Hourly Ozone and PM_{2.5} Prediction Using Meteorological Data – Alternatives for Cities with Limited Pollutant Information

Felipe Cifuentes^{1*}, Angel Gálvez¹, Carlos M. González¹,
Mauricio Orozco-Alzate², Beatriz H. Aristizábal^{1*}

¹Hydraulic Engineering and Environmental Research Group, Universidad Nacional de Colombia Sede Manizales, Cra 27 64-60 Bloque H Palogrande. Manizales, Colombia

²Department of Informatics and Computing, Universidad Nacional de Colombia Sede Manizales, Km 7 vía al Magdalena Bloque S La Nubia. Manizales, Colombia

ABSTRACT

Using statistical models, the average hourly ozone (O₃) concentration was predicted from seven meteorological variables (Pearson correlation coefficient, $R = 0.87\text{--}0.90$), with solar radiation and temperature being the most important predictors. This can serve to predict O₃ for cities with real time meteorological data but no pollutant sensing capability. Incorporating other pollutants (PM_{2.5}, SO₂, and CO) into the models did not significantly improve O₃ prediction ($R = 0.91\text{--}0.94$). Predictions were also made for PM_{2.5}, but results could not reflect its peaks and outliers resulting from local sources. Here we make a comparative analysis of three different statistical predictor models: (1) Multiple Linear Regression (MLR), (2) Support Vector Regression (SVR), and (3) Artificial Neuronal Networks (ANNs) to forecast hourly O₃ and PM_{2.5} concentrations in a mid-sized Andean city (Manizales, Colombia). The study also analyzes the effect of using different sets of predictor variables: (1) Spearman coefficients higher than ± 0.3 , (2) variables with loadings higher than ± 0.3 from a principal component analysis (PCA), (3) only meteorological variables, and (4) all available variables. In terms of the O₃ forecast, the best model was obtained using ANNs with all the available variables as predictors. The methodology could serve other researchers for implementing statistical forecasting models in their regions with limited pollutant information.

Keywords: Tropospheric ozone, Particulate matter, Hourly concentrations, Andean city, Support Vector Regression, Artificial Neuronal Network

1 INTRODUCTION

Understanding the patterns of the complex relationships between meteorology and air pollution is of great interest in air quality prediction (Zeng *et al.*, 2020). Many of the studies predict the daily concentrations of pollutants (Weizhen *et al.*, 2014; Franceschi *et al.*, 2018; Mehdipour *et al.*, 2018; Sihag *et al.*, 2019). However, understanding the hourly variation and their relationship with the meteorological variables (e.g., solar radiation and temperature) would be important for studying the emission patterns and prediction (Sekar *et al.*, 2015a; Franceschi *et al.*, 2018; Cuesta *et al.*, 2020). The option of using only meteorological variables to forecast pollutant concentrations could provide concentrations of pollutants in areas where only meteorological variables are monitored, or as an alternative when air pollutant measurements fail; indeed, according to the World Meteorological Organization (WMO) (2020), the measurement of meteorological variables has better geographic coverage around the world.

Numerical (deterministic) and statistical approaches are the two most common methods for air quality forecasting. The deterministic approach involves the use of chemical transport models (CTMs), which simulate dynamics of the atmosphere from the mathematical representation of

OPEN ACCESS

Received: July 30, 2020

Revised: April 4, 2021

Accepted: June 2, 2021

* Corresponding Authors:

Felipe Cifuentes

fcifuentesc@unal.edu.co

Beatriz H. Aristizábal

bharistizabalz@unal.edu.co

Publisher:

Taiwan Association for Aerosol
Research

ISSN: 1680-8584 print

ISSN: 2071-1409 online

© Copyright: The Author(s).

This is an open access article distributed under the terms of the [Creative Commons Attribution License \(CC BY 4.0\)](https://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are cited.

different physical and chemical mechanisms (Hoshyaripour *et al.*, 2016). CTMs are mostly use for forecasting over extender areas (Zhang *et al.*, 2012; Zeng *et al.*, 2020). These deterministic-type models fail to explain the nonlinearity and heterogeneity of atmospheric processes (Vlachogianni *et al.*, 2011).

In contrast, statistical models can deal with nonlinear data and are less strict with the input requirements. The statistical approach uses field-measurements (e.g., meteorological and air quality data) to build a model (linear or nonlinear) able to predict air pollutant concentrations in response to changes of predictor variables (Jia *et al.*, 2019). These models use ground-level data and establish relationships among variables, providing a better insight into the particularities of the study area. This characteristic is considered simple, economical, and easy to implement; being a valuable alternative in cities of emerging countries, where forecasting exercises are scarce and the missing data is a common issue (Jia *et al.*, 2019).

