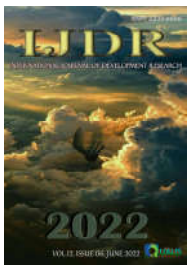
ISSN: 2230-9926

Available online at http://www.journalijdr.com

IJDR

International Journal of Development Research
Vol. 12, Issue, 06, pp. 56463-56465, June, 2022
https://doi.org/10.37118/ijdr.24619.06.2022



RESEARCH ARTICLE OPEN ACCESS

FORECASTS FOR PM10 IN MACUCO, BRAZIL

Igor Campos de Almeida Lima, Marcello Montillo Provenza, and Paulo Henrique Couto Simões

Department of Statistics, Rio de Janeiro State University, Rio de Janeiro, Brazil

ARTICLE INFO

Article History:
Received 17th March, 2022
Received in revised form
06th April, 2022
Accepted 14th May, 2022
Published online 22nd June, 2022

Key Words:
PM10, ARIMA model,
ETS model,
AR-NN model.

Corresponding author:
Igor Campos de Almeida Lima

ABSTRACT

Pollutant parameter measurements are constantly carried out at monitoring stations. The object of this work is the time series of the content of particulate matter with a diameter smaller than 10 micrometers in the municipality of Macuco/RJ. The time series has no normality, stationarity, or seasonality. The database was separated into two periods: between January 2015 and December 2019 for training and between January and December 2020 for testing. The ARIMA(1,0,0), ETS(A,N,N) and AR-NN(1,1) models were adjusted. The RMSE and MAE metrics indicate that ARIMA(1,0,0) presents the best predictions.

Citation: ***Igor Campos de Almeida Lima, Marcello Montillo Provenza, and Paulo Henrique Couto Simões.*** "Forecasts for pm10 in macuco, brazil*", International Journal of Development Research*, 12, (06), 56463-56465.

INTRODUCTION

The oldest air pollutant monitoring system is located in the state of Rio de Janeiro. The region has 95 monitoring stations, divided into 34 semi-automatic stations and 61 automatic stations. One of them is in the municipality of Macuco. Monitoring of environmental pollutants can be measured through some parameters such as particulate matter and concentrations of carbon monoxide, lead, sulfur dioxide, nitrogen dioxide and ozone, ammonia, and others. (CONAMA, 2018). Once the station performs the measurement of pollutant parameters, the data is collected and disseminated. Therefore, it may be in the interest of researchers and companies to make future projections in relation to atmospheric pollutants, since this is a problem that affects society.

The objective of this work is to compare three stochastic prediction methods for the content of particulate matter with a diameter of fewer than 10 micrometers (PM10) in the municipality of Macuco in the state of Rio de Janeiro between 2015 and 2020: Box-Jenkins, Error Trend Seasonal, and Autoregressive Neural Network models.

MATERIALS AND METHODS

The main objective of time series analysis is to forecast the data. It allows future values of a series to be predicted based on its present and past values. There can be four elements in a time series: trend, cycle, seasonality, and residual (BOX *et al*., 2015).

The histogram, Q-Q plot, and the Shapiro-Wilk (W) test are used to verify the normality of the data. The box plot provides the observation of outliers. The Augmented Dickey-Fuller (ADF), Wald-Wolfowitz (Z), and Kruskal-Wallis (T_1) tests are used to verify the presence of stationarity, trend, and seasonality, respectively (SIEGEL & CASTELLAN JR, 1975).

$$W = \frac{\left(\sum_{i=1}^{n} a_i y_{(i)}\right)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (1)$$

$$ADF = \delta + bt + \lambda Y_{t-1} + \sum_{s=1}^{p} \beta_s \Delta Y_{t-1} + \varepsilon_t \quad (2)$$

$$Z = \frac{R - \left(\frac{2n_1 n_2}{n_1+n_2} + 1\right)}{\sqrt{\frac{2n_1 n_2(2n_1 n_2 - n_1 - n_2)}{(n_1+n_2)^2(n_1+n_2-1)}}} \quad (3)$$

$$T_1 = \frac{12}{N(N+1)} * \sum_{i=1}^{k} \frac{R_j^2}{n_j} - 3(N+1) \quad (4)$$

