute terms) and then applies an average to this data set given by Equation 7 where n is the number of samples and $e_i$ is the error obtained for each iteration i. Therefore, the lower the MAE value, the better;

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|e_i| \quad (7)$$

- RMSE - Mean squared error. Calculates the standard deviation of the sample, that is, the difference between what was predicted and what was measured, therefore, the lower the RMSE value, the better. As can be seen in Equation 8, n is the number of samples and is the error obtained for each iteration i;

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}e_i^2} \quad (8)$$

- MAPE - Mean absolute percentage error, essentially being the mean absolute error (MAE) in percentage terms.

## RESULTS AND DISCUSSIONS

### Program created to generate the data prediction models

The result of this research was a program used to generate data prediction models, as well as to perform the prediction. Thus, Figure 5 shows its structure in detail through a graph of discharge rates.

This structure allowed the program to perform its function successfully, in an optimized way. The step that consumed the most time and processing was installing the library packages, but this step is only performed the first time the program uses these functions. Therefore, in general, this structure with its own programming language can be recommended for the development of future applications, as it generates lightweight files that are quick to be executed even on notebook-type personal computers and is easy to implement and, due to its modularity, can be adjusted for other variables.

### Simulations generated from the ARIMA model

For the nodes where pressures were measured (nodes 14 to 116), the average, maximum and minimum of the performance metrics were calculated.

With these data, it is observed that the ARIMA model constructed has a good representation, that is, it fits well to the data sets. This statement is based on MAPE, since taking the average of this indicator across all nodes in the network that measured pressure, there is an error of only 2.48% in the training period and there is still a maximum error of 5.93%.

In the training stage, the program itself builds the model, using data from the training set as a reference. Therefore, as the program at this stage has this set as a reference, it is expected that the performance metrics will be low, since the main objective of the model is to minimize these metrics.

By analyzing the test set, the program generates the data prediction without having a reference to reach as an answer. The model generates the prediction and then the program calculates the difference between the generated prediction and the test set that was previously separated. Error measurements were higher as the model tries to predict data for an uncertain future. But even so, the data shows that good predictions were obtained, as average errors of 8.54% (MAPE) were calculated.

Considering all the analyzes provided above, the two best predictions generated by the ARIMA prediction model from the point of view of performance metrics were selected, which were node 14 and node 111, presented in Figure 6 and Figure 7. And the worst prediction (node 51), according to statistical metrics presented in Figure 8. It is important to highlight that there is uncertainty in the forecasts (blue line), explained by the light gray (95% of confidence) and blue (90% of confidence) shaded areas in the graphs, as well as remember that the comparison of forecasts being better or worse is based on the metrics, which give an idea of how well the model fits the predictions.

Combining the graphical analysis of the data with the analysis of statistical performance metrics, it was identified that the node that obtained the best result was 111, as the graph showed that the peak demand movements were well represented and between the nodes with the same characteristics as the forecast result, they obtained the lowest RMSE (1.97 m) and MAPE (1.65%).

It is possible to verify that the worst forecast from the point of view of statistical performance indicators was for node 51. However, it is possible to notice that the actual measured data underwent a change because the behavior of the measured pressure did not follow the trend of the previous days. Otherwise, this forecast would likely have much better statistical performance indicators. This anomaly in the pressure behavior occurred due to the functioning of the system at the time of measurement, therefore the prediction was impaired due to such an event.

### Simulations generated from the MLP model

To understand the dataset through an overview, you can check at supplemental material (tables), which show the average, maximum and minimum of the statistical performance metrics, for the nodes where pressures in m were measured.

By observing in supplemental material (tables), it is possible to identify that the results obtained presented considerable errors, with an average value of 16.47% (MAPE), and a maximum value of approximately 60%. By analyzing the statistics of the

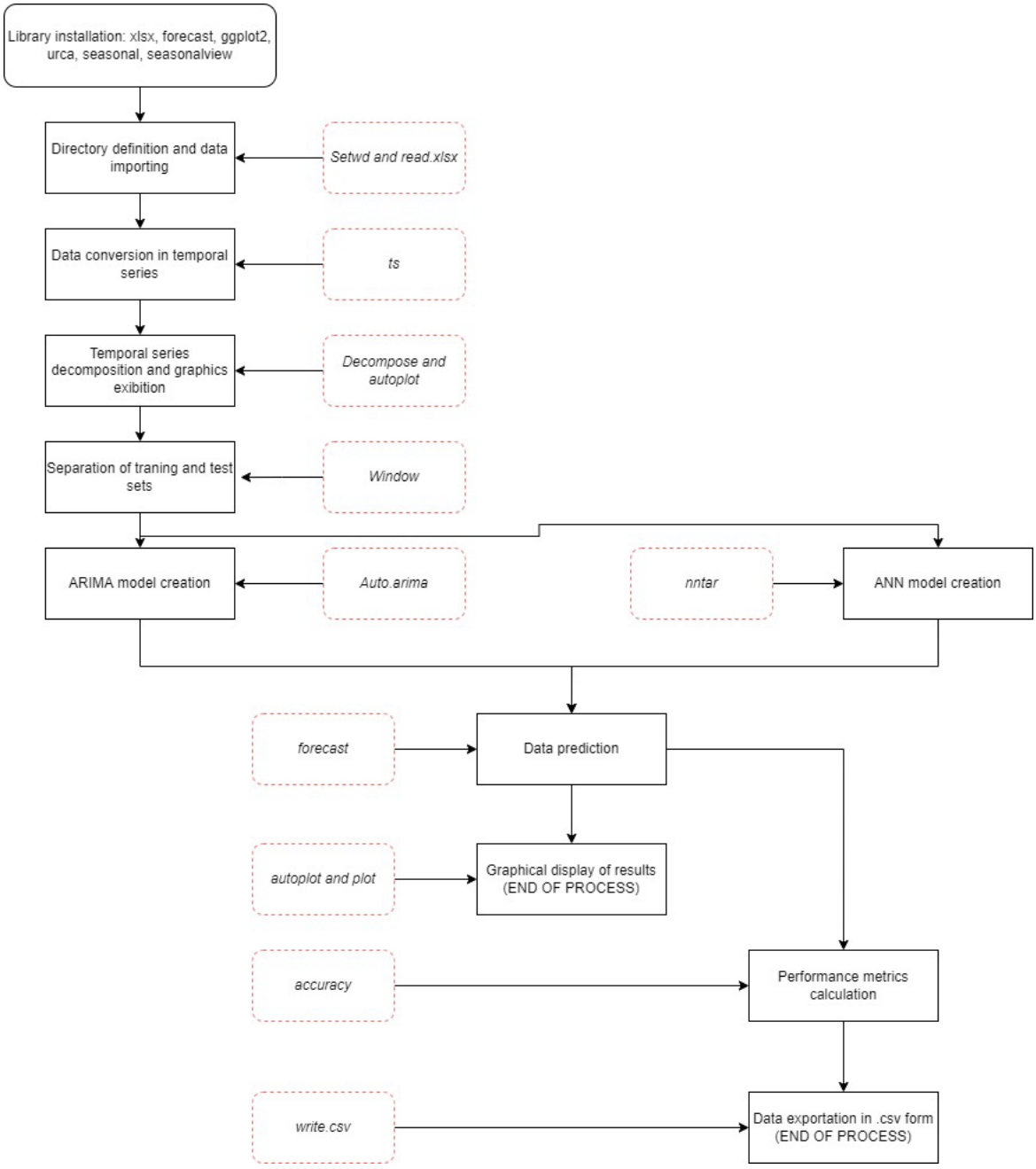


**Figure 5.** Program developed in R language for data prediction.

performance metrics, it is now possible to identify that the model did not fit the data set.

Furthermore, nodes 51, 72, 103, 112, 113, 116 and Vila Nova presented MLP models that were not able to make the forecast, that is, they did not follow the peak demand movements. Thus, the best predictions for the MLP models were selected, which were for nodes 14 and 109, represented in Figure 9 and Figure 10. The worst prediction generated by the MLP model was also selected, which was node 113, represented in Figure 11.