A key factor in the development of the statistical forecasting models is selecting adequate predictor variables. Using a reduced number of representative features as input parameters can reduce the computational cost as well as improve the performance of the model. One common approach for selecting features consists of using statistical methods to evaluate the relationship between the target variable and the input variables, selecting the variables that exhibit strong relationships. The most common statistical methods for selecting numeric features are Pearson correlation (linear data) and spearman correlations (nonlinear data). However, other techniques such as principal component analysis (PCA) have also been tested (Franceschi *et al.*, 2018).

There are a wide variety of statistical models that have been used in air pollution studies; from the traditional and simple multiple linear regression (MLR) method (Dimitriou, 2016) to more sophisticated techniques such as artificial neural networks (ANNs), and support vector regression (SVR) for nonlinear systems (Karimian *et al.*, 2019). The advantages of using ANNs and SVR methods for air quality prediction have been proposed in previous studies. For example, Kisi *et al.* (2017) obtained a better accuracy using least square SVR when forecasting monthly SO₂ concentrations in Delhi, India. Mogollón-Sotelo *et al.* (2021) validate the capability of SVR techniques to forecast PM_{2.5} in an area of complex terrain such as the city of Bogota, Colombia. Similarly, Guo *et al.* (2020) discussed the benefits of using ANNs for PM_{2.5} forecasting, compared to multi-variate and linear regression analyses.

In tropical cities, the weather patterns are complex. According to the altitude, weather conditions such as solar radiation present high values (e.g., higher than 1200 W m²), and temperature tends to be low (Franceschi *et al.*, 2018) with significance changes during the day. An example of these complex characteristics is the medium-sized Andean city of Manizales, Colombia, located on the western slope of the Cordillera Central (2150 m.a.s.l), 28 km away from the Nevado del Ruiz volcano (5321 m.a.s.l). The city is characterized by maximum solar radiation around ~1300 W m⁻² and daily average air temperatures near to 18°C, with daily variations of temperature between 14 to 23°C, and solar radiation between 0 to 897 W m⁻². Additionally, the city present typically mountain—valley airflows patterns with low average wind speed (< 2 m s⁻¹), where the air masses flow upslope during the day due to convective heat and buoyancy forces, and downslope at night. Most of the primary pollutants in the city are emitted by a common source, which corresponds to the internal combustion vehicles, with a typical diurnal cycle related to the rush-hours, and a reduction of activities in the early morning (CORPOCALDAS and UNAL, 2020; Cuesta *et al.*, 2020).

This study is divided into two main sections. First, different methods for selecting adequate features as predictor variables for forecasting O₃ and PM_{2.5} hourly concentrations in Manizales, Colombia were evaluated. Secondly, the different sets of predictor variables were tested with different statistical predictor models. The different combinations of predictors—regression models were used to forecast a month (January 1, 2020, to January 30, 2020) of hourly O₃ and PM_{2.5} concentrations. The modeling results were compared with surface measurements to assess which option provides the most accurate forecast. Finally, the best performing models for O₃ and PM_{2.5} forecasting were used to predict one additional month (May 1, 2020, to May 31, 2020) to test the models under different atmospheric conditions (Different chronological period and reduce emissions due to COVID-19 mobility restrictions in Manizales).

The results showed the sensitivity of O₃ prediction to meteorological variables, and the reduce impact that including pollutant concentrations (PM_{2.5}, SO₂, and CO) has on improving the forecasting accuracy. This shows the possibility of developing O₃ forecasting models exclusively

from meteorological predictors. On the contrary, $PM_{2.5}$ predictions are most sensitive to variations in SO_2 and CO concentrations, nonetheless, the models follow the average trend of $PM_{2.5}$ concentrations but are unable to forecast the peaks of concentrations associate with particular emission scenarios.

The statistical exploratory techniques, such as spearman correlations and PCA reduce the complexity of the models and improve accuracy. Furthermore, the results show the sensitivity of the different regression techniques to changes in the predictor variables and assess which combination of predictors–regression model offers better performance. This study can guide the development of future statistical forecasting models, which can be applied to deal with missing data in emerging countries with low budgets in air quality management, and to developed forecasting tools in areas that could not afford a more robust system, such as deterministic forecasting models.

2 METHODS

Fig. 1 shows the methodology framework proposed to develop the statistical forecasting models for O_3 and $PM_{2.5}$ concentrations (hourly scale) in Manizales, Colombia. First, the initial preparation of the raw air quality and meteorological data was performed, including the deletion of outliers and a min-max normalization (Wilks, 2019). The next step consisted of the selection of the predictor variables used in the models. For this purpose, four different methodologies were employed: (1) variables that exhibit Spearman coefficients higher than ± 0.3 , in relation to O_3 and $PM_{2.5}$, (2) variables with loadings higher than ± 0.3 and related to O_3 and $PM_{2.5}$ concentrations, during a principal component analysis (PCA), (3) only meteorological variables, and (4) all available variables.