The graphs of the Autocorrelation Function (FAC) and the Partial Autocorrelation Function (FACP) are used to admit the need to perform differentiation in the data series (BOX *et al*., 2015). The Exponential Smoothing State Space (ETS) model is a "special" class of exponential smoothing. ETS uses a three-character string recognition method: the first letter indicates the error type (A, M or Z), the second the trend type (N, A, M, or Z), and the third the type of error. seasonality (N, A, M, or Z). In these cases: N = none, A = additive, M = multiplicative and Z = automatic selection (HYNDMAN & ATHANASOPOULOS, 2018). The equation of the observations is given by:

$$Y_t = z_t\alpha_t + \varepsilon_t \quad (5)$$

And the equation of state is:

$$\alpha_t = T_t\alpha_{t-1} + R_t n_t \quad (6)$$

where t is the state vector; εt is the uncorrelated noises; nt is the serially uncorrelated noise vector; zt, Tt, and Rt are system matrices.
The Box-Jenkins methodology fits Autoregressive Integrated Moving Averages (ARIMA) models to a time series. In general, the models are parsimonious, that is, the parameters contain small values and the predictions are accurate. The objective of the methodology is to determine three components: p (autoregressive parameters), d (integration processes), and q (moving average parameters). Thus, ARIMA (p,d,q) is formed (BOX *et al*., 2015).

$$W_t = \phi_1 W_{t-1} + \cdots + \phi_p W_{t-p} + \varepsilon_t - \theta_1\varepsilon_{t-1} - \cdots - \theta_q\varepsilon_{t-q}, where\ W_t^d = \Delta^d Z_t \quad (7)$$

The Autoregressive Neural Network is a union of the autoregressive model with the neural network (AR-NN). In addition to the lagged values of the autoregressive process, neural networks usually contain a linear and a non-linear part for model adjustment (HAYKIN *et al*., 2009).

$$Z_t = \alpha_0 + \sum_{i=1}^{p}\alpha_i Z_{t-i} + \sum_{j=1}^{h}\Psi\left(\gamma_{0j} + \sum_{i=1}^{p}\gamma_{ij}Z_{t-i}\right)\beta_j + \varepsilon_t \quad (8)$$

where i indicates the number of lags, j the number of hidden layers, . is a nonlinear activation function (also called hidden neurons), ij are the parameters, 0j is the bias, j are the weights of hidden neurons, and t is the random component.

The Root Mean Error Square (RMSE) and the Mean Absolute Error (MAE) were used to assess the goodness of fit and compare the models with each other (SILVA *et al*., 2015).

$$RMSE = \sqrt{\frac{\sum(Z_t - \widehat{Z_t})^2}{n}} \quad (9)$$

$$MAE = \frac{\sum|Z_t - \widehat{Z_t}|}{n} \quad (10)$$

RESULTS

Figure 1 presents the time series of PM10 particles found in Macuco between January 2015 and December 2020. Visually, it can be assumed that the data are stationary. Table 1 reveals the statistics of the analyzed data. In all, there were 72 months between January 2015 and December 2020. The mean of the data was 25.2 ± 6.9. Figure 2 presents the histogram, Q-Q plot, and box-plot of the data. At first, there is no normality in the data, a hypothesis confirmed by the Shapiro-Wilk test (p-value = 0.0005).

Table 1. PM10 time-series statistics

Statistics	Values
Sample	72
Average	25.2
Standard deviation	6.9
Minimum	14.8
Q1	20.2
Median	24.6
Q3	29.9
Maximum	50.7

Table 2. Tests for the time series components

Statistis	p-value	Result
Dickey-Fuller	0.011	Not stationary
Wald-Wolfowitz	0.003	Have tendency
Kruskal-Wallis	0.089	Not seasonality

Table 3. Shapiro-Wilk test for residuals

Models	p-value	Result
ARIMA(1,0,0)	0.0004	Distribuição desconhecida
ETS(A,N,N)	0.3748	Distribuição normal
AR-NN(1,1)	0.0004	Distribuição desconhecida

Table 4. Forecast metrics

Models	RMSE	MAE
ARIMA(1,0,0)	7.74	7.13
ETS(A,N,N)	10.59	9.74
AR-NN(1,1)	7.99	7.32

The box plot also reveals that there are two outliers. Table 2 shows the statistical tests applied to the components of the PM10 time series. The Dickey-Fuller Augmentation reveals that the data are not stationary, confirmed by the Wald-Wolfowitz test indicating a trend. Therefore, the visual stationarity hypothesis of Figure 1 is discarded. Kruskal-Wallis points out that there is no seasonality.

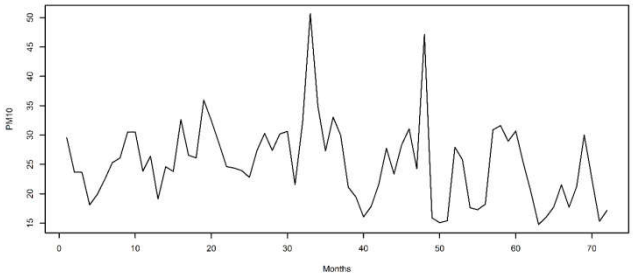


Figure 1. PM10 time series between 2015 and 2020

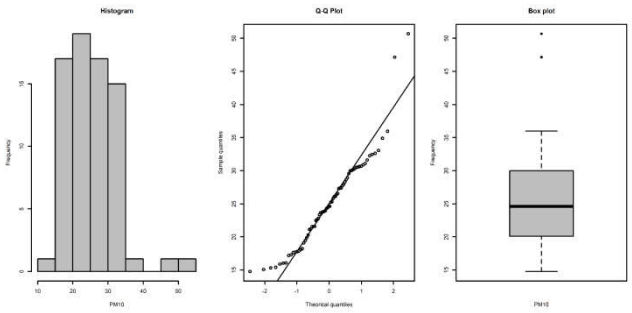


Figure 2. Histogram, Q-Q plot, and box-plot of the series

Figure 2: Histogram, Q-Q plot, and box-plot of the series. The ACF and PACF were used to verify the possibility of performing

integrations in the data series for ARIMA modeling. Through Figure 3, it is observed that the FAC has a cyclic behavior and the FACP is truncated at lag 1. Thus, the indicated model is ARIMA(1,0,0), also called AR(1).

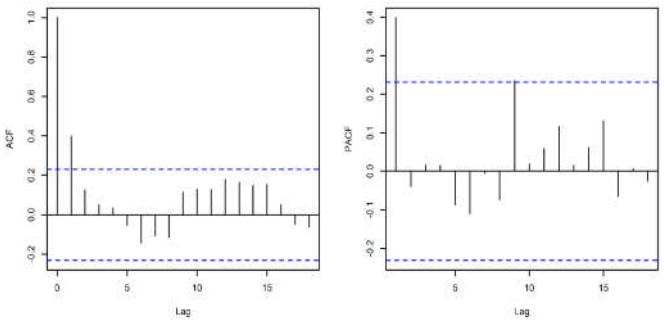


Figure 3. ACF and PACF

In addition to ARIMA (1,0,0) the ETS(A,N,N) and AR-NN(1,1) models were also adjusted. The training base corresponded to the period between January 2015 and December 2019. Residues were analyzed by their respective histograms and box-plot (Figure 3), in addition to the Shapiro-Wilk test (Table 3). Histograms do not appear to be normally distributed. The ARIMA, ETS, and AR-NN models have three, one, and two outliers respectively. Through the Shapiro-Wilk test, it is observed that the ETS model is the only one that presents residuals with a normal distribution (p-value > 0.05).

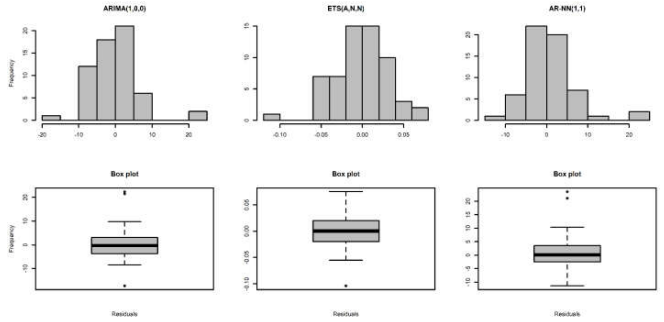


Figure 4. Histogram and box plot of residuals from the fitted models

Figure 4 presents two graphs with the models' predictions. The test base corresponded to the period between January and December 2020. Graph 1 shows the entire series between 2015 and 2020. Graph 2 reveals the forecast for the months of 2020. To evaluate the fit of the models, the forecast metrics for the test base were used. ARIMA(1,0,0) obtained the best results (RMSE = 7.74 and MAE = 7.13).

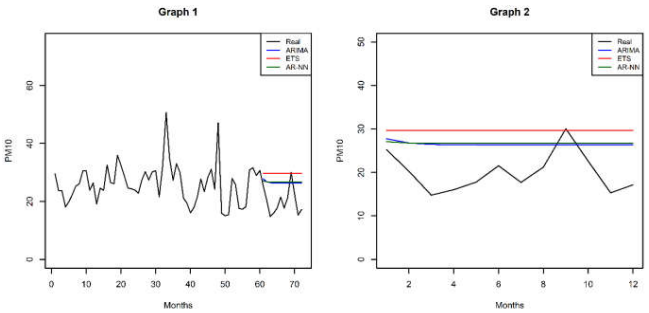


Figure 5. Forecast for 2020

Despite the normalized residuals, the ETS(A,N,N) presented the worst metrics. Through the results obtained (Table 4), it can be seen that the three models did not have good adjustments. One hypothesis analyzing the series graph (Figure 1) is that, with the COVID-19 pandemic in 2020, gas emissions were reduced (Table 2 - Wald-Wolfowitz Test) due to the industrial and economic slowdown. This may have caused the forecasts to not adjust properly. Most model estimates were above the actual values recorded in 2020 (Figure 4).

CONCLUSION

The PM10 time series does not have normality in the data. Over the 72 months (the period between 2015 and 2020) the box plot revealed that there are two outliers. The series also lacks stationarity and seasonality. Data were separated into two periods. Between January 2015 and December 2019 it was used as a training base. The ARIMA(1,0,0), ETS(A,N,N) and AR-NN(1,1) models were adjusted. Among them, only the ETS(A,N,N) obtained normality in the residuals. The test base, used to make the forecasts, corresponded to the year 2020 (between January and December). The RMSE and MAE metrics indicate that ARIMA(1,0,0) presents the best predictions, even though the model does not have residuals with normal distribution.

REFERENCES

BOX, George EP *et al*. Time series analysis: forecasting and control. John Wiley & Sons, 2015.

CONAMA. Resolution n.º 491, 11/19/2018. Provides for air quality standards in Brazil.

HAYKIN, S. *et al*. Neural networks and learning machines. vol. 3 Pearson. Upper Saddle River, NJ, EUA, v. 30, p. 39-40, 2009.

HYNDMAN, Rob J.; ATHANASOPOULOS, George. Forecasting: principles and practice. OTexts, 2018.

SIEGEL, Sidney; CASTELLAN JR, N. John. Non-parametric statistics for behavioral sciences. Artmed Editora, 1975.

SILVA, Madson Tavares *et al*. Application of the SWAT model to estimate the flow in the sub-medium São Francisco river basin. Revista Brasileira de Geografia Física, v. 8, n. 6, p. 1615-1627, 2015.

****