After analyzing all the data presented for the MLP models, it was verified that the same situation mentioned above, in the case of the ARIMA models, occurred for node 14. In other words, this node was the one that presented the lowest RMSE (1.66 m) and MAPE (1.76%), however, from the graphical analysis, it

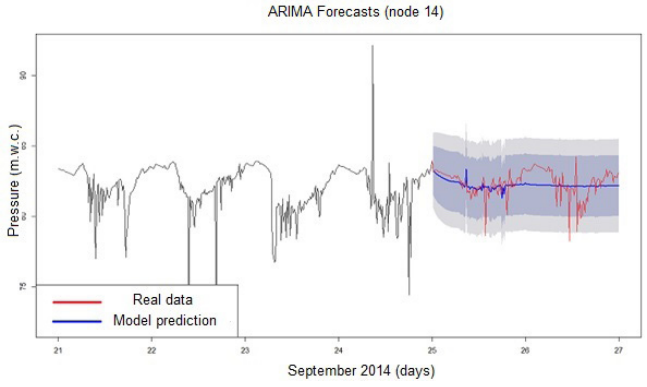


**Figure 6.** Forecast generated by the mathematical model ARIMA, in node 14.

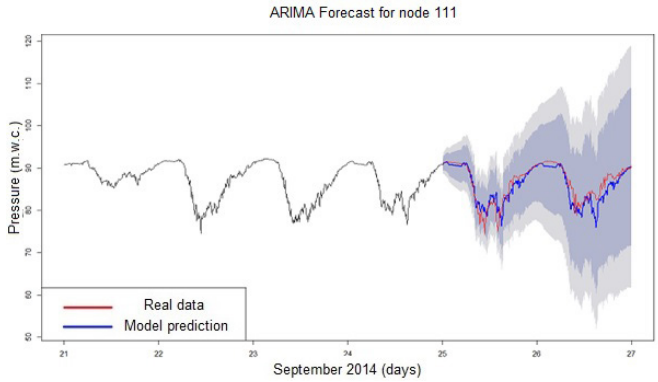


**Figure 7.** Forecast generated by the mathematical model ARIMA, at node 111.

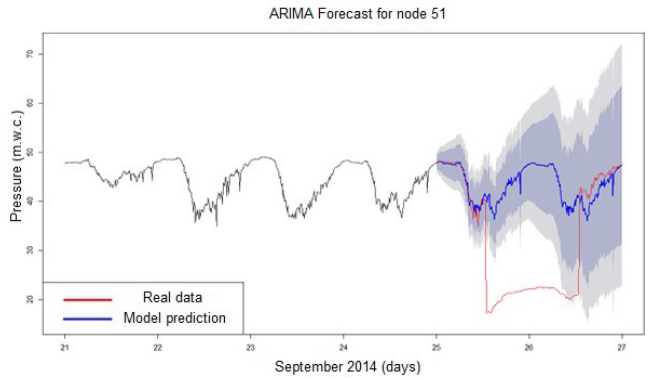


**Figure 8.** Forecast generated by the mathematical model ARIMA, at node 51.

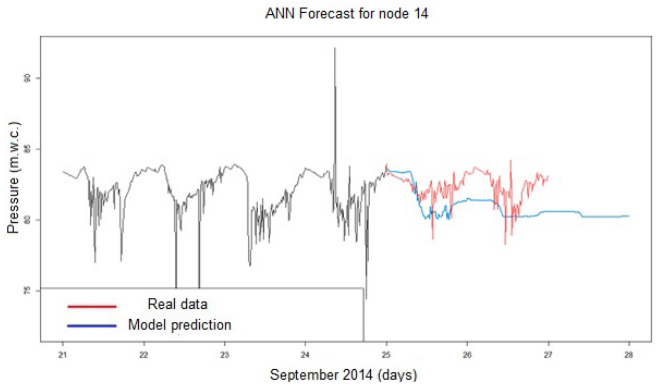


**Figure 9.** Forecast generated by the MLP model, at node 14.

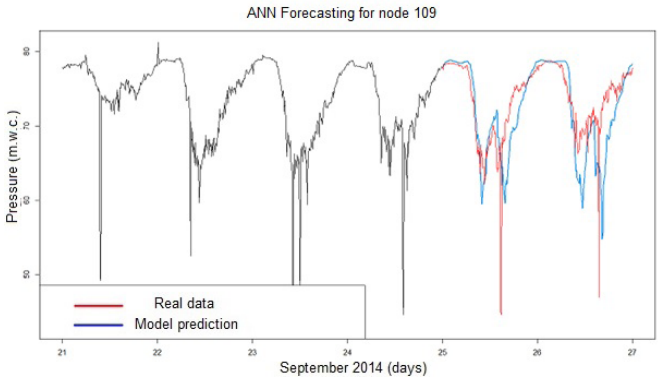


**Figure 10.** Forecast generated by the MLP model, at node 109.

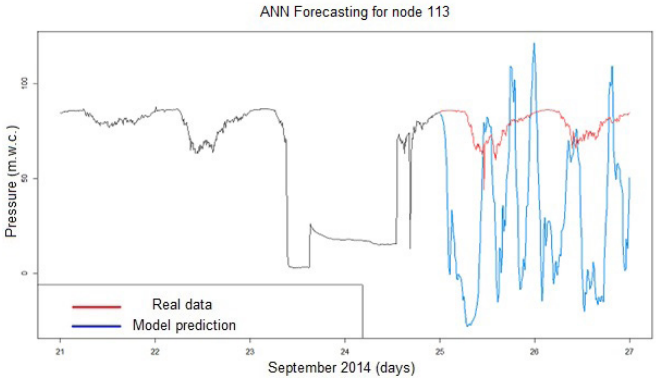


**Figure 11.** Forecast generated by the MLP model, at node 113.

was clear that this node was not the one that presented the best prediction. This was expected, as the data set from node 14 shows little variation and the forecast results were around the average, failing to represent peak demand movements.

Finally, it was possible to observe that the node with the best prediction was 109, as it was able to represent the peaks in demand of network users, their consumption pattern and presented the lowest RMSE (3.63 m) and MAPE (36%), when compared to the other nodes. Node 113 had the worst forecast, but here again it was found that the forecast was hampered by events related to network operation, which caused anomalous pressure values and generated data which, in turn, caused disturbances in the MLP mode.

## Comparison of performance of ARIMA × MLP models

For better understanding, the results were separated into:

- Satisfactory: Forecasts that were able to express peak demand movements and consumption patterns of network users;
- Unsatisfactory: Forecasts that somehow failed to represent, in one or more forecast periods, peak demand movements and consumption patterns of network users;
- No significant forecast: The forecast results did not keep up with the data in the testing period and are considered non-significant.

Thus, it was possible to count the actions as shown in Figure 12.

In terms of the volume of forecasts considered satisfactory, the ARIMA model stands out, presenting more than twice as many good forecasts. As for unsatisfactory results, that is, those in which the forecasts came to follow the data measured in some period of the forecast, but which in general did not go well, the number is very close. Therefore, it appears that the reason for the ARIMA model to stand out in relation to the MLP model, for this specific work, were the models that did not present a significant prediction.

Table 2 shows that the statistical performance metrics are in agreement with the visual analysis, as on average the RMSE and MAPE of the MLP models are twice that of the ARIMA models.

**Table 2.** Comparison between statistical performance metrics.

| AVERAGE | MLP | | ARIMA | |
|---|---|---|---|---|
| | RMSE (m) | MAPE (%) | RMSE (m) | MAPE (%) |
| Training | 1.52 | 1.88 | 2.34 | 2.48 |
| Test | 11.32 | 16.47 | 5.02 | 8.54 |
| MAX | RMSE | MAPE | RMSE | MAPE |
| Training | 2.83 | 4.35 | 4.68 | 5.93 |
| Test | 56.64 | 59.96 | 16.06 | 55.17 |
| MIN | RMSE | MAPE | RMSE | MAPE |
| Training | 0.43 | 0.56 | 0.95 | 0.67 |
| Test | 1.66 | 1.76 | 0.88 | 0.79 |

**Source:** Author (2020).

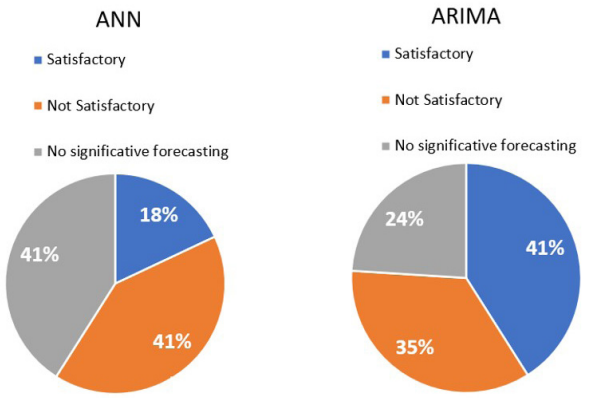


**Figure 12.** Comparison between satisfactory and unsatisfactory results and those that did not present prediction.

Finally, taking into account all the analyzes presented above, it is possible to state that for this specific work, considering the applied artificial neural network (unidirectional with hidden layer, which is provided by the nnetar function, from the prediction package) the best model for prediction of hydraulic parameters using only its own time series, this is the ARIMA model, given its simplicity of implementation through the prediction library in R language, standing out as a tool for evaluating losses caused by excess pressure on water distribution networks with potential for future development aimed at developing an application or computer program.

## CONCLUSIONS

In general, it was possible to verify that these prediction techniques are viable for these types of data. Furthermore, it was found that among the ARIMA models, the one that presented the best prediction was the model applied to node 111, as it obtained the lowest RMSE (1.97 m) and MAPE (1.65%) values. In the case of the MLP models, it was node 109 with the lowest RMSE (3.63 m) and MAPE (3.36%) values.

Even so, the importance of combining graphical analysis with the analysis of performance indicators was justified, as for node 14 the lowest values of statistical performance indicators occurred. However, the graphs indicate that the forecast was unable to identify peak demand movements and consumption patterns. This fact was caused by the low pressure variation at the node, and the forecasts present values close to the average pressure.

After all these steps, it appears that both techniques were able to generate data predictions and the ARIMA model was determined as the one that best represents the data prediction for this water supply network in question. It was also possible to observe that the MLP used in this work is not the best for this type of forecast. Therefore, for future work, it is recommended to use more robust neural networks that allow adjusting the activation functions of neurons, as well as the number of hidden layers.

Finally, ARIMA models are simpler to generate, more tools can build this type of model and they have been applied for longer, therefore they have the advantage of having more knowledge added over the years. Therefore, this type of model represents well the prediction of a real water supply network, as it has an average MAPE of 8.54%.

For future work, it is suggested to evaluate the performance of hybrid models such as the combination of Genetic Algorithms and MLP. It is also suggested to use MLP models with a greater number of hidden layers, since the nntar function uses only one hidden layer to create the model, as well as the possibility of using Long Short-Term Deep Learning Neural Networks Memory type to evaluate new predictions against the ARIMA model.

## ACKNOWLEDGEMENTS

Alex Takeo Yasumura Lima Silva thanks the Federal University of Itajubá for the Institutional Doctoral Scholarship nº 23088.037588/2021-72. Sara Maria Marques thanks CAPES for the Master's Scholarship nº 8887.614858/2021-00. Authors thanks Redecope FINEP Project – MCT (Ref. 0983/10) - Ministry of Science and Technology, entitled "Development of efficient technologies for hydro-energy management in water supply systems"; and Program "Pesquisador Mineiro" from FAPEMIG for the PPM - 00755-16. NUMMARH – Center for Modeling and Simulation in Environment and Water Resources and Systems (NUMMARH) of UNIFEI.

## REFERENCES

Adamowski, J., Fung Chan, H., Prasher, S. O., Ozga-Zielinski, B., & Sliusarieva, A. (2012). Comparison of multiple linear and nonlinear regression, autoregressive integrated moving average, artificial neural network, and wavelet artificial neural network methods for urban

water demand forecasting in Montreal, Canada. *Water Resources Research*, *48*(1), 1-14. http://doi.org/10.1029/2010WR009945.

Alizadeh, Z., Yazdi, J., Mohammadiun, S., Hewage, K., & Sadiq, R. (2019). Evaluation of data driven models for pipe burst prediction in urban water distribution systems. *Urban Water Journal*, *16*(2), 136-145. http://doi.org/10.1080/1573062X.2019.1637004.

Amaral, H. L. M. (2020). *Desenvolvimento de uma nova metodologia para previsão do consumo de energia elétrica de curto prazo utilizando redes neurais artificiais e decomposição de séries temporais* (Tese de doutorado). Escola Politécnica, Universidade de São Paulo, São Paulo.

Awad, M., & Zaid-Alkelani, M. (2019). Prediction of water demand using artificial neural networks models and statistical model. *International Journal of Intelligent Systems and Applications*, *11*(9), 40-55. http://doi.org/10.5815/ijisa.2019.09.05.

Bo, Z., Yezheng, L., & Feifei, Z. (2021). Annual water consumption forecast of Hefei based on ARIMA model. *Academic Journal of Computing & Information Science*, *4*(3), 88-93.

Bueno, R. L. S. (2018). *Time series econometrics* (2nd ed., 360 p.). São Paulo: Cengage Learning.

Chai, T., & Draxler, R. R. (2014). Root mean square error (RMSE) or mean absolute error (MAE)? Arguments against avoiding RMSE in the literature. *Geoscientific Model Development*, *7*(3), 1247-1250. http://doi.org/10.5194/gmd-7-1247-2014.

Dragulescu, A., & Arendt, C. (2020). *xlsx: read, write, format Excel 2007 and Excel 97/2000/XP/2003 files. R package version 0.6.3*. Vienna: R Foundation for Statistical Computing. Retrieved in 2023, May 19, from https://CRAN.R-project.org/package=xlsx

Du, H., Zhao, Z., & Xue, H. (2020). ARIMA-M: a new model for daily water consumption prediction based on the autoregressive integrated moving average model and the markov chain error correction. *Water*, *12*(3), 1-20. http://doi.org/10.3390/w12030760.

Gharabaghi, S., Stahl, E., & Bonakdari, H. (2019). Integrated nonlinear daily water demand forecast model (case study: City of Guelph, Canada). *Journal of Hydrology*, *579*(1), 1-18. http://doi.org/10.1016/j.jhydrol.2019.124182.

Ghosal, P. S., Javaregowda, A., Gupta, A. K., & Singh, D. P. (2019). A novel framework of multivariate modeling of water distribution network through 3 3 factorial design and artificial neural network. *Journal of Environmental Science and Health. Part A, Toxic/Hazardous Substances & Environmental Engineering*, *54*(6), 551-562. http://doi.org/10.1080/10934529.2019.1571308.

Guarnaccia, C., Longobardi, A., Mancini, S., & Viccione, G. (2020). Drinking water tank level analysis with ARIMA models: a case study. *Environmental Sciences Proceedings*, *2*(1), 33.

Hyndman, R., Athanasopoulos, G., Bergmeir, C., Caceres, G., Chhay, L., O'Hara-Wild, M., Petropoulos, F., Razbash, S., Wang, E., & Yasmeen, F. (2020). *Forecast: forecasting functions for time series and linear models. R package version 8.12*. Retrieved in 2023, May 19, from http://pkg.robjhyndman.com/forecast

Jang, D., & Choi, G. (2017). Estimation of non-revenue water ratio using MRA and ANN in water distribution networks. *Water*, *10*(1), 2. http://doi.org/10.3390/w10010002.

Jetmarova, H., Sultanova, N., & Savic, D. (2017). Lost in optimization of water distribution systems? A literature review of system operation. *Environmental Modelling & Software*, *93*, 209-254. http://doi.org/10.1016/j.envsoft.2017.02.009.

Kamiński, K., Kamiński, W., & Mizerski, T. (2017). Application of artificial neural networks to the technical condition assessment of water supply systems. *Ecological Chemistry and Engineering. S*, *24*(1), 31-40. http://doi.org/10.1515/eces-2017-0003.

Lima, G., Brentan, B. M., Manzi, D., & Luvizotto Junior, E. (2018). Metamodel for nodal pressure estimation at near real-time in water distribution systems using artificial neural networks. *Journal of Hydroinformatics*, *20*(2), 486-496. http://doi.org/10.2166/hydro.2017.036.

Lopez Farias, R., Puig, V., Rodriguez Rangel, H., & Flores, J. (2018). Multi-model prediction for demand forecast in water distribution networks. *Energies*, *11*(3), 660. http://doi.org/10.3390/en11030660.

Lorente-Leyva, L. L., Pavón-Valencia, J. F., Montero-Santos, Y., Herrera-Granda, I. D., Herrera-Granda, E. P., & Peluffo-Ordóñez, D. H. (2019). Artificial neural networks for urban water demand forecasting: a case study. *Journal of Physics: Conference Series*, *1284*(1), 012004. http://doi.org/10.1088/1742-6596/1284/1/012004.

Pandey, P., Bokde, N. D., Dongre, S., & Gupta, R. (2021). Hybrid models for water demand forecasting. *Journal of Water Resources Planning and Management*, *147*(2), 04020106. http://doi.org/10.1061/(ASCE)WR.1943-5452.0001331.

Porto, V. C., Souza Filho, F. A., Carvalho, T. M. N., Studart, T. M. C., & Portela, M. M. (2021). A GLM copula approach for multisite annual streamflow generation. *Journal of Hydrology*, *598*, 126226. http://doi.org/10.1016/j.jhydrol.2021.126226.

Shirkoohi, M. G., Doghri, M., & Duchesne, S. (2021). Short-term water demand predictions coupling an artificial neural network model and a genetic algorithm. *Water Science and Technology: Water Supply*, *21*(5), 2374-2386. http://doi.org/10.2166/ws.2021.049.

Vieira, L. T. S. (2019). *Análise e avaliação do comportamento de parâmetros hidráulicos de uma rede de distribuição de água do Sul de Minas Gerais* (Dissertação de mestrado). Universidade Federal de Itajubá, Itajubá. Retrieved in 2020, April 25, from https://repositorio.unifei.edu.br/xmlui/bitstream/handle/123456789/1967/dissertacao_2019076.pdf?sequence=1&isAllowed=y

Wu, Y., & Liu, S. (2017). A review of data-driven approaches for burst detection in water distribution systems. *Urban Water Journal*, *14*(9), 972-983. http://doi.org/10.1080/1573062X.2017.1279191.

Xu, G., Cheng, Y., Liu, F., Ping, P., & Sun, J. (2019). A water level prediction model based on ARIMA-RNN. In *Proceedings of the 5th IEEE International Conference on Big Data Service and Applications (BigDataService)* (pp. 221-226), Newark, CA, USA. New York: IEEE.

Xu, Z., Ying, Z., Li, Y., He, B., & Chen, Y. (2020). Pressure prediction and abnormal working conditions detection of water supply network based on LSTM. *Water Science and Technology: Water Supply*, *20*(3), 963-974. http://doi.org/10.2166/ws.2020.013.

Zhou, X., Tang, Z., Xu, W., Meng, F., Chu, X., Xin, K., & Fu, G. (2019). Deep learning accurately pinpoints burst locations in water distribution networks. *Water Research*, *166*, 115058. http://doi.org/10.1016/j.watres.2019.115058.

Zubaidi, S. L., Dooley, J., Alkhaddar, R. M., Abdellatif, M., Al-Bugharbee, H., & Ortega-Martorell, S. (2018). A Novel approach to predicting monthly water demand by combining singular spectrum analysis with neural networks. *Journal of Hydrology*, *561*, 136-145. http://doi.org/10.1016/j.jhydrol.2018.03.047.

Zubaidi, S. L., Al-Bugharbee, H., Muhsen, Y. R., Hashim, K., Alkhaddar, R. M., & Hmeesh, W. H. (2019). The prediction of municipal water demand in Iraq: a case study of baghdad governorate. In *Proceedings of the 12th International Conference on Developments in eSystems Engineering (DeSE)* (pp. 274-277), Kazan, Russia. New York: IEEE.

## Authors contributions

André Carlos da Silva: Development of ARIMA and ANN algorithms, data processing, obtaining and processing of results.

Fernando das Graças Braga da Silva: Creator of the work, obtaining and reviewing data, reviewing results.

Victor Eduardo de Mello Valério: Assistance in the development of ARIMA and ANN algorithms, treatment and review of results.

Alex Takeo Yasumura Lima Silva: Review of results, writing of the article.

Sara Maria Marques: Review of the ARIMA and ANN algorithms, review of the results.

José Antonio Tosta dos Reis: Writing of the article, review of the ARIMA and ANN algorithms, review of the results., reviewing the data.

**Editor-in-Chief:** Adilson Pinheiro

**Associated Editor:** Carlos Henrique Ribeiro Lima