Each group of predictor variables was evaluated separately as inputs for the statistical models using three regression techniques: (1) MLR, (2) SVR, and (3) ANNs. All models underwent a training, validation, and testing stage. Hence, a partition of the original dataset was required for training, validation, and test, as specified in Fig. 1. Each model, with its corresponding input variables, was developed and tested for forecasting an entire month (January 1, 2020, to January 30, 2020) of hourly O_3 and $PM_{2.5}$ concentrations. Model predictions were compared against ground observations to assess the predictive capabilities of the different models. Finally, the models that offered the best performance were used to predict an additional month of O_3 and $PM_{2.5}$ concentrations (May 1, 2020, to May 31, 2020) to evaluate the models over a different chronological period, with different emission scenarios (e.g., first month of COVID-19 mobility restrictions in the city).

The comparisons were made base on different performance statistics such as Mean Bias (MB), Mean Absolute Error (MAE), Root Mean Squared Error (RMSE), Normalized Root Mean Squared Error (NRMSE), and Pearson correlation coefficient (R). See the definition of this metrics in Equations S1 to S5.

2.1 Study Area and Raw Data Collection

Fig. S1 shows the study area covering the urban area of Manizales, Colombia. It also shows the location of monitoring stations and an orographic illustration of airflows patterns (Cuesta *et al.*, 2020). The air masses around the city flow upslope during the day due to convective heat and buoyancy forces, and downslope at night (CORPOCALDAS and UNAL, 2020; Cuesta *et al.*, 2020). Manizales, Caldas–Colombia is a medium-sized Andean city located on the western slope of the Cordillera Central (2150 m.a.s.l), 28 km away from the Nevado del Ruiz volcano (5321 m.a.s.l), which is located to the Southeast of the city. By 2018, the city had a population of 400434 inhabitants (DANE, 2019), in an urban area of 54 km². The city is characterized by tropical weather with low average wind speed ($< 2 \text{ m s}^{-1}$), high relative humidity ($> 60\%$), low variation of temperature (12–24°C), atmospheric pressure around 785–795 h Pa, maximum solar radiation around $\sim 1300 \text{ W m}^{-2}$ and total annual rainfall around 1755 mm.

Air quality five-minute records were obtained from a central station of the city (5°4′6.5″N, 75°31′1.7″W, 2155 m.a.s.l. See flag “GOB” in Fig. S1). The nearby area is influenced by street cannons, high circulation of private and public vehicles, and a near industrial source (food industry). According to the last local emission inventory (base year 2017), on-road vehicle emissions were

Fig. 1. Methodology framework.

the most significant source of pollution in the city, contributing to more than 80% of emissions of all criteria pollutants; only for the case of SO₂ this contribution was 12%. (CORPOCALDAS and UNAL, 2019).

The five-minutes air quality measurements were obtained using the following equipment: Carbon monoxide (CO) infrared absorption analyzer, method gas filter correlation (GFC) (Thermo 48i - ACSCC), tropospheric ozone (O₃) UV absorption analyzer (Teledyne - Model T400), sulfur dioxide (SO₂) UV fluorescence analyzer (Teledyne - Model T100), Particulate matter PM_{2.5} photometer (Air Pollution Monitor 2, APM-2) (CORPOCALDAS and UNAL, 2020; Cuesta *et al.*, 2020).

On the other hand, meteorological five-minute records were extracted from a nearby meteorological station (~2 km) (5°3'46.6"N, 75°30'2.1"W, 2183 m.a.s.l. See the point called "HOS" in Fig. S1). The variable records showed stability during the whole period of analysis. The information was downloaded from The Caldas Environmental Data and Indicators Center (CDIAC), a public online platform (<http://cdiac.manizales.unal.edu.co>). The meteorological variables used were temperature (T), solar radiation (SR), relative humidity (RH), precipitations (Ppt), wind speed (ws), the meridional (wd_i) and zonal (wd_j) components of the wind direction, estimated from its measurement in degrees. Finally, air quality and meteorological data records were collected from October 1, 2019, to January 30, 2020, and May 1, 2020, to May 31, 2020. A full description of the periods of time used in each stage of the process can be consulted in Table S1.

2.2 Data Preparation

The raw data were averaged to an hourly time scale only if over 75% of the five-minute records were captured during each hour. Fig. 1 shows the total number of hourly average records obtained during the study period. For air quality data, the values were converted to $\mu\text{g m}^{-3}$ units at standard conditions (25°C and 1.013 h Pa). The atypical values were deleted following a univariate approach, in which observations laying outside 3 times the interquartile range (IQR) were removed from the dataset.

Finally, the dataset was scaled using a min-max normalization to guarantee that all the variables studied vary in a range from 0 to 1, following the approach of other similar studies (Voukantsis *et al.*, 2011; Franceschi *et al.*, 2018). The normalization helps to compare all data on a similar scale and get the same initial weight during the mathematical analysis (Wilