27th International Congress of Mechanical Engineering
December 4-8, 2023, Florianópolis, SC, Brazil

COB-2023-2036
COMBINED MACHINE LEARNING AND DECOMPOSITION MODELS BY PREDICTING URBAN WATER CONSUMPTION IN CURITIBA

Wilton Sergio Morais Brayner
Mechanical Engineering Graduate Program (PPGEM), Pontifical Catholic University of Paraná (PUCPR), Imaculada Conceição, Curitiba, Paraná, Brazil
wilton.brayner@yahoo.com.br
Viviana Cocco Mariani
Mechanical Engineering Graduate Program (PPGEM), Pontifical Catholic University of Paraná (PUCPR) and Federal University of Paraná (UFPR), Coronel Francisco Heraclito dos Santos, Curitiba, Paraná, Brazil
viviana.mariani@pucpr.br
Anderson Schamne
Gustavo Rafael Collere Possetti
Sanitation Company of Paraná (SANEPAR), Dr Celso Luiz Peixoto Ribas, Curitiba, Paraná, Brazil
andersons@sanepar.com.br, gustavorcp@sanepar.com.br

Abstract. *Forecasting water consumption is an extremely important factor for the planning of the agencies that distribute water to each neighborhood in each city. Recently, sanitation companies have invested in automating their water supply systems, which extract and store water consumption data in real-time. To make correct decisions regarding demand forecasting, it is necessary to have prior knowledge of consumption throughout the days considering seasonality such as holidays, weekends, also different needs according to the seasons of the year. According to the Institute for Research and Urban Planning in Curitiba now is formed seventy-five neighborhoods being Bairro Alto one of the most populated of the city, with forty-two thousand inhabitants, which represents 2.15% of the total population, the neighborhood studied in this research. This paper aims to present different short-term water consumption forecasting models using different forecasting techniques such as Holt-Winters, Autoregressive Integrated Moving Average (ARIMA), Random Forest, and XGBoost for different forecasting horizons, one, seven, fourteen, and twenty-eight steps ahead for daily data in one time series about water distribution for SANEPAR to Bairro Alto in Curitiba. These predictions are based on historical data collected over three consecutive years to predict urban water consumption in the previously mentioned neighborhood. The partial results obtained were 0.39 for the R^2 with daily data and 0.96 for hourly data considering the ARIMA model and 0.28 for the R^2 with daily data and 0.92 for hourly data considering the Holt-Winters model.*

Keywords: *Time Series Forecasting, Machine learning, Decomposition models, Water supply.*

1. INTRODUCTION

Water is a vital element for the preservation of human life and obtaining a forecast model of its consumption to optimize its distribution brings several advantages, such as the identification of peaks, trends, and seasonality, which makes the distribution according to demand, reducing costs and increasing productivity (MACHADO et al., 2016).

In recent years, sanitation companies have invested in the automation of water supply systems, and this has provided access to information on flow, pressure, frequency, and demand forecast for water use in several cities in Brazil. Once this automation has been achieved, it is possible to analyze the operation of the Water Supply Systems (SAA) to improve the services for capturing, treating, and distributing water, consequently reducing related costs, such as energy, chemicals, and equipment wear (Falkenberg, 2005).

According to the IBGE (2019), in recent years, the scarcity of water resources has been news present in the reality of Brazil. The alerts of water crisis and water rotation were constant in each residence, which made the state governments look with a sense of urgency to the analysis of water distribution demand, through their historical data of supply of the last years.

Forecast techniques based on historical data were applied for this purpose, among which the Holt-Winters models (Lima et al., 2019), ARIMA (Babu, 2014), Random Forest (Tyralis, 2017), and XGBoost, as the work developed by Sun (2022). In this work, data from the Sanitation Company of Parana (SANEPAR) for water supply in Bairro Alto, in the city of Curitiba, for the years 2018, 2019, and 2020 were extracted for forecasting analysis of time series. Due to the Corona Virus (COVID-19) pandemic, there was disorderly behavior in the series in 2020, this year will not be

considered for analysis. According to the Institute for Research and Urban Planning (IPPUC, 2010), the city of Curitiba has seventy-five neighborhoods and Bairro Alto has an area of seven square kilometers and is among the most populous in the city, with forty-two thousand inhabitants, which represents 2.15% of the total population.

In Brazil, at the beginning of the 2000s, studies were also built regarding the forecasting of time series. Falkenberg (2005) compares several techniques such as artificial neural networks, multiple linear regression, and models like Box and Jenkins to forecast water consumption in three areas with different consumption profiles in the city of Ponta Grossa, Paraná. By applying these forecasting techniques to regions with different consumption characteristics, he found that not always a technique or a single model is a good solution for all cases. It concludes that, with the application of the data, it is possible to obtain a short-term forecast with boosted results to the point of being used in the optimization of water supply systems. Rodríguez (2010) compared ARIMA models with Neural Network models for forecasting urban water demand in populated cities. He proposed these models to study the consumption of a city in the southeast of the United States because it presents different consumption patterns than those that occur in Spanish and Mediterranean cities, where the climate component influences consumption. These cities were urbanized and densely populated, where they use gardens, and public parks, among others, requiring the use of irrigation systems, concluding that the meteorological component was not very relevant.

Manfrin (2020) analyzed water demand forecast models for the city of Joinville using time series analysis, an exponential smoothing model, and a Seasonal Autoregressive Integrated Moving Average (SARIMA) model for each of the analyzed categories (residential, commercial, industrial, public, and total). Their results showed that the SARIMA models performed better in forecasting in the commercial, industrial, public, and total categories, while the exponential smoothing method proved to be more appropriate for forecasting in the residential category. The results obtained in the prediction of the total category indicate that making predictions considering a unified urban consumption can be more assertive.

The remainder of this paper is organized as follows. Section 2 comprises the dataset analyzed and its features. Section 3 comprises a brief description of the dataset analyzed and its features. Section 4 describes the concepts of classic approaches, forecasting models, and evaluation methods employed in this study. Section 5 summarizes the methodology applied. In Section 6, the main results are presented, and discussions are detailed. Last, Section 7 concludes the study by presenting insights about the results and proposing directions for future research.

2. DATASET DESCRIPTION

The dataset regards the variables of a water station system of Bairro Alto, a residential area located in Curitiba, Parana state, Brazil. The dataset was retrieved from the Sanepar and is composed of 9 variables, where there are one target variable and eight input exogenous ones, comprising the years 2018 to 2020, in a daily sampling. Table 1 are presented the variables, their acronyms, and unit measures, respectively.

Table 1. Output and inputs of the water station system dataset.

Type	Description	Acronym	Measurement unit
Output	Water reservoir level	RES	(m)
Input	Suction pump #1	B1	(Hz)
Input	Suction pump #2	B2	(Hz)
Input	Suction pump #3	B3	(Hz)
Input	Input flow	VENT	(m^3/h)
Input	Gravity flow	VGRA	(m^3/h)
Input	Hold flow	VREC	(m^3/h)
Input	Suction pressure	PSUC	(mH_2O)
Input	Hold pressure	PREC	(mH_2O)

Spearman's correlation was used to analyze the data, and for daily and hourly information, the relationship between variables is described in Figure 1:

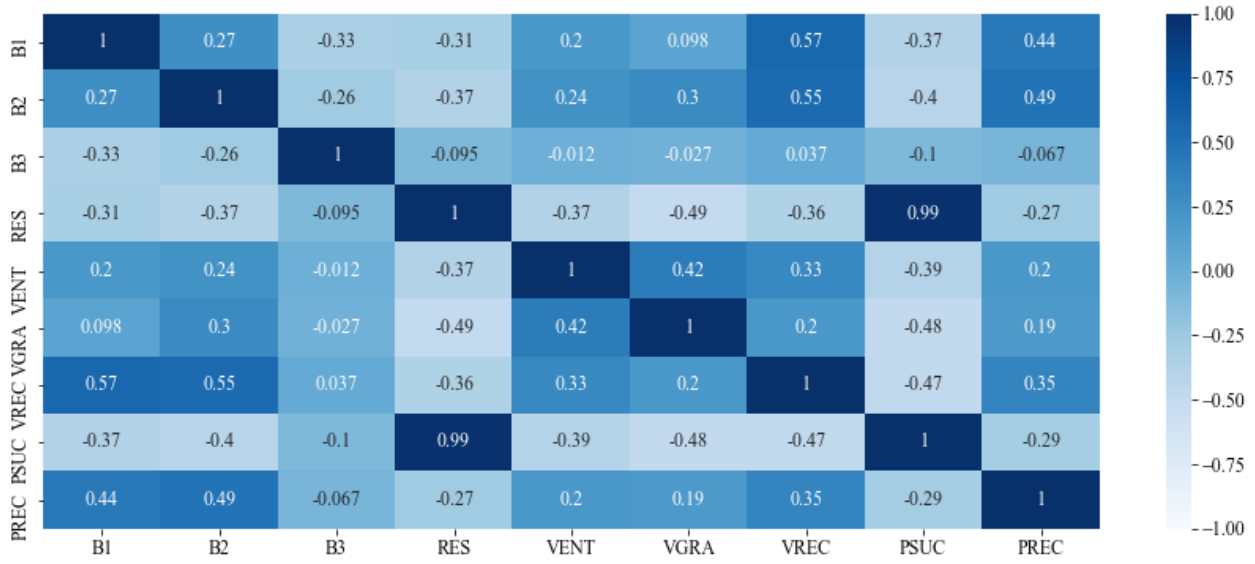


Figure 1. Spearman correlation of the daily data sample in Bairro Alto.

3. METHODS

This section presents the main aspects of the methods proposed in this study. The signal decomposition approaches are presented followed by the description of the forecasting models and the evaluation methods. This study aims at the use of time series forecasting models aiming at the planning and sizing of the SANEPAR water reservoir levels present at the base of Bairro Alto, in Curitiba. As shown in Section 2, the Holt-Winters, ARIMA, Random Forest, and XGBoost models are recognized as fundamental for forecasting short-term consumption, as they bring the advantage of working with time series that show seasonality. Exponential dissection methods are methods that use exponentially decreasing weights with the age of observations.

3.1 Holt-Winters

According to Costa (2015), the Holt-Winters exponential smoothing, or smoothing model is univariate, using only the data series itself to perform the forecast, and, due to its accuracy and robustness, it is applied in several areas, such as energy consumption. electricity and sales forecasts. However, before applying it to forecasts, it is necessary to estimate the initial values of the level, trend, and each seasonality. It is also necessary to determine the parameters known as the model's damping constants. This model is often used to forecast time series that have a trend and seasonality and can be formulated in an additive or multiplicative way. The first is more suitable for series that show constant variance over time (additive seasonality).

3.2 ARIMA

ARIMA (Autoregressive Integrated Moving Average) models are highly structured and applied in water demand studies, especially in a short or very short-term context. They are appropriate for modeling the seasonal pattern in water use, as well as fixed patterns that may occur for daily water use (Billings and Jones, 2008).

ARIMA models come from the combination of autoregression (AR) models, where the variable of interest is projected using a linear combination of past values of the variable, and MA models, where the error values are a weighted moving average of the last errors of the forecast. The I (for "integrated") indicates that the data values were replaced by the difference between their values and the previous values (and this differentiation process may have been performed more than once) (Hyndman and Athanasopoulos, 2013). For the AR, MA, ARMA, and ARIMA models, conditions of stationarity, mean equal to zero, and constant variance are required (Box, Jenkins and Reinsel, 2008). In the case of non-stationary series, the influence of seasonality and trend of time series can be eliminated through differentiation and mathematical transformations, such as the application of logarithms. Later, the transformations can be inverted, and forecasts obtained for the original series by applying a SARIMA model, which adds a seasonal component to the equation (Hyndman and Athanasopoulos, 2013).

The Box-Jenkins methodology investigates the autocorrelation between series values at different successive time points. Autocorrelation patterns, in general, make it possible to identify one or several possible models for the time series (Khashei and Bijari, 2010). When observing the autocorrelation within a period of one year, a season of seasonality is considered, and the original series can be adjusted by a seasonal ARIMA model (Martins and Werner, 2014). The SARIMA models consist of a non-seasonal part (p, d, q) and a seasonal part (P, D, Q). These models are the

most requested for the description of seasonal time series, demonstrating success in their applications in the last 30 years (Chen and Wang, 2007).

3.3 Random Forest

Random Forest (RF) is a joint learning method for classification and regression that operates by building multiple decision trees at training time and producing the class, which is the mode of outputs generated by individual trees (Breiman and Cutler, 2001). The term originates from random decision forests that were first proposed by Ho (1995). This method combines Breiman's bagging idea and random feature selection, introduced by Ho (1995) and Geman (1997) to build a collection of decision trees with controlled variation. According to Stevens (2012), RF algorithms are attractive for the following reasons:

- Can operate with regression and multiclass classification;
- Are fast for training and testing;
- Depend on one to two tuning parameters;
- Have a built-in estimate of the generalization error;
- Can be used for high-dimensional problems.

According to Cutler and Stevens (2012), RF is a joint learning method based on decision trees (each tree depends on a collection of random variables). Thus, for a random p dimension vector $X=(X_1,\ldots, X_p)T$, representing the actual input values, and a random variable Y, representing the actual response value, an unknown joint distribution P_xy is assumed (X, Y). In this sense, the goal is to find a function f(x) to predict Y.

3.4 XGBoost

According to Cai-Xia (2021), XGBoost, a type of reinforcement algorithm, which brings together multiple learning algorithms to achieve better predictive performance than any of the constituent learning algorithms alone, has excelled in many fields. Compared with the traditional algorithm, XGBoost applies a second-order Taylor expansion to the loss function and simultaneously implements the first derivative and second derivative. Furthermore, a regularization term is added to the objective function, which improves the generalization of a single tree and reduces the complexity of the objective function. To sum up, XGBoost has attracted researchers' attention due to its high speed, excellent sorting effect, and ability to enable custom loss functions.

The impact of this algorithm has been widely recognized in various machine learning and data mining challenges. This is because the package offers very good results on different types of problems. Problems solved include sales forecasting, web text classification, customer behavior forecasting, product categorization, etc. All this success is justified by the scalability of the algorithm in all scenarios. Furthermore, XGBoost runs ten times faster than existing popular solutions. This is possible since its algorithm consists of optimization and combinations of techniques such as Gradient Boosting, which uses the gradient descent algorithm to minimize errors in conjunction with ensemble techniques. That is, the implementation of a set of gradient-optimized decision trees is considered (Chen and Guestrin, 2016).

In addition to these characteristics, Chen and Dhaliwal (2018) mention that there are other advantages to using XGBoost, such as:

- Capacity for parallel processing, in which the algorithm uses all the cores of the machine that execute it, making it highly effective for high-level data pre-processing.
- Accepts various types of data as input, which can be a dense or sparse matrix and local files;
- The algorithm is customizable, supporting custom objective and evaluation functions;
- Runs on multiple platforms: AWS, Azure, Python, etc.
- Well-prepared to detect and deal with missing or missing values;
- Support for multiple programming languages.

3.5 Evaluation Metrics

According to PURCOTE (2009), there are many performance measures for analyzing the quality of a time series. However, most describe the difference between the actual and predicted values of the series, known as errors or residuals. For the evaluation of forecast accuracy, the Coefficient of Determination (R^2), Mean Squared Error (MSE), Mean Absolute Percentage Error (MAPE), and Root Mean Error Square (RMSE) will be adopted in this work.

The R^2 measures the degree of association between observed and calculated values; is a number between 0 and 1 and calculated by the equation below:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - y)^2}, \quad (1)$$

where y_i is the actual value, $(\hat{y}_i)$ are the estimated value, and y is the average of N measured samples.

The value of R2 is located between 0 and 1, when its value is equal to 1, there is a perfect approximation of the model to the measured data of the system. The MSE consists of calculating the average of the squared errors, squaring the forecast errors, adding, and dividing them by the number of periods. Therefore, considering that y_i are the actual values, $(\hat{y}_i)$ are the estimated values and n is the number of periods, the MSE calculation is described by the equation below:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{2}$$

Due to the calculation of the mean squared error, large errors stand out when compared to smaller magnitude errors. MSE values equal to zero indicate the exact adequacy of the model for the measured data of the system. To evaluate the magnitude of the error in the historical series, the MAPE calculates the forecast errors in absolute percentages.

Considering y_i the actual values of the series, $(\hat{y}_i)$ the predicted values and N the number of forecast periods, the formula for calculating the MAPE is given by the equation below:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} * 100. \tag{3}$$

The RMSE evaluates the size of the forecast error and is calculated by averaging the errors of each forecast about the true value, squared. The closer the RMSE is to zero, the better the forecast model. Its calculation is defined by the equation below:

$$\text{RMSE} = 1/n\sqrt{\sum_{i=1}^{n}(\hat{y}_i - y_i)^2}. \tag{4}$$

The MAE, in turn, is calculated from the average of the absolute errors, that is, the module of each error is used to avoid underestimation, because the value is less affected by especially extreme points (outliers) through the equation:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|\, y_i - \hat{y}_i|. \tag{5}$$

This measure is used in time series, as there are cases in which the negative error can reset the positive one or give an idea that the model is accurate. But here, only the distance from the actual value is calculated, regardless of whether it is above or below. Given the metrics, it is important to remember that for each model, we can use one or more combined metrics to analyze it, in addition, it can be thought that depending on the chosen metric or metrics, something is waived (whether it is accuracy, treatment of outliers or similar).

In summary, the water reservoir level data passes through a decision. The data are pre-processed by the daily average approaches and split into components, and they go to the training phase to be learned by the forecasting models, otherwise, the data go directly to the training phase. Also, in the training phase, the input exogenous variables are added to the learning process. Next, the predictions are generated and passed through a second decision, the signal is reconstructed and sent to the test phase, otherwise, the predictions go to the test phase. In the last phase, the predictions of the analyzed models are evaluated by performance metrics and hypothesis tests to verify the effectiveness and accuracy of the predictions. All models are compared with each other, and a final prediction is chosen according to the results of evaluation methods.

4. RESULTS AND DISCUSSIONS

This section presents the daily values, obtained by averaging the 24 daily data. Analyzing the data in figure 29, the variables B3 and PSUC were disregarded, due to the low and high correlation, respectively, with the output variable RES. Analyzing figures 1, the values for p, q and d of the ARIMA model are found to be 3, 1 and 2, respectively.

For the analysis of the time series and predictions with the ARIMA, Holt-Winters, Random Forest, and XGBoost models, 70% of the Exposures for training and 30% of the Exposures for testing were used, as they were the most

common divisions found in the literature review. For the daily data, 700 Executions were used, with 481 data in training and 219 data in the test. Figures 2 to 5 show the behavior of the RES forecasts, in m³, using the Holt-Winters, ARIMA, RF, and XGBoost techniques, respectively.

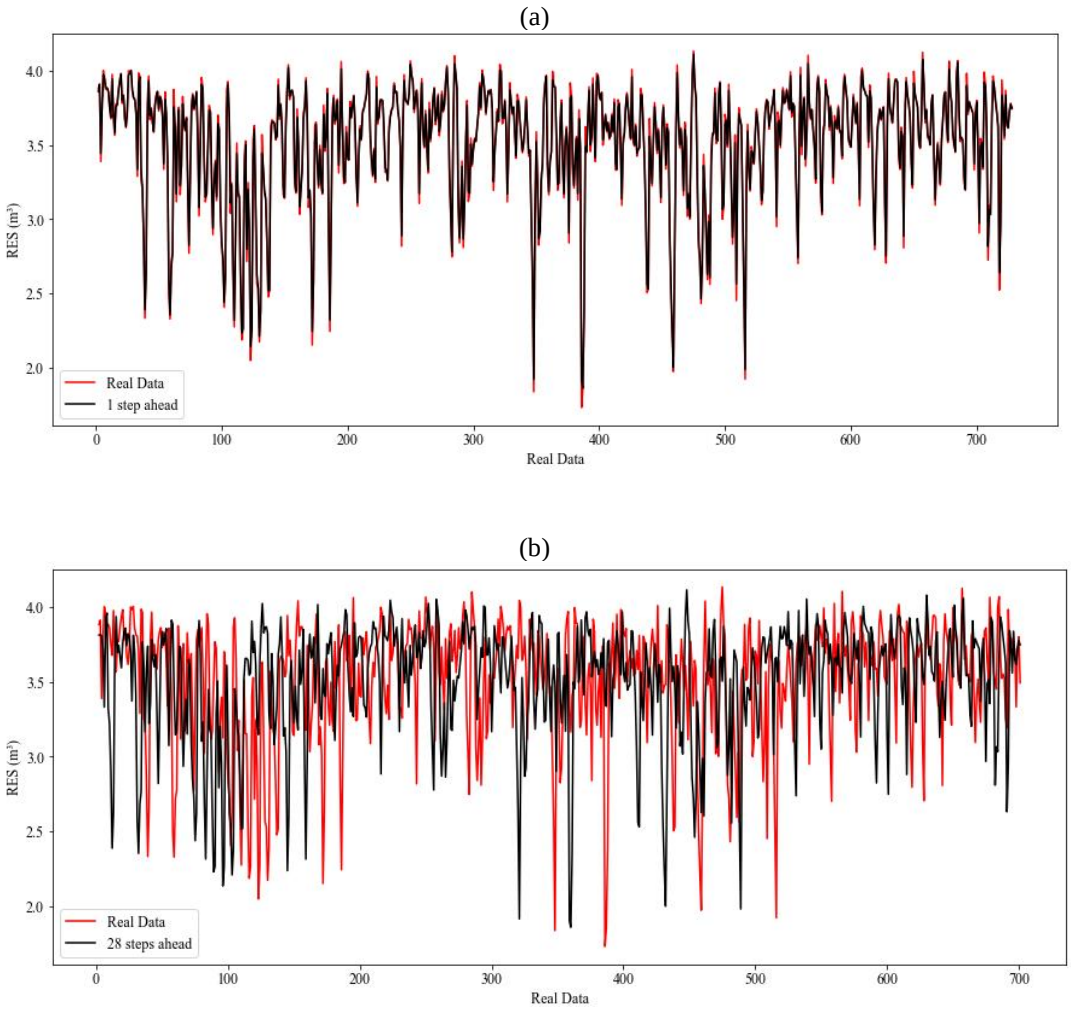


Figure 2. Daily RES data predictions considering the Holt-Winters model for (a) 1 step ahead, (b) 7 steps ahead, (c) 14 steps ahead, and (d) 28 steps ahead.

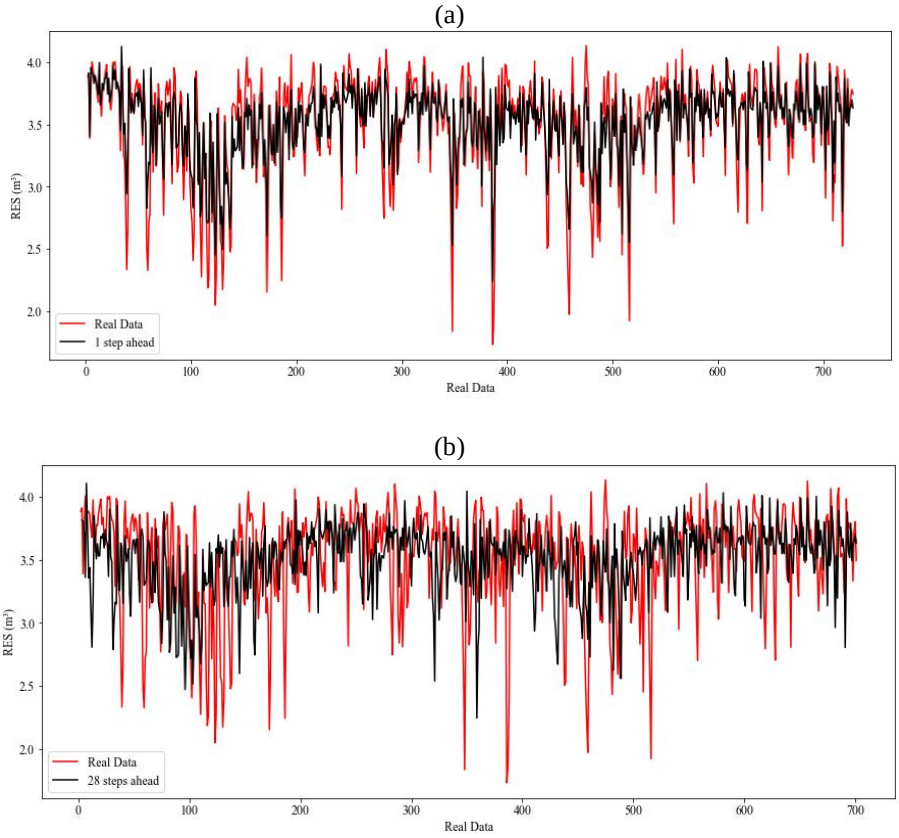


Figure 3. Daily RES data predictions considering the ARIMA model for (a) 1 step ahead, (b) 7 steps ahead, (c) 14 steps ahead, and (d) 28 steps ahead.

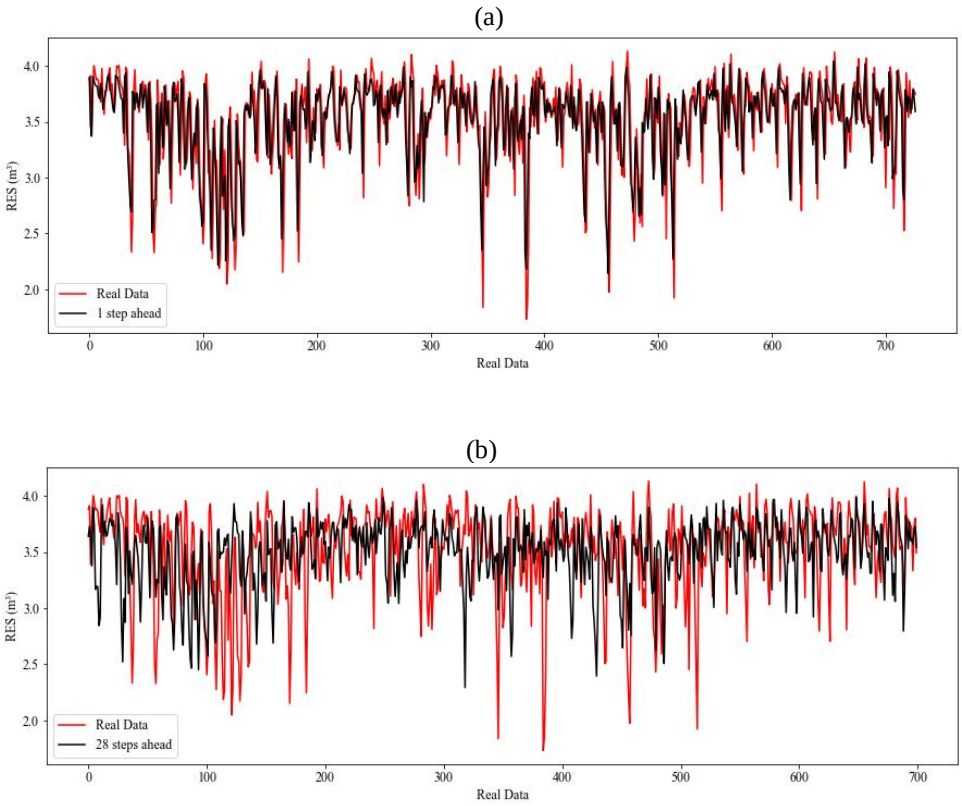
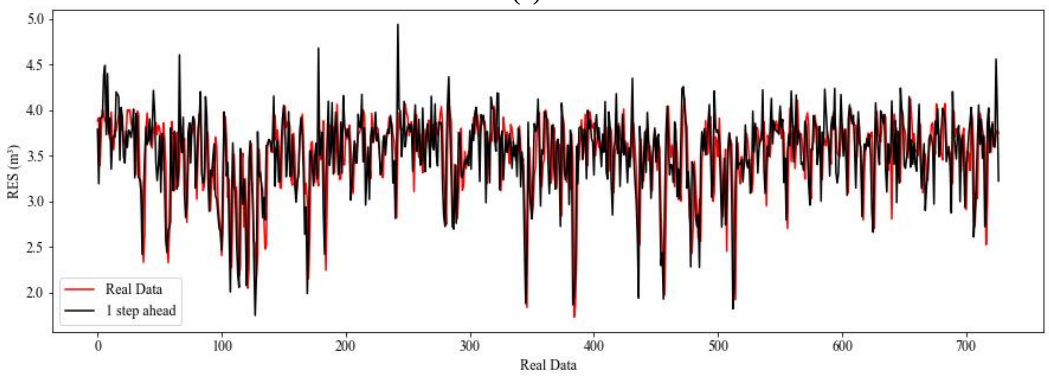


Figure 4. Daily RES data predictions considering the RF model for (a) 1 step ahead, (b) 7 steps ahead, (c) 14 steps ahead, and (d) 28 steps ahead.

(a)



(b)

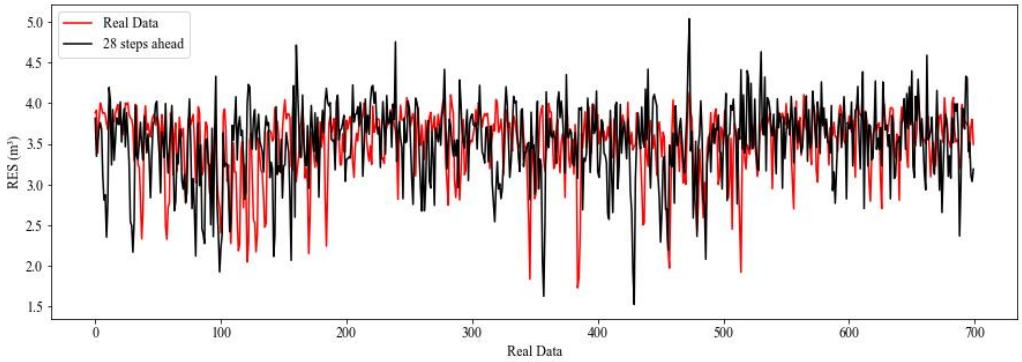


Figure 5. Daily RES data predictions considering the XGBoost model for (a) 1 step ahead, (b) 7 steps ahead, (c) 14 steps ahead, and (d) 28 steps ahead.

Analyzing figures 2 to 5, it is concluded that for 1 step forward, the Holt-Winters and ARIMA models are closer to the real time series data, where the forecast was more efficient, as seen in table 2. In figure 5 it is observed that the XGBoost model was closer to the real data, however, as the forecast horizon increases, the data predicted by the model moves away from the real data, which represents a decrease in effectiveness reflected in table 2, representing the statistical performance metrics. Table 2 describes the performance metrics RMSE, R^2, MSE, MAE and MAPE for all forecast models at different horizons.

Table 2. Result of statistical performance metrics from daily data.

		Training				Test			
Forecast	Metric	Holt-Winters	ARIMA	RF (TPE)	XGBoost (TPE)	Holt-Winters	ARIMA	RF (TPE)	XGBoost (TPE)
1	RMSE	0.379	0.370	**0.105**	0.257	0.416	0.389	0.257	**0.250**
	R^2	0.244	0.279	**0.947**	0.652	**0.697**	0.486	0.244	0.676
	MAE	0.279	0.253	**0.073**	0.190	0.311	0.298	0.196	**0.188**
	MSE	0.144	0.137	**0.010**	0.066	0.173	0.152	0.066	**0.063**
	MAPE	8.600	7.900	**2.300**	6.000	9.100	8.350	**5.600**	5.400
7	RMSE	0.389	0.379	**0.158**	0.325	0.460	0.390	**0.310**	0.365
	R^2	0.241	0.281	**0.871**	0.460	0.071	**0.488**	0.228	0.436
	MAE	0.285	0.254	**0.121**	0.247	0.345	0.295	**0.239**	0.282
	MSE	0.149	0.139	**0.030**	0.105	0,211	0.152	**0.096**	0.133
	MAPE	9.100	8.100	**3.900**	7.700	10.00	8.500	**6.700**	8.000
14	RMSE	0.395	0.383	**0.173**	0.377	0.462	0.390	0.336	**0.333**
	R^2	0.237	0.283	**0.847**	0.266	0.088	**0.489**	0.098	0.404
	MAE	0.291	0.256	**0.117**	0.277	0.346	0.299	**0.276**	0.283
	MSE	0.154	0.141	**0.025**	0.142	0.213	0.152	0.113	**0.111**
	MAPE	9.500	8.350	**3.800**	8.900	10.10	8.590	**7.700**	7.800
28	RMSE	0.399	0.394	**0.174**	0.340	0.467	0.389	**0.344**	0.411
	R^2	0.220	0.265	**0.842**	0.395	0.133	**0.485**	0.068	0.211
	MAE	0.297	0.261	**0.125**	0.259	0.350	0.302	**0.289**	0.314
	MSE	0.158	0.145	**0.030**	0.116	0.218	0.152	**0.119**	0.169
	MAPE	9.700	8.600	**4.000**	8.200	10.20	8.700	**8.000**	9.100

Regarding the training data, for 1 step forward, the RF model presented the best statistical metrics, such as RMSE, R^2, MSE and MAPE, with emphasis on the R^2 value being 0.947. Increasing the forecast horizons, to 7 steps ahead, the RF model presented the best metrics, highlighting the R^2 value of 0.847. Increasing the horizon to 14 steps ahead, RF continued to be the best time series forecast model, with a value of 0.871 for R^2. For 28 steps ahead, the RF model continued with the best performance metrics, presenting a value of 0.847 for R^2. Regarding the test data, considering 1 step forward, it can be concluded that the XGBoost model presented better performance compared to the other models, where it had the best performance metrics, highlighting the value of 0.250 for RMSE. Considering 7 steps ahead, we have the most efficient RF model, as it presents 4 of the 5 metrics with the best efficiency. Regarding the other forecast horizons, for 14 steps ahead, the RF and XGboost models presented the best results, with the two models presenting 2 best metrics each. For 28 steps ahead, the RF model was the one with the best performance metrics.

5. CONCLUSION AND FUTURE RESEARCH

This work aimed to forecast reservoir level time series (RES) in Bairro Alto, in the city of Curitiba, using classical models and machine learning models. Evaluating the proposed objectives: (i) bibliographic research was carried out on computational methods applied to water supply, mainly associated with the prediction of time series; (ii) the use of classical and machine learning models was implemented to determine the level of the water reservoir in the neighborhood and (iii) the methods were applied to the daily average and hourly data, using the performance metrics RMSE, R2, MAE, MSE and MAPE.

Other ways exist, such as using a model for each of the components, similar to Moreno et al. (2020a), or use a network with multiple inputs and multiple outputs, as used by Hu et al. (2021), in a wind speed time series forecasting application. For the daily SANEPAR data sample from 2018 and 2019, we consider different forecast horizons compared to the hourly data sample. For the training data, considering 1 step forward, the RF model demonstrated the most efficient performance metrics, with emphasis on the R^2 value being 0.947. Considering the other forecast horizons, for 7, 14 and 28 steps ahead the RF model demonstrated efficient performance metrics. For the test data, considering 1 step forward, the XGBoost model was the one with the best performance, and for the other steps forward, the RF model had the largest number of the most efficient metrics. The TPE optimization technique was responsible for improving the accuracy of the time series forecast.

Among the steps that will still be evaluated for the final work, the following can be highlighted:

- Check seasonal components using the classic SARIMA technique;
- Evaluate the decomposition method: Empirical Wavelet Transform (EWT)
- Use a network with multiple outputs;
- Evaluate different steps forward.

6. REFERENCES

Babu, C. N.; Reddy, B. E. 2014. "A moving-average filter based hybrid ARIMA–ANN model for forecasting time series data". https://doi.org/10.1016/j.asoc.2014.05.028

Billings, B., Jones, C. 2008. "Forecasting Urban Water Demand". 2 ed. Denver: American Waterworks Association. P. 340.

Billings, R., Agthe, D. 1998. "State-space versus multiple regression for forecasting urban water demand". Journal of Water Resource Planning and Management, v. 124, n. 2, p. 113-117.

Box, G. E. P., Jenkins, G. M. 1976. "Time series analysis forecasting and control". San Francisco: H. Day.

Bradford, W., Bridgeman, J., Gaterell, M. 2010. "A review of the 1892 water demand forecasts for Birmingham". *Engineering History and Heritage*, v. 164, p. 39- 49.

Breiman, L., Cutler, A. 2001. "Random Forests. Article".

Brendan, B. M.; Luvizotto JR., E. Herrera, M.; Izquierdo, J.; PerezGarcia, R. 2017. "Hybrid regression model for near real-time urban water demand forecasting". Journal of Computational and Applied Mathematics, v. 309, p. 532-541.

Cai-Xia, Shu-Yi, Bao-Jun, Wei, Wu. 2021. *Time series analysis of hemorrhagic fever with renal syndrome in mainland China by using an XGBoost forecasting model*.

Chen, J., Yang, C., Zhu, H., Li, Y, Gui, W. 2018. *A novel variable selection method based on stability and variable permutation for multivariate calibration*. Chemometrics and Intelligent Laboratory Systems, v. 182, p. 188–201

Chen, J., Wang, W., Huang, C., 1995. "Analysis of an adaptive time-series AutoRegressive moving-average (ARMA) model for short-term load forecasting". Electric Power Systems Research vol. 34, 187-196.

Chen, T.Q., Guestrin, C. 2016. *XGBoost: A Scalable Tree Boosting System*.

Costa, M., . Goldberger, A. L. 2014. "Dynamical glucometry: Use of multiscale entropy analysis in diabetes". Chaos, New York, v. 24, p. 033139.

Dhaliwal, S.; Nahid, A.; ABBAS, R. 2018. "Effective intrusion detection system using boost". Information, Multidisciplinary Digital Publishing Institute, v. 9, n. 7, p. 149.

Falkengerg, A. V. 2011. *Forecast of urban water consumption in the short term (in Portuguese).* Masters dissertation. Curitiba: Universidade Federal do Paraná.

Geman, D., Yali, A. 1997. "Shape Quantization and Recognition with Randomized Trees", 1997.

Ho, T. K. 1995. "Random decision forests". IEEE Electronic Library (IEL) Conference Proceedings. DOI: 10.1109/ICDAR.1995.598994

Hyndman, R. Athanasopoulos, G. Bergmeir, C. Caceres, G. Chhay, L. O'hara-Wild, M. Petropoulos, F. Razbash, S. Wang, E. Yasmeen, F. 2019. "Forecast: Forecasting functions for time series and linear models". R package version 8.4, 2018. Disponível em: < http://pkg.robjhyndman.com/forecast>.

Kazemi, A., Foroughi, A., Hosseinzadeh, M. 2012. "A multi-level fuzzy linear regression model for forecasting industry energy demand of Iran". Procedia-Social and Behavioral Sciences, Elsevier, v. 41, p. 342–348.

Lima, S., Gonçalvez, M., Costa, 2019. M. "Time Series forecasting using Holt-Winters Exponential Smoothing: an application to economic data".

Manfrin, D. C. 2020. *Water demand forecasting models for the city of Joinville using time series analysis (in Portuguese)*. Masters dissertation.

Martins, V., Werner, L. 2014. "Comparison of individual forecasts and their combinations: a study with industrial series (in Portuguese)". Scielo.

Morettin, P., Toloi, C. M. 2004. *Time series analysis (in Portuguese)*. São Paulo: Edgard Blucher, 2004.

Rodríguez, C. A. E. 2010. *Short-term prediction of urban water demand in densely populated areas (in Spanish).* Universidade Politecnica de Valencia

Sun, C., Chen, Z., Qin, Y., Wang, B. 2009. "Multi-step Time Series Forecasting Based on Informer-XGBoost-GA". DOI 10.1088/1742-6596/2333/1/012009

Tyralis, H., Papacharalmpous, G. 2017. "Variable Selection in Time Series Forecasting Using Random Forests". Department of Water Resources and Environmental Engineering, School of Civil Engineering, National Technical University of Athens, Iroon Polytechniou 5, 157 80 Zografou, Greece.

7. RESPONSIBILITY NOTICE

The authors are the only responsible for the printed material included in this paper.