Box-Jenkins stochastic models for studying air pollutants in a Latin American megacity

A D Pinzón-Hassan¹, V Tique-Ortiz¹, and C A Zafra-Mejía¹
¹Grupo de Investigación en Ingeniería Ambiental, Universidad Distrital Francisco José de Caldas, Bogotá, Colombia

E-mail: vtiqeo@correo.udistrital.edu.co, czafra@udistrital.edu.co

Abstract. The objective of this paper is to show the development of Box-Jenkins stochastic models to study the behavior of air pollutants concentrations in the megacity of Bogotá, Colombia. Information was collected from 10 stations in the city’s air quality monitoring network over a ten-year period. The temporal relationship between air pollutants, their spatial variation, and the occurrence of extreme pollution episodes was studied using Box-Jenkins models. The results showed that the moving average term of the models developed was the main indicator of spatial distribution for the daily pollutant concentrations. In the case of atmospheric particulate matter < 10 µm, the following spatial distribution was identified in the megacity: northwestern, center-southwest, and southeast. For atmospheric particulate matter < 2.5 µm: north, center, and southwest. For ozone: northwest, center, and south. Maximum hourly particulate matter concentrations were observed between 6–10 a.m., and between 11 a.m. - 4 p.m. for ozone. Monthly, the highest particulate matter concentrations were observed in February (14.1%), January (13.5%), and March (12.2%). In the context of atmospheric physics, this study was relevant for the following findings: The usefulness of Box-Jenkins models in simulating the temporal behavior of air pollutants, and for their adequate performance in detecting urban spatial trends.

1. Introduction

Air pollution has become a public health problem of great magnitude, especially in the large cities of the world. The city of Bogotá, Colombia, is one of the Latin American cities with the highest air pollution due to its economic and social dynamics [1]; several air quality studies conducted in the megacity have reported significant associations between air pollutants such as particulate matter < 10 µm (PM_{10}), particulate matter < 2.5 µm (PM_{2.5}), and ozone (O₃), and negative effects on human health [2]. Due to the air pollution risk in the megacity, environmental authorities use air quality modelling tools in order to establish actions and protocols for action in the face of alert and emergency episodes [3]. However, there is a lack of alternative simulation approaches to assess air quality in this study megacity. Several authors [4] have suggested the use of Box-Jenkins stochastic models to analyze the temporal behavior of air pollutants and generate projections, since they represent a possibility of providing inputs to strengthen the systems of air quality monitoring and control in the megacity.

The time series analysis is a fundamental tool for the study of phenomena as complex as those observed in the atmosphere physics, specifically those related to urban air pollutants; among the proposed approaches, the Box-Jenkins stochastic models stand out, due to their wide use in the time series analysis of air pollutants [5]. These models perform the prediction assuming that the model structure remains time-invariant, namely, that in the future the model is still suitable for modeling the
time series [6]. Box-Jenkins models consist of an autoregressive polynomial of order ‘p’, a difference polynomial of order ‘d’, and a moving average polynomial of order ‘q’ [7].

The main objective of this paper is to show the development of Box-Jenkins stochastic models to study the behavior of PM$_{2.5}$, PM$_{10}$, and O$_3$ concentrations in the megacity of Bogotá, Colombia; in the context of atmospheric physics, this study will be relevant for the following aspects: (i) use of Box-Jenkins models to simulate the temporal behavior of urban air pollutants; and (ii) to assess their performance in detecting spatial trends in atmospheric pollutant concentrations.

2. Materials and methods
This section initially shows a description of the study megacity and the equipment used during the sampling of air pollutants. Lastly, the methodology used for the development and validation of Box-Jenkins models is described.

2.1. Study site
The research area was demarcated by 10 stations that formed the air quality-monitoring network of Bogotá, Colombia: “Carvajal, Centro de Alto Rendimiento (C.A.R)”, Guaymaral, Kennedy, Las Ferias, Puente Aranda, San Cristóbal, Suba, Tunal, and Usaquén; in these stations, the monitoring of air pollutants (PM$_{10}$, PM$_{2.5}$, and O$_3$) and meteorological variables (rainfall, wind speed and direction, temperature, solar radiation, relative humidity, and atmospheric pressure) was carried out; on average, the monitoring stations were located at an elevation of 2600 m.a.s.l.

2.2. Information collection
The study periods for each air pollutant were adjusted depending on the availability of information: PM$_{10}$ between 2007-2017, PM$_{2.5}$ between 2014-2017, and O$_3$ between 2009-2017. Information for each pollutant was collected under an hourly timescale, with Met One Bam 1020 continuous monitoring equipment for PM$_{10}$, ThermoScientific FH62C14 for PM$_{2.5}$, and Teledyne API 400E/Ecotech ML9811 for O$_3$. All information was downloaded from the air quality monitoring network platform of Bogotá city, Colombia (http://201.245.192.252:81/home/map). Likewise, all weather information was downloaded from this platform.

2.3. Information analysis
From the time series of each study variable, the moving averages were calculated under the following timescales: daily, weekly, and monthly. Namely, the original time series were transformed under the previous timescales. Subsequently, Spearman’s coefficient (rs) was used to study the relationship between monitoring stations based on the air pollutants considered. The correlations using Spearman’s coefficient were evaluated according to the ranges recommended by Hernández and Fernández [8]. For the development of Box-Jenkins models, the methodology of these authors was applied [9]; this iterative procedure was carried out in four stages: identification, estimation, verification, and model prediction. The development of the models was carried out using the Expert-modeling Tool of the IBM-SPSS V.25.0.0 software [10].

The statistics used to verify the adjustment of the models developed were the following: determination coefficient ($R^2$), root-mean square error (RMSE), mean absolute percentage error (MAPE), Bayesian information criterion (BIC), and p-value of the Lung-Box statistic ($Q'$); in this regard, a p-value greater than 0.05 in $Q'$ indicated that Box-Jenkins model was properly developed [9]. The results of the models were used as input to identify additive and innovative outliers using the IBM-SPSS V.25.0.0 software [10]. These outliers were contrasted with respect to the environmental regulations in force in Colombia (Resolution 2254 of 2017) [11] using frequency distribution tables.

3. Results and discussion
This section initially shows an analysis of the selection of the best simulation timescale for the air pollutants under study. Subsequently, we display the Box-Jenkins models developed and a spatial
analysis based on their temporal structure. Lastly, an analysis of the extreme pollution episodes detected using Box-Jenkins models is shown.

3.1. Simulation timescales
The results showed that the Box-Jenkins models for PM and O₃ under a monthly timescale were the best compared to the adjustment statistics considered (see Table 1 and Table 2). The models under the daily and weekly timescales met the validation phase; however, the adjustment statistics (R², RMSE, MAPE, and p-value of Q’) were not better compared to the monthly timescale. The descending order in the adjustment of the models was as follows: monthly > weekly > daily. Thus, the results suggested that the monthly timescale was the best for studying the temporal behavior of the air pollutants considered in this study (PM₁₀, PM₂.₅, and O₃). However, in practice, this timescale may not be useful for declaring alert or emergency status for air pollution. Namely, in practice a rapid response to these extreme episodes of air pollution is required, where the daily timescale is possibly the most appropriate.

Table 1. Box-Jenkins models for PM₁₀ concentrations under a monthly timescale.

| Station  | p  | d  | q  | Transf. | R²  | % RMSE | % MAPE | P-value | BIC  |
|----------|----|----|----|--------|-----|--------|--------|---------|------|
| Carvajal | 2  | 1  | 9  | NT     | 0.998 | 0.041 | 0.042 | 0.089 | -6.383 |
| C.A.R.   | 1  | 1  | 8  | SR     | 0.998 | 0.022 | 0.055 | 0.097 | -7.652 |
| Guaymaral| 1  | 1  | 9  | NT     | 0.998 | 0.017 | 0.042 | 0.360 | -8.088 |
| Kennedy  | 3  | 1  | 3  | NT     | 0.998 | 0.032 | 0.035 | 0.090 | -6.895 |
| Las Ferias | 2  | 1  | 3  | NT     | 0.998 | 0.021 | 0.044 | 0.396 | -7.761 |
| Puente Aranda | 1  | 1  | 8  | NT     | 0.998 | 0.030 | 0.041 | 0.149 | -6.998 |
| San Cristóbal | 1  | 1  | 10 | NT    | 0.998 | 0.017 | 0.051 | 0.061 | -8.100 |
| Suba     | 1  | 1  | 8  | NT     | 0.998 | 0.022 | 0.034 | 0.763 | -7.591 |
| Tunel    | 2  | 1  | 7  | NT     | 0.998 | 0.026 | 0.045 | 0.074 | -7.254 |
| Usaquén  | 1  | 1  | 13 | NT     | 0.998 | 0.016 | 0.038 | 0.057 | -8.297 |

*Transformation type, LN = Natural logarithm, NT = no transformation, and SR = Square root.

Table 2. Box-Jenkins models for PM₂.₅ concentrations under a monthly timescale.

| Station  | p  | d  | q  | Transf. | R²  | % RMSE | % MAPE | P-value | BIC  |
|----------|----|----|----|--------|-----|--------|--------|---------|------|
| Carvajal | 1  | 1  | 11 | LN     | 0.999 | 0.013 | 0.013 | 0.239  | -8.656 |
| C.A.R.   | 3  | 1  | 2  | NT     | 0.999 | 0.013 | 0.056 | 0.054  | -8.751 |
| Guaymaral| 1  | 2  | 3  | NT     | 0.999 | 0.010 | 0.050 | 0.055  | -9.220 |
| Kennedy  | 1  | 1  | 9  | LN     | 0.999 | 0.017 | 0.046 | 0.076  | -8.191 |
| Las Ferias | 2  | 1  | 6  | NT     | 0.999 | 0.013 | 0.062 | 0.094  | -8.613 |
| San Cristóbal | 2  | 1  | 4  | NT    | 0.999 | 0.006 | 0.049 | 0.089  | -10.177 |
| Suba     | 1  | 2  | 4  | NT     | 0.999 | 0.013 | 0.048 | 0.130  | -8.719 |
| Tunel    | 1  | 1  | 10 | SR     | 0.999 | 0.017 | 0.060 | 0.060  | -8.203 |
| Usaquén  | 2  | 1  | 2  | NT     | 0.999 | 0.011 | 0.062 | 0.367  | -9.087 |

*Transformation type, LN = Natural logarithm, NT = no transformation, and SR = Square root.

3.2. Spatial distribution
The results showed by a Spearman’s correlation analysis (rs), the existence of significant direct correlations between all monitoring stations for observed PM and O₃ concentrations (rs between 0.30 - 0.92); this for the daily, weekly, and monthly timescales. Thus, the results suggested weak to very strong correlations between the air pollutants concentrations observed in the monitoring stations of the megacity. This coincided with the Box-Jenkins models developed for PM and O₃ concentrations. Namely, similarity was observed in the temporal structure of the models developed with respect to the autoregressive term (p); this similarity was most evident under the daily timescale. All models developed showed a first-order autoregressive term (p = 1) (Table 3). In other words, the concentrations of PM₁₀, PM₂.₅, and O₃ were influenced by the concentrations observed the day before. Due to the practical need to respond quickly to extreme episodes of air pollution, the daily timescale was selected to study the spatial distribution of air pollutants. Indeed, this spatial distribution of air pollutants was analyzed from
the similarity or difference in the daily temporal structure of the Bok-Jenkins models generated for each sector of the megacity.

The moving average term \( q \) of the Box-Jenkins models suggested a spatial distribution based on the daily behavior of air pollutants (Table 3). Indeed, this term had a similar or very close magnitude between nearby monitoring stations. For the case of PM\(_{10}\), the following spatial distribution was identified in the megacity: northwestern (1,1,4), center-southwest (1,1,6-9), and southeast (1,1,4). In relation to the PM\(_{2.5}\) concentrations, the following spatial distribution was detected: north (1,1,10), center (1,1,4), and south-west (1,1,9). This spatial distribution was possibly related to wind direction and speed, and land use [12]. The results showed the following spatial distribution for O\(_3\): northwest (1,1,4), center (1,1,9), and south (1,1,5). The temporal structure of the models developed between nearby monitoring stations was similar.

For example, in the northwest of the megacity, the nearby stations of Suba and Guaymaral showed a model of Box-Jenkins (1,1,4) for O\(_3\). Indeed, in this area of the megacity, the highest values of solar radiation were observed (monthly-daytime average: 532.9 W/m\(^2\)) compared to the central and southern zones (monthly-daytime average: 510.3 W/m\(^2\) and 425.0 W/m\(^2\), respectively). This possibly influenced the higher O\(_3\) concentrations to be observed. However, this northwestern area of the megacity showed the lowest concentrations of O\(_3\) precursors. This coincided with that reported by other authors in places subject to strong solar radiation [13]. Lastly, the difference term \( d \) of the developed models suggested a weak decreasing trend in daily concentrations of air pollutants under study \((d = 1)\).

### Table 3. Terms of daily Box-Jenkins models for air pollutants under study.

| Station       | PM\(_{10}\) |   | PM\(_{2.5}\) |   | O\(_3\) |   |
|---------------|-------------|---|-------------|---|--------|---|
|               | p | d | q | p | d | q | p | d | q |
| Carvajal      | 1 | 1 | 9 | 1 | 1 | 9 | 1 | 1 | 5 |
| C.A.R.        | 1 | 1 | 6 | 1 | 1 | 4 | 1 | 1 | 9 |
| Guaymaral     | 1 | 1 | 5 | 1 | 1 | 10 | 1 | 1 | 4 |
| Kennedy       | 1 | 1 | 6 | 1 | 1 | 9 | - | - | - |
| Las Ferias    | 1 | 1 | 6 | 1 | 1 | 4 | 1 | 1 | 9 |
| Puente Aranda | 1 | 1 | 6 |- | - | - | 1 | 1 | 10 |
| San Cristóbal | 1 | 1 | 4 | 1 | 1 | 10 | 1 | 1 | 5 |
| Suba          | 1 | 1 | 5 | 1 | 1 | 10 | 1 | 1 | 4 |
| Tunal         | 1 | 1 | 4 | 1 | 1 | 9 | 1 | 1 | 5 |
| Usaquén       | 1 | 1 | 4 | 1 | 1 | 10 | 1 | 1 | 5 |

#### 3.3. Extreme pollution episodes

In the identification of outliers using Box-Jenkins models, two types were considered: additive and innovative. An additive outlier appeared as an unexpectedly high or low value that occurred for a single observation and subsequent observations were not affected. An innovative outlier was characterized by an initial impact that extended over the following observations. The results showed that 12.4% and 8.67% of the total outliers for PM\(_{10}\) and PM\(_{2.5}\) exceeded the Colombian limit for 24 h, respectively (PM\(_{10}\): 75 μg/m\(^3\) and PM\(_{2.5}\): 37 μg/m\(^3\); Resolution 2254 of 2017 [11]).

For PM\(_{2.5}\) – 24 h, the excesses were related to prevention status (38 μg/m\(^3\) - 55 μg/m\(^3\)) and alert status (56 μg/m\(^3\) - 150 μg/m\(^3\)) in air quality according to Colombian regulations. In the case of PM\(_{10}\) – 24 h and O\(_3\) – 8 h, the excesses only remained in the prevention status (PM\(_{10}\): 155 μg/m\(^3\) - 254 μg/m\(^3\); O\(_3\): 139 μg/m\(^3\) - 167 μg/m\(^3\)). With the time series of air pollutants that did not have a normative reference value, as were the weekly and monthly series, the identified outliers were compared with the multi-year mean.

The frequency distribution of outliers showed the hours and months where the greatest amounts of maximum outliers were observed. The extreme hourly episodes of PM associated with prevention and alert status were observed between 7 a.m. - 8 a.m. (Figure 1). The occurrence frequencies for these additive and innovative outliers were 48.5% and 44.1%, respectively. The results also showed that the time interval where maximum outliers were most observed for O\(_3\) was between 11 a.m. - 4 p.m. The
additive and innovative outliers had an occurrence frequency of 64.0% and 67.5%, respectively. These results in the identification of maximum outliers for PM and O$_3$ using Box-Jenkins models, suggested that these started as additives, and subsequently became innovative due to their effects on future observations.

![Figure 1. Hourly occurrence frequency (%) of maximum additive and innovative outliers for PM$_{10}$.](image1)

In relation to the months in which more outliers were observed for air pollutants, a high occurrence was observed from January to March for PM$_{10}$. The occurrence frequency of additive and innovative outliers for these months was 45.4% and 43.1%, respectively. The order of importance in the occurrence of outliers for PM$_{10}$ was February > January > March (Figure 2). In the case of PM$_{2.5}$, it was observed that outliers tended to be concentrated in September, followed by October and November.

The occurrence frequency of additive and innovative outliers in the month of September was 20.3% and 32.5%, respectively. In the case of O$_3$, the results showed that the maximum outliers tended to occur during the months of February and October. The occurrence frequency of outliers (additive and innovative) for these months was 28.9%. It is important to mention that in the months where the greatest amount of minimal outliers were observed for PM$_{10}$, PM$_{2.5}$, and O$_3$ concentrations, the highest wind speeds occurred. This according to the monthly historical behavior of wind speed; some researchers reported similar results [14].

![Figure 2. Monthly occurrence frequency (%) of outliers for PM$_{10}$.](image2)

Therefore, the results hinted at the influence of wind direction on the occurrence of atypical concentrations or episodes of alert for air pollution in the study megacity. This is possibly because the synoptic winds represented a component with greater influence from the northeast and, locally, the winds entered from the west, which possibly caused a stagnation of the ventilation in the megacity [15]. Indeed, there was evidence of other atmospheric phenomena that could possibly increase the concentrations of air pollutants in the megacity such as abnormal wind regimes influenced by the settlement of hurricanes in the Atlantic Ocean, and PM drag due to the presence of wildfires and open burning. In addition, the PM contribution from the Sahara Desert was reported, and from extreme events related to the dust resuspension from unpaved areas [16].
4. Conclusions

The results show that the best timescale to simulate the temporal behavior of PM$_{10}$, PM$_{2.5}$, and O$_3$ concentrations is the monthly one. This is compared to the adjustment statistics (determination coefficient, root-mean square error, and mean absolute percentage error) of the daily and weekly timescales. However, in practice, the monthly time scale may not be useful for declaring alert or emergency status due to air pollution. In these cases, the response must be faster and that is when the daily timescale is possibly the most useful.

Temporal analysis using Box-Jenkins models shows that PM$_{10}$, PM$_{2.5}$, and O$_3$ concentrations are influenced by the concentrations observed the day before. In other words, extreme episodes of air pollution may persist for two days in the study megacity. This from the analysis of the autoregressive term of the developed models.

The findings suggest that the moving average term of the models allows analyzing the spatial distribution of air pollutants. It is observed that nearby monitoring stations have a similar magnitude in this term. Namely, there is a similar variation in the air pollutants concentrations from nearby stations. Indeed, the detail in spatial analysis depends on the number of monitoring stations considered.

Within the framework of atmospheric physics, the findings of this research are useful for deepening knowledge regarding the use of Box-Jenkins models to study the temporal behavior of air pollutants, their spatial distribution, and the occurrence of extreme pollution episodes in megacities.

In this study, the following future research line is visualized: To perform a comparative analysis using Box-Jenkins models between air pollutant concentrations and variables associated with atmospheric physics (rainfall, temperature, solar radiation, and wind direction and speed). This is to evaluate its influence on the spatiotemporal behavior of air pollutant concentrations.

References

[1] Mangones S, Jaramillo P, Rojas N, Fischbeck P 2020 Air pollution emission effects of changes in transport supply: the case of Bogotá, Colombia *Environmental Science and Pollution Research* 27(29) 35971

[2] Hernández-Flórez L J, Aristizabal-Duque G, Quiroz L, Medina K, Rodríguez-Moreno N, Sarmiento R, Osorio-Garcia S D 2013 Contaminación del aire y enfermedad respiratoria en menores de cinco años de Bogotá, 2007 *Revista de Salud Pública* 15 552

[3] Zárate E, Belalcázar L, Clappier A, Manzi V, Van den Bergh H 2007 Air quality modelling over Bogotá, Colombia: Combined techniques to estimate and evaluate emission inventories *Atmospheric Environment* 41(29) 6302

[4] Naseem F, Rashid A, Izhar T, Khawar M I, Bano S, Ashraf A, Adnan M N 2018 An integrated approach to air pollution modeling from climate change perspective using ARIMA forecasting *Journal of Applied Agriculture and Biotechnology* 2(2) 37

[5] Gopu P, Panda R R, Nagwani N K 2021 Time series analysis using ARIMA model for air pollution prediction in Hyderabad city of India *Soft Computing and Signal Processing* ed Reddy V S, Prasad V K, Wang J, Reddy K T V (Singapore: Springer) p 47

[6] Sahai A K, Rath N, Sood V, Singh M P 2020 ARIMA modelling & forecasting of COVID-19 in top five affected countries *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 14(5) 1419

[7] Mishra S, Swain D K, Mishra D, Santra G H 2021 Time series analysis of rainfall for Puri district of Odisha using ARIMA modelling *Intelligent and Cloud Computing*, vol. 2 ed Mishra D, Buuya R, Mohapatra P, Patnaik S (Singapore: Springer) p 419

[8] Hernandez-Sampieri R 2014 *Metodología de la Investigación* (México: Mc Graw Hill)

[9] Box G E P, Jenkins G M, Reinsel G C, Ljung G M 2016 *Time Series Analysis: Forecasting and Control* (New Jersey: John Wiley & Sons)

[10] International Business Machines Corporation (IBM) 2021 *Statistical Package for Social Sciences (SPSS), IBM Forecasting 25* (United States of America: International Business Machines)

[11] Ministerio de Ambiente y Desarrollo Sostenible 2017 *Resolución 2254 de 2017* (Colombia: Ministerio de Ambiente y Desarrollo Sostenible)

[12] Ramírez O, Sánchez de la Campa A M, Amato F, Catacolí R A, Rojas N Y, de la Rosa J 2018 Chemical composition and source apportionment of PM$_{10}$ at an urban background site in a high–altitude Latin American megacity (Bogota, Colombia) *Environmental Pollution* 233 142
[13] Nassif W G, Al-Taai O T, Abbood Z M 2020 The Influence of solar radiation on ozone column weight over Baghdad city *IOP Conf. Ser.: Mater. Sci. Eng.* **928**(1) 072089:1

[14] Molina-Gómez N I, Calderón-Rivera D S, Sierra-Parada R, Diaz-Arévalo J L, López-Jiménez P A 2021 Analysis of incidence of air quality on human health: a case study on the relationship between pollutant concentrations and respiratory diseases in Kennedy, Bogotá *Int. J. Biometeorol.* **65**(1) 119

[15] Secretaría Distrital de Ambiente (SDA) 2020 *Informe Mensual Calidad del Aire RMCAB Junio de 2020* (Bogotá: Secretaría Distrital de Ambiente)

[16] Guzmán J S, Hernández H T, Mora J M, Johan J E P, Vanegas S 2019 Spatial and temporal analysis of particulate material concentrations in Bogota: A year of operation of an independent network *Congreso Colombiano y Conferencia Internacional de Calidad de Aire y Salud Pública (CASAP 2019)* (Barranquilla: Universidad del Norte)