scientific reports

OPEN

Spatio-temporal visualization and forecasting of PM_{10} in the Brazilian state of Minas Gerais

Kim Leone Souza da Silva[1], Javier Linkolk López-Gonzales[2,3]✉, Josue E. Turpo-Chaparro[2], Esteban Tocto-Cano[3] & Paulo Canas Rodrigues[1]

Air pollution due to air contamination by gases, liquids, and solid particles in suspension, is a great environmental and public health concern nowadays. An important type of air pollution is particulate matter with a diameter of 10 microns or less (PM_{10}) because one of the determining factors that affect human health is the size of particles in the atmosphere due to the degree of permanence and penetration they have in the respiratory system. Therefore, it is extremely interesting to monitor and understand the behavior of PM_{10} concentrations so that they do not exceed the established critical levels. In this work, we will study the PM_{10} concentrations in all available monitoring stations in the Brazilian state of Minas Gerais. To better understand its behavior, we will provide a spatio-temporal visualization of the PM_{10} concentrations. Besides the descriptive and visualization analysis, we consider six standard and advanced time series models that will be used to fit and forecast PM_{10} concentrations, with application to three locations, one in Belo Horizonte, the Minas Gerais state capital, and the monitoring stations with the lowest and highest average PM_{10} concentration levels.

The human impact on the planet is remarkable, and the attempt to reduce these impacts is increasingly urgent. Air pollution, for example, is directly related to the environment and human health[1–3]. One of the determining factors that affect human health is the size of particles in the atmosphere due to the degree of permanence and penetration they have in the respiratory system[4–7]. As the health impact is directly related to the particle size, monitoring the PM_{10} concentrations, particulate materials smaller than or equal to 10 micrometers, is very important[8,9]. Based on the annual average of PM_{10}, the World Health Organization (WHO) ranked Ahvaz in Iran as the most polluted city in the world at 372 $\mu g/m^3$[10].

In this scenario, in Europe, the Apheis project has developed guidelines for analyzing and collecting data on air quality, and public health impacts[11]. The study presented the health impact in 19 Eastern and Western European cities. The results indicate that reducing long-term PM_{10} exposure by 5 $\mu g/m^3$ could prevent approximately 3300–7700 premature deaths annually. The Apheis project also showed that in urban Europe, current air pollution has a non-negligible impact on public health and that even in cities with low air pollution, preventive measures can reduce damage[12]. For its part, in Brazil, in the metropolitan region of São Paulo (MASP), 40% of PM_{10} emissions come from mobile sources[13–15]. In addition, ozone and PM_{10} are the pollutants with the greatest impact on air quality at MASP[16,17]. A study carried out in the Jânio Quadros and Maria Maluf tunnels in São Paulo indicates that the emission of heavy diesel vehicles is the major source of $PM_{2.5}$ fine particulate matter[14]. Likewise, a study in the metropolitan region of Lima, the capital of Peru, proposed a space-time visualization to analyze PM_{10} levels, showing that the highest concentrations of PM_{10} were recorded near hills and high-traffic roads and unpaved streets[18].

In particular, in Brazil, through Conama Resolution No. 005/1989, the National Air Quality Control Program (Pronar) was created[19]. This program attempts to build the foundations for a national air quality protection policy[20]. However, although Pronar is the beginning of a national air quality policy, it has great legal fragility because its legal basis is hierarchically inferior to the already established laws. Furthermore, there is a clear asymmetry between the country's regions and most of the air quality management instruments are located in southeast Brazil[21].

[1]Department of Statistics, Federal University of Bahia, Salvador, Brazil. [2]UPG Ingeniería y Arquitectura, Escuela de Posgrado, Universidad Peruana Unión, Lima, Peru. [3]Facultad de Ingeniería y Arquitectura, Universidad Peruana Unión, Lima, Peru. ✉email: javierlinkolk@gmail.com

Several studies with air pollution data were developed with this challenge, using statistical models to address both model fit[22–24], and model forecast[25–27] of air pollution in Brazil. For example, in Itabira (a city in the Brazilian state of Minas Gerais), an increase of 10 $\mu g/m^3$ of PM_{10} was associated with an increase in respiratory diseases in the emergency room, concluding that an increase in PM_{10} levels has a major impact on the exposed population[28]. Likewise, a study carried out in the Greater Vitória Region (the capital of the Brazilian state of Espírito Santo) used the seasonal auto-regressive integrated moving average with exogenous factors (SARIMAX) model to better understand and predict the behavior of PM_{10} concentrations, noting that both wind speed and rainfall were statistically significant and helped to improve the model fit[29]. On the other hand, in a study carried out in the Brazilian State of Rio Grande do Sul, it was presented that the auto-regressive integrated moving average with exogenous factors (ARMAX) model, with the inclusion of the exogenous variables (Carbon Monoxide and Sulfur Dioxide), obtained better performance when compared to the autoregressive integrated moving average (ARIMA), simple exponential smoothing, and Holt-Winters models, for PM_{10} prediction[29].

Meanwhile, other methods for time series forecasting, including neural networks[30–33] and deep learning have also been used to forecast variables related to air pollution. For example, one study proposed predictive models for $PM_{2.5}$ concentration with a model that combined the fast Fourier transform and long short-term memory neural network (FFT-LSTM) and proved to be superior to the traditional LSTM and extended long short-term memory recurrent neural network (LSTM) models[34]. Another study used a deep learning algorithm integrating convolutional neural networks (CNNs) and LSTM neural networks to predict $PM_{2.5}$ concentrations[35]. Cordova et al.[36] studied the spatio-temporal behavior of air quality in Metropolitan Lima, evaluated and predicted the PM_{10} concentrations using the recurrent artificial neural network LSTM, based on the past values of this pollutant and three meteorological variables obtained from five monitoring stations. It is important to notice that the PM_{10} concentrations have nonlinear behavior and fluctuate strongly in spatio-temporal scales[37] due to the nonlinear character of the atmospheric wind speed[38]. Consequently, to manage this strong variability, in this paper, we consider different forecasting models.

Although many studies have been made to better understand the behavior and to forecast PM_{10} concentrations, no comprehensive study that includes all monitoring stations in the Brazilian state of Minas Gerais has been made. In this paper, we will analyze the spatio-temporal dynamics of PM_{10} concentrations in all available monitoring stations in the Brazilian state of Minas Gerais. Then, we compare classical parametric models and neural networks to forecast the PM_{10} concentrations, whose results can be useful for governmental agencies and policymakers to decide on specific policies and actions to improve air quality.

The rest of the paper is structured as follows. The following section describes the data collection, data cleaning, and the methods and models used for PM_{10} forecasting. The section "Results and discussion" presents the descriptive analysis and the main findings of this research regarding model fit and model forecasting. Finally, the section "Conclusions" provides the main conclusions of this paper, together with some recommendations for future research.

Materials and methods

Data collection and data cleaning. The data used in this work was collected by the State Foundation for the Environment of the Brazilian state of Minas Gerais. The data was collected hourly and the last five years available were considered (between 2015 and 2019) in all 58 monitoring stations. The data are publicly available per municipality, monitoring station, and year. For each combination municipality/monitoring station/year, the data is available in a csv file that includes the hourly information on pollutant levels such as PM_{10} and $PM_{2.5}$, as well as meteorological data such as temperature, wind direction, rainfall, atmospheric pressure, wind speed, radiation, and relative humidity. The first step of the analysis was to organize, clean, and store the database, which is always a challenging operation when dealing with real data. In this case, the main challenges were:

- The lack of information for some variables in several stations.
- A significant amount of missing values.

From the available 58 monitoring stations, we decided to discard those with a percentage of missing values above 35% in the PM_{10} data. In addition, one station that did not have data for 2015 was also discarded. Thus, we proceeded with data from 29 air quality monitoring stations distributed throughout the Brazilian state of Minas Gerais. The locations of the 29 monitoring stations considered in this study can be seen in Fig. 1, with more stations in areas with higher population density, resulting in some overlapped points in the map. Figure 6S of the "Supplementary material" shows a heat map of the missing PM_{10} values in each monitoring station, and Table 1S gives the rate of the missing values. The next stage was the imputation of missing values, which was done by using the function `na_kalman` of the package `imputeTS` in the `R` software[39]. At the end of the process, we obtain a database of PM_{10} concentrations with 43824 hourly observations (rows) for each of the 29 stations (columns) available in the "Supplementary material". Table 1S of the "Supplementary material", presents detailed information for each monitoring station, including code, station name, company responsible for the monitoring station, longitude, latitude, and the rate of missing values.

Models for time series forecasting. Time series models are very important and can be useful in many areas of knowledge that collect time-dependent data[40,41]. They can be used both to understand the underline process that generated the data and to predict future observations[42,43]. Predictions can be for a short term (e.g., 1 h ahead) and for a long term (e.g., 720 h–1 month ahead). Despite the forecasting horizon, forecasting is an important aid to effective planning, and policy-making[44]. In this study, six models for time series forecasting of PM_{10} levels are considered and briefly described in the sequence.

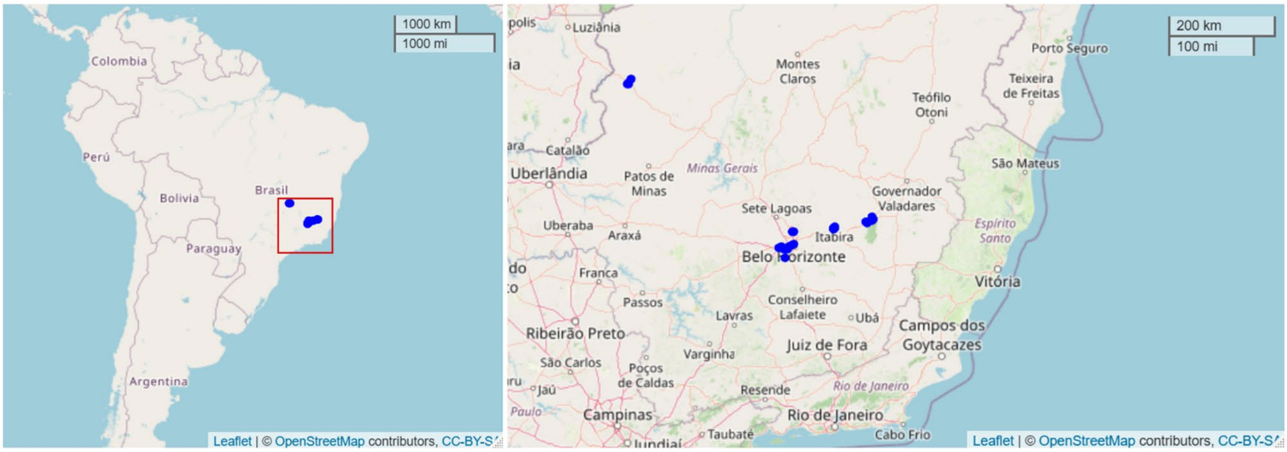

Figure 1. Map of South America (left) and map with the location of the monitoring stations (right). The source map was made with the R package leaflet, version 2.1.1.

Seasonal Naive. The Seasonal Naive (SNAIVE) model is an extension of the NAIVE model that considers a seasonal component of period T in the time series[45] and can be written as

$$\widehat{Y}(t+h|t) = Y(t+h-T), \tag{1}$$

where t is the length of the time series, h is the forecasting horizon, T is the seasonal period, $\widehat{Y}(t+h|t)$ is the prediction h steps ahead, and $Y(t+h-T)$ is the observed value T observations before the length of the series, t, minus the forecasting horizon, h. This means that the seasonal naive model estimates the out-of-sample forecast as the last observation at the same seasonal point. When considering $T=1$, the NAIVE model is obtained. This model was adjusted using the `snaive` function of the package `forecast` in the software `R`.

Seasonal Naive + Decomposition. Let us consider the three-part decomposition of the time series $Y(t)$ of length t,

$$Y(t) = T(t) + S(t) + R(t), \tag{2}$$

where $T(t)$ is the trend of the time series, $S(t)$ is the seasonal component, and $R(t)$ is the rest/residual of the time series. Although several techniques are available to estimate the components in the decomposition, we consider the STL (Seasonal and Trend decomposition using Loess) for its versatility and robustness. The model Seasonal Naive + decomposition firstly removes the seasonality $S(t)$ of the time series $Y(Y)$,

$$\widehat{Y}(t) = Y(t) - S(t), \tag{3}$$

and then uses the NAIVE model to forecast the time series with the seasonal adjustment, which is added to the seasonal adjustment of the last time period of the time series to obtain the final forecast. The decomposition and forecasts can be obtained by using the `stl` and `naive` functions of the `R` software.

Exponential Smoothing + Decomposition. Exponential smoothing is one of the most used and well-known methods for time series forecasting[46]. The forecast h steps ahead for the simple exponential smoothing can be written as:

$$\widehat{Y}(t+h|t) = \alpha y(t) + \alpha(1-\alpha)y(t-1) + \alpha(1-\alpha)^2 y(t-2) + \cdots, \tag{4}$$

with $\alpha \in [0,1]$. In this way, the forecasts are obtained as a weighted average of past observations, with the weights decreasing exponentially as we go back in time. Various versions of exponential smoothing have been proposed to deal with trends and seasonality in time series. In this work, we use the exponential smoothing model automatically selected for the seasonally adjusted series. Further details about exponential smoothing algorithms can be found in[46].

SARIMA. The seasonal autoregressive integrated moving average (SARIMA) models are among the most widely used methods for time series forecasting. They are an extension of the autoregressive integrated moving average (ARIMA) model that adds a linear combination of seasonal values and/or forecast errors. Let $Y(t)$ be a time series. The $SARIMA(p,d,q)(P,D,Q)_s$ model can be written as

$$(1-B)^d(1-B^s)^D\Phi(B^s)\phi(B)Y(t) = \Theta(B^s)\theta(B)\varepsilon(t) \tag{5}$$

where B is the lag operator given by $B^k = Y(t-k)/Y(t)$, $\Phi(B) = 1 - \phi_1 B^1 - \phi_2 B^2 z \cdots - \phi_p B^p$ is an autoregressive (AR) polynomial function of order p with vector of coefficients $\Phi' = [\phi_1, \phi_2, \ldots, \phi_p]$, $\Theta(B) = 1 + \theta_1 B^1 + \theta_2 B^2 + \cdots + \theta_q B^q$ is a moving average (MA) polynomial of order q with vector of coefficients $\Theta' = [\theta_1, \theta_2, \ldots, \theta_q]$, $\Phi(B^s) = 1 - \phi_{s,1}B^s - \phi_{s,2}B^{2s} - \cdots - \phi_{s,p}B^{ps}$ and $\Theta(B^s) = 1 + \theta_{s,1}B^s - \theta_{s,2}B^{2s} - \cdots - \theta_{s,q}B^{qs}$ are seasonal polynomial functions of order P and Q, respectively, that satisfy the stationarity and invertibility conditions, d is the number of differences needed to stationarize the series, D is the number of seasonal differences and $\varepsilon(t)$ is white noise, defined as a sequence of uncorrelated random variables with zero mean and constant variance over time, $\varepsilon_t \sim RB(0, \sigma_\varepsilon^2)$. The parameter estimates of the SARIMA model can be obtained with the `arima` function of the `R` software.

NNETAR and NNETAR + Decomposition. The Neural Network AutoRegression (NNETAR) model is an artificial neural network (ANN). ANNs are mathematical models based on the behavior of the brain that allow for complex nonlinear relationships between the response variable and its predictors[44]. A neural network comprises an input, output, and hidden layers. In the hidden layers, we find the weights (W_i), bias (b), and the activation function, which help to convert the input data into the expected output. The weights are the parameters that will determine the intensity with which each neuron affects the other. On the other hand, bias is a parameter used to adjust the output along with the weighted sum of the neuron's inputs. In each neuron, there will be an activation process through the z function[47]. This process is illustrated by Eq. (6):

$$z = \sum_{i=1}^{k} W_i X + b \tag{6}$$

The forecasts using the NNETAR model and the NNETAR in the seasonally adjusted time series using the STL decomposition can be obtained with the `nnetar` function of the `forecast` package in the `R` software. The model receives the last observations up to time t and performs the forecast for time $t + 1$. To obtain more predictions, the same process is repeated iteratively.

Accuracy measures. To evaluate the performance of the models, two types of accuracy measures will be considered, one for the model fit (using the training data) and another for the model forecast (using the train set). Two accuracy measures will be used. Equation (7) defines the root mean squared error (RMSE) and Eq. (8) defines the symmetric mean absolute percent error (SMAPE). In contrast to the mean absolute percentage error, the SMAPE provides a value with upper and lower bounds, with values between zero and one.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \widehat{y}_i)^2} \tag{7}$$

$$\text{SMAPE} = \frac{100\%}{n}\sum_{i=1}^{n}\frac{|y_i - \widehat{y}_i|}{|\widehat{y}_i| + |y_i|} \tag{8}$$

In both equations, n is the number of observations (i.e., length of the train or test data), y_i, $i = 1, \ldots, n$ are the observed real values, and $\widehat{y}_i$ are the estimated or forecast values.

Results and discussion

Descriptive analysis. The database includes 43824 hourly observations (5 y between 2015 and 2019) of PM_{10} concentrations in 29 monitoring stations in the Brazilian state of Minas Gerais. Being a large dataset results in a big challenge for its visualization. To better visualize and understand the behavior and patterns of the data, several strategies were used. The weekly average in each monitoring station is presented in Fig. 1S of the "Supplementary material". In addition, boxplots per hour of the day, per day of the month, per month of the year, and per year are also presented in Figs. 2S–5S of the "Supplementary material", respectively. In these plots, specific trends and patterns are visible, particularly, along the day, along the months, and along the years.

To present further results, without doing an exhaustive analysis, three monitoring stations were selected. The first is located in Belo Horizonte (BH1), the state capital city and the most populous city in the state, with its main sources of atmospheric pollution being traffic and industry. To consider the full range of the observed data in the 29 monitoring stations, the other two monitoring stations that were selected are those with the lowest (Itabira4) and the highest (S.J.daLapa2) average concentration of PM_{10} among the available stations. Figure 2 shows the weekly average behavior of these three monitoring stations. It is possible to notice that the concentrations of BH1 and Itabira4 are very similar, with emphasis on the year 2019, where the BH1 station shows a significant increase in the average weekly concentration of PM_{10}. Among the 29 considered monitoring stations, BH1 and Itabira4 are among those with the lowest average pollution levels. São José da Lapa (S.J.daLapa2), located north of the metropolitan region of Belo Horizonte, has PM_{10} concentrations well above the weekly average of the other two stations, which is likely due to lime and crushed stone factories located in the region. The average concentration of PM_{10} in S.J.daLapa2 is 49.9 $\mu g/m^3$ against 25.37 $\mu g/m^3$ and 22.13 $\mu g/m^3$ in BH1 and Itabira4, respectively.

Figure 3 shows the behavior of the hourly, daily, monthly, and annual concentration of PM_{10} at the BH1 station. The hourly plot shows a higher concentration between 7 and 10 a.m. and between 6 and 10 p.m. In the monthly plot, higher concentrations of PM_{10} are observed between June and October. There is also an increase in concentrations in the years 2018 and 2019. Figure 4 shows the behavior of the hourly, daily, monthly, and annual concentration of PM_{10} at the Itabira4 station. The hourly graph shows a higher concentration between 6 and 9 a.m. and at the end of the day between 6 and 11 p.m. In the monthly plot, higher PM_{10} concentrations are observed between June and October. Figure 5 shows the behavior of the hourly, daily, monthly, and annual

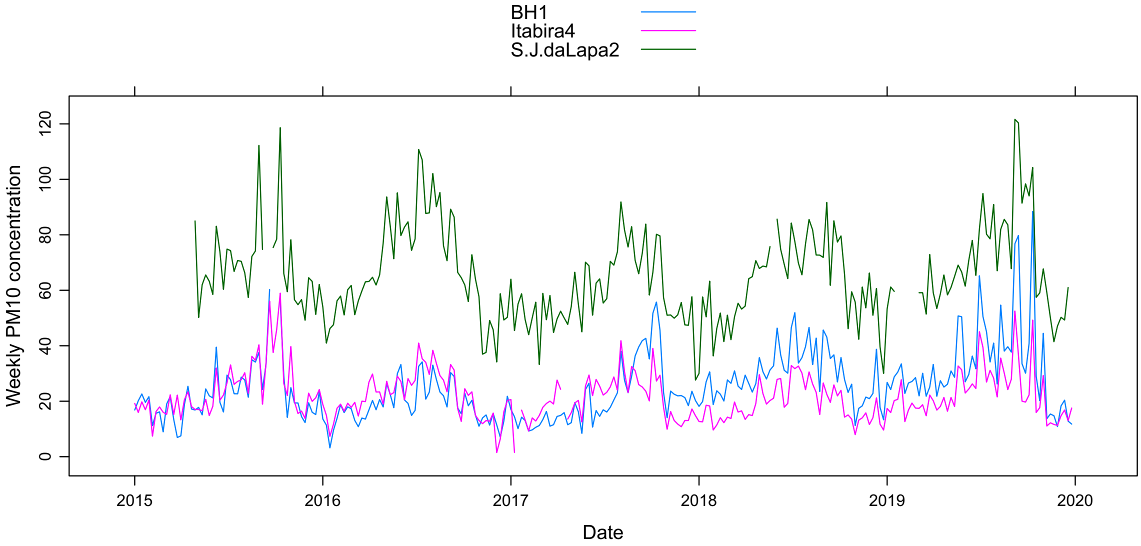

Figure 2. Average weekly concentration of PM_{10} (in $\mu g/m^3$) between 2015 and 2019 for one monitoring station located in the Brazilian state capital of Minas Gerais, Belo Horizonte (BH1, blue), the monitoring station with the lowest average of PM_{10} concentrations (Itabira4, pink), and the monitoring station with the highest average of PM_{10} concentrations (S.J.daLapa2, green).

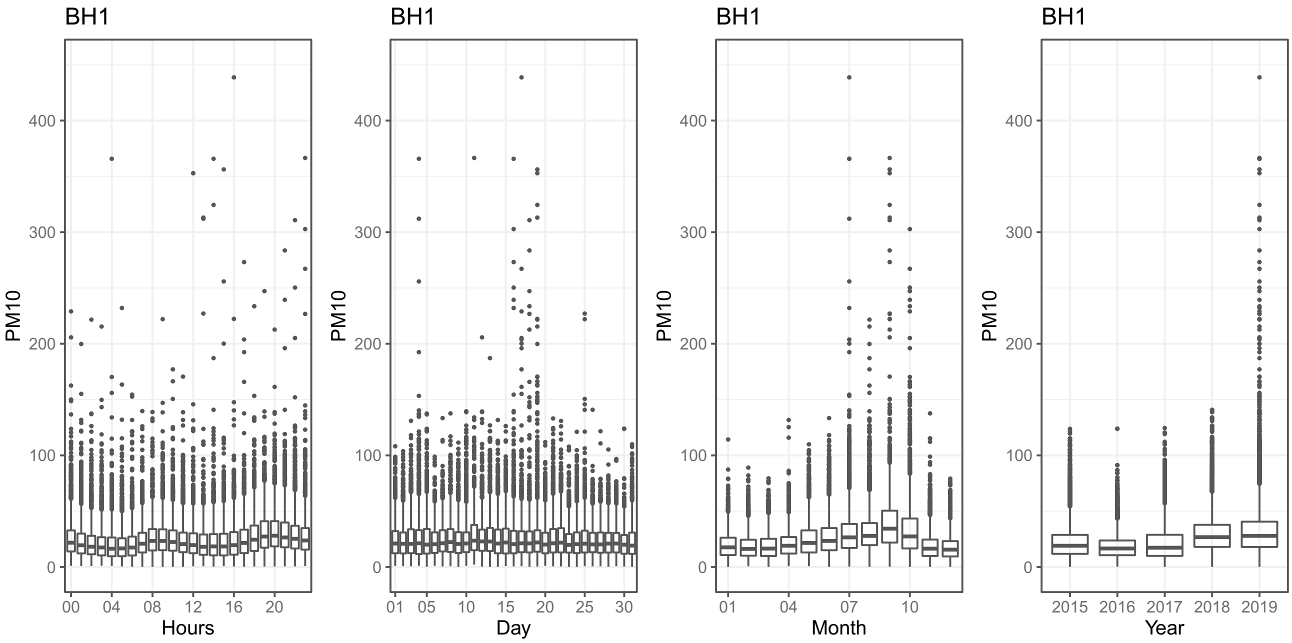

Figure 3. Hourly, daily, monthly and yearly boxplots for the monitoring station located in the Brazilian state capital of Minas Gerais, Belo Horizonte (BH1), respectively.

concentration of PM_{10} at the S.J.daLapa2 station. The hourly graph shows a higher concentration between 6 and 9 a.m. and between 5 and 11 p.m. In the monthly plot, higher concentrations of PM_{10} are also observed between June and October.

All boxplots for the hourly, daily, monthly, and annual behavior of the 29 monitoring stations can be seen in Figs. 2S–5S, of the "Supplementary material", respectively.

Model fit. The six models defined above were used for model fit, considering the data from the three monitoring stations described in the previous subsection (BH1, Itabira 4, and S.J.daLapa2). Table 1 shows the results of the two accuracy measures, RMSE and SMAPE, for the model fit of each model, in the data from the three monitoring stations. Based on the RMSE, the best fit was obtained by the model NNETAR for the Itabira4 and S.J.daLapa2 monitoring stations, while the best model for BH1 was the NNETAR+Decomposition. When con-

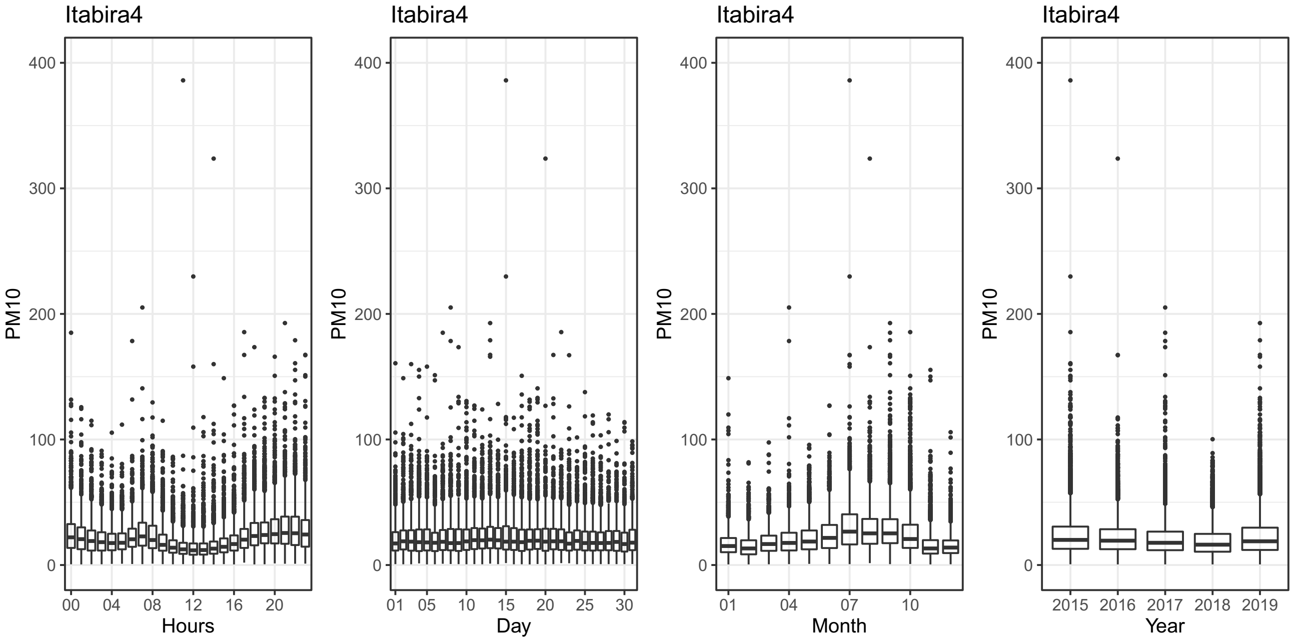

Figure 4. Hourly, daily, monthly and yearly boxplots for the monitoring station with the lowest average of PM_{10} concentration (Itabira4), respectively.

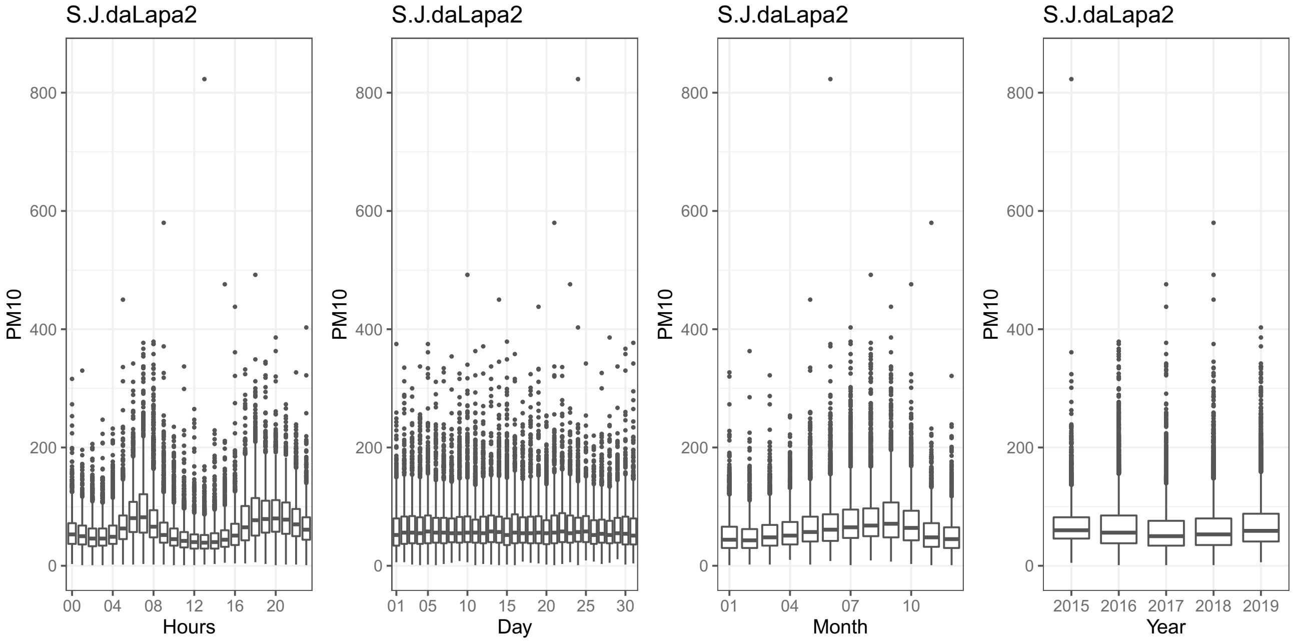

Figure 5. Hourly, daily, monthly and yearly boxplots for the monitoring station with the highest average of PM_{10} concentration (S.JdaLapa2), respectively.

sidering the SMAPE, the results for Tabira4 and S.J.daLapa2 do not change, but for BH1, the best model was the Naive+Decomposition.

Model forecasting. A similar procedure for model fit, now considering the test data, was done for the model forecast. The same six models were used, considering the data from the three monitoring stations. Table 2 shows the results of the two accuracy measures, RMSE and SMAPE, for the forecasts using each of the six models for the data from the three monitoring stations, BH1, Itabira4, and S.J.daLapa2. The accuracy measures were obtained by considering the last 14 days (336 observations) of each time series as test data. From the analysis of Table 2, it can be seen that the best model to forecast the PM_{10} concentrations in BH1 is SARIMA. For the monitoring station with the highest PM_{10} average, S.J.daLapa2, the Exponential Smoothing + Decomposition

	BH1		Itabira4		S.J.daLapa2	
	RMSE	SMAPE	RMSE	SMAPE	RMSE	SMAPE
Seasonal Naive	18.53	0.47	14.73	0.43	39.58	0.38
Naive + Dec	8.94	0.23	8.78	0.25	29.92	0.27
Exp Smoothing + Dec	8.75	0.27	8.61	0.25	28.97	0.27
SARIMA	8.66	0.26	8.36	0.26	27.14	0.28
NNETAR	8.49	0.27	8.15	0.24	26.16	0.26
NNETAR + Dec	8.46	0.27	8.21	0.24	26.78	0.26

Table 1. Accuracy measures (RMSE and SMAPE) for the model fit of each of the five considered models, in the data from the three monitoring stations, BH1, Itabira4, and S.J.daLapa2.

	BH1		Itabira4		S.J.daLapa2	
	RMSE	SMAPE	RMSE	SMAPE	RMSE	SMAPE
Seasonal Naive	11.24	0.58	9.85	0.57	34.31	0.41
Naive + Dec	8.74	0.35	7.11	0.32	30.10	0.43
Exp Smoothing + Dec	9.06	0.36	7.11	0.30	26.45	0.28
SARIMA	8.23	0.35	7.02	0.31	31.05	0.35
NNETAR	10.41	0.39	11.52	0.52	47.00	0.65
NNETAR + Dec	8.64	0.34	9.39	0.47	33.34	0.50

Table 2. Accuracy measures (RMSE and SMAPE) for the model forecast of each of the five considered models, in the data from the three monitoring stations, BH1, Itabira4, and S.J.daLapa2.

was the best forecasting model. On the other hand, for Itabira4, the best forecasting model was the Exponential Smoothing + Decomposition based on the SMAPE and the SARIMA based on the RMSE.

Conclusion

The approach presented in this paper provided a spatio-temporal and descriptive analysis of the behavior of the PM_{10} concentrations in 29 monitoring stations in the Brazilian state of Minas Gerais. The use of boxplots per hour of the day, per day of the month, per month of the year, and per year, allowed us to find specific trends and patterns. Besides the seasonal patterns, an increase in the PM_{10} concentrations was visible in BH1 from 2018 and especially at the end of 2019. S.J.daLapa2 is the monitoring station with the highest average concentration of PM_{10}, likely due to lime and crushed stone factories located in the region, with an average concentration of 49.9 $\mu g/m^3$ against 25.37 $\mu g/m^3$ and 22.13 $\mu g/m^3$ in BH1 and Itabira4, respectively.

For the modeling and forecast part of the paper, six standard and more advanced models for time series were considered, as well as three monitoring stations: BH1, the capital city of the Brazilian state of Minas Gerais, and the monitoring stations with the lowest and highest average PM_{10} concentration levels. The overall best models for model fit were the NNETAR and NNETAR+decomposition, and the overall best models for forecasting were the SARIMA and Exponential Smoothing + decomposition. This difference could be because of the small difference in RMSE and SMAPE between several models in the model fit.

Although the methodologies used in this study have been widely used for time series forecasting in general and to forecast PM_{10} concentrations in particular, no comprehensive study including all monitoring stations in the Brazilian state of Minas Gerais has been made. Therefore the results and analyses presented in this paper, both in terms of model fit to better understand the historical behavior and of model forecast to predict the coming hours and days are of great potential relevance for local governments and policymakers to understand the dynamics of the PM_{10} concentrations and take the necessary action to improve the environment and public health.

Some of the limitations of this study that can be considered as future working directions are: (1) the forecasting models discussed in this paper might not fully capture the whole signal in the data and others, e.g., based on deep learning[48,49] and hybrid methods[50,51], can be considered for all 29 monitoring stations in the Brazilian state of Minas Gerais to better understand the overall behavior; (2) the modeling and forecasting are based on univariate time series models and without geographical information, that can potentially be improved when considering multivariate and station-temporal models[52]; and (3) the influence of climate variables such as temperature, wind speed, radiation, and humidity, is not accessed in this paper, but their use might help to improve the forecasts and the spatio-temporal modeling approach as covariates.

Data availability

The data is available as supplementary material for this paper.

Received: 7 October 2022; Accepted: 21 February 2023
Published online: 25 February 2023

References

1. Martins, L. C. *et al.* Poluição atmosférica e atendimentos por pneumonia e gripe em São Paulo, Brasil. *Revista de Saúde Pública* **36**, 88–94 (2002).
2. Goudarzi, G. *et al.* Health risk assessment on human exposed to heavy metals in the ambient air PM_{10} in Ahvaz, Southwest Iran. *Int. J. Biometeorol.* **62**, 1075–1083 (2018).
3. Makri, A. & Stilianakis, N. I. Vulnerability to air pollution health effects. *Int. J. Hygiene Environ. Health* **211**, 326–336 (2008).
4. Idani, E. *et al.* Characteristics, sources, and health risks of atmospheric PM_{10}-bound heavy metals in a populated Middle Eastern City. *Toxin Rev.* **39**, 266–274 (2020).
5. Wang, J., Hu, Z., Chen, Y., Chen, Z. & Xu, S. Contamination characteristics and possible sources of PM_{10} and $PM_{2.5}$ in different functional areas of Shanghai, China. *Atmos. Environ.* **68**, 221–229 (2013).
6. Guarnieri, M. & Balmes, J. R. Outdoor air pollution and asthma. *Lancet* **383**, 1581–1592 (2014).
7. Anderson, J. O., Thundiyil, J. G. & Stolbach, A. Clearing the air: A review of the effects of particulate matter air pollution on human health. *J. Med. Toxicol.* **8**, 166–175 (2012).
8. Roy, D., Seo, Y.-C., Kim, S. & Oh, J. Human health risks assessment for airborne PM_{10}-bound metals in Seoul, Korea. *Environ. Sci. Pollut. Res.* **26**, 24247–24261 (2019).
9. Maesano, C. *et al.* Impacts on human mortality due to reductions in PM_{10} concentrations through different traffic scenarios in Paris, France. *Sci. The Total. Environ.* **698**, 134257 (2020).
10. Maleki, H., Sorooshian, A., Goudarzi, G., Nikfal, A. & Baneshi, M. M. Temporal profile of PM_{10} and associated health effects in one of the most polluted cities of the world (Ahvaz, Iran) between 2009 and 2014. *Aeolian Res.* **22**, 135–140 (2016).
11. Medina, S., Le Tertre, A. & Saklad, M. The Apheis project: Air pollution and health—A European information system. *Air Qual. Atmos. Heal.* **2**, 185–198 (2009).
12. Medina, S., Plasencia, A., Ballester, F., Mücke, H. & Schwartz, J. Apheis: Public health impact of PM_{10} in 19 European cities. *J. Epidemiol. Community Heal.* **58**, 831–836 (2004).
13. Pérez-Martínez, P. J., de Fátima Andrade, M. & de Miranda, R. M. Traffic-related air quality trends in São Paulo, Brazil. *J. Geophys. Res. Atmos.* **120**, 6290–6304 (2015).
14. Sánchez-Ccoyllo, O. R. *et al.* Vehicular particulate matter emissions in road tunnels in Sao Paulo, Brazil. *Environ. Monitoring Assess.* **149**, 241–249 (2009).
15. Ribeiro, H. & de Assunção, J. V. Historical overview of air pollution in São Paulo Metropolitan Area, Brazil: Influence of mobile sources and related health effects. *WIT Trans. Built Environ.* **52**,10 (2001).
16. Bravo, M. A. & Bell, M. L. Spatial heterogeneity of PM_{10} and O_3 in São Paulo, Brazil, and implications for human health studies. *J. Air Waste Manag. Assoc.* **61**, 69–77 (2011).
17. De Freitas, E. D., Martins, L. D., da Silva Dias, P. L. & de Fátima Andrade, M. A simple photochemical module implemented in rams for tropospheric ozone concentration forecast in the metropolitan area of Sao Paulo, Brazil: Coupling and validation. *Atmos. Environ.* **39**, 6352–6361 (2005).
18. Encalada-Malca, A. A., Cochachi-Bustamante, J. D., Rodrigues, P. C., Salas, R. & López-Gonzales, J. L. A spatio-temporal visualization approach of PM_{10} concentration data in Metropolitan Lima. *Atmosphere* **12**, 609 (2021).
19. do Meio Ambiente, C. N. Institutes the national air quality control programee. Tech. Rep., Official Journal of the Federative Republic of Brazil (1989).
20. do Meio Ambiente, C. N. Sets standards of primary and secondary air quality and even the criteria for acute episodes of air pollution. Tech. Rep., Official Journal of the Federative Republic of Brazil (1990).
21. Artaxo, P. O estado da qualidade do ar no brasil. *Work. Pap. WRI Brasil* 32 (2021).
22. Costa, A. F., Hoek, G., Brunekreef, B. & Ponce de Leon, A. C. Air pollution and deaths among elderly residents of Sao Paulo, Brazil: An analysis of mortality displacement. *Environ. Health Perspectives* **125**, 349–354 (2017).
23. Bravo, M. A., Son, J., De Freitas, C. U., Gouveia, N. & Bell, M. L. Air pollution and mortality in São Paulo, Brazil: Effects of multiple pollutants and analysis of susceptible populations. *J. Exposure Sci. Environ. Epidemiol.* **26**, 150–161 (2016).
24. Chiarelli, P. S. *et al.* The association between air pollution and blood pressure in traffic controllers in Santo André, São Paulo, Brazil. *Environ. Res.* **111**, 650–655 (2011).
25. Ventura, L. M. B., de Oliveira Pinto, F., Soares, L. M., Luna, A. S. & Gioda, A. Forecast of daily $PM_{2.5}$ concentrations applying artificial neural networks and holt-winters models. *Air Qual. Atmos. Heal.* **12**, 317–325 (2019).
26. Leão, M. L. P., Zhang, L. & da Silva Júnior, F. M. R. Effect of particulate matter ($PM_{2.5}$ and PM_{10}) on health indicators: Climate change scenarios in a Brazilian Metropolis. *Environ. Geochem. Heal.* **44***,* 1–12 (2022).
27. Habermann, M. & Gouveia, N. Application of land use regression to predict the concentration of inhalable particular matter in São Paulo City, Brazil. *Engenharia Sanit. e Ambiental* **17**, 155–162 (2012).
28. Braga, A. L. F., Pereira, L. A. A., Procópio, M., André, P. A. D. & Saldiva, P. H. D. N. Association between air pollution and respiratory and cardiovascular diseases in Itabira, Minas Gerais State. *Brazil. Cadernos de Saúde Pública* **23**, S570–S578 (2007).
29. Pinto, W. D. P., Reisen, V. A. & Monte, E. Z. Previsão da concentração de material particulado inalável, na região da grande vitória, ES, Brasil, utilizando o modelo sarimax. *Engenharia Sanitária e Ambiental* **23**, 307–318 (2018).
30. Schornobay-Lui, E. *et al.* Prediction of short and medium term PM_{10} concentration using artificial neural networks. *Manag. Environ. Qual. An Int. J.* **30**, 414–436 (2018).
31. Neto, P. S. D. M. *et al.* Neural-based ensembles for particulate matter forecasting. *IEEE Access* **9**, 14470–14490 (2021).
32. Albuquerque Filho, F. S. D., Madeiro, F., Fernandes, S. M., de Mattos Neto, P. S. & Ferreira, T. A. Time-series forecasting of pollutant concentration levels using particle swarm optimization and artificial neural networks. *Química Nova* **36**, 783–789 (2013).
33. Lei, T. M., Siu, S. W., Monjardino, J., Mendes, L. & Ferreira, F. Using machine learning methods to forecast air quality: A case study in Macao. *Atmosphere* **13**, 1412 (2022).
34. Yu, T. *et al.* Study on the regional prediction model of $PM_{2.5}$ concentrations based on multi-source observations. *Atmos. Pollut. Res.* **13**, 101363 (2022).
35. Li, J., Xu, G. & Cheng, X. Combining spatial pyramid pooling and long short-term memory network to predict $PM_{2.5}$ concentration. *Atmos. Pollut. Res.* **13**, 101309 (2022).
36. Cordova, C. H. *et al.* Air quality assessment and pollution forecasting using artificial neural networks in Metropolitan Lima-Peru. *Sci. Rep.* **11**, 1–19 (2021).
37. Plocoste, T., Calif, R. & Jacoby-Koaly, S. Temporal multiscaling characteristics of particulate matter PM_{10} and ground-level ozone O_3 concentrations in caribbean region. *Atmos. Environ.* **169**, 22–35 (2017).
38. Calif, R. & Schmitt, F. G. Multiscaling and joint multiscaling description of the atmospheric wind speed and the aggregate power output from a wind farm. *Nonlinear Process. Geophys.* **21**, 379–392 (2014).
39. Hyndman, R. J. & Khandakar, Y. Automatic time series forecasting: The forecast package for r. *J. Stat. Softw.* **27**, 1–22 (2008).
40. Harvey, A. C. *Forecasting, structural time series models and the Kalman filter* (Cambridge University Press, 1990).
41. Zhang, G. P. Time series forecasting using a hybrid arima and neural network model. *Neurocomputing* **50**, 159–175 (2003).

42. Liao, T. W. Clustering of time series data—A survey. *Pattern Recognit.* **38**, 1857–1874 (2005).
43. Bell, M. L., Samet, J. M. & Dominici, F. Time-series studies of particulate matter. *Annu. Rev. Public Heal.* **25**, 247–280 (2004).
44. Hyndman, R. J. & Athanasopoulos, G. *Forecasting: Principles and Practice* (OTexts, 2018).
45. Box, G. E., Hillmer, S. C. & Tiao, G. C. Analysis and modeling of seasonal time series. in *Seasonal analysis of economic time series*, 309–344 (NBER, 1978).
46. Sulandari, W., Suhartono, Subanar & Rodrigues, P. C. Exponential smoothing on modeling and forecasting multiple seasonal time series: An overview. *Fluctuation Noise Lett.* **20**, 2130003 (2021).
47. Rodrigues, P. C., Awe, O. O., Pimentel, J. S. & Mahmoudvand, R. Modelling the behaviour of currency exchange rates with singular spectrum analysis and artificial neural networks. *Stats* **3**, 137–157 (2020).
48. Sako, K., Mpinda, B. N. & Rodrigues, P. C. Neural networks for financial time series forecasting. *Entropy* **24**, 657 (2022).
49. Coelho, Leite *et al.* Statistical and artificial neural networks models for electricity consumption forecasting in the Brazilian industrial sector. *Energies* **15**, 588 (2022).
50. Sulandari, W., Subanar, S., Lee, M. H. & Rodrigues, P. C. Time series forecasting using singular spectrum analysis, fuzzy systems and neural networks. *MethodsX* **7**, 101015 (2020).
51. Sulandari, W. *et al.* Indonesian electricity load forecasting using singular spectrum analysis, fuzzy systems and neural networks. *Energy* **190**, 116408 (2020).
52. Rodrigues, P. C. & Mahmoudvand, R. The benefits of multivariate singular spectrum analysis over the univariate version. *J. Frankl. Inst.* **355**, 544–564 (2018).

Acknowledgements

P.C. Rodrigues acknowledges financial support from the CNPq grant "bolsa de produtividade PQ-2" 309359/2022-8, Federal University of Bahia, and CAPES-PRINT-UFBA, under the topic "Modelos Matemáticos, Estatísticos e Computacionais Aplicados às Ciências da Natureza".

Author contributions

All authors participated in the conceptualization, methodology, software, and manuscript writing.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-30365-w.

Correspondence and requests for materials should be addressed to J.L.L.-G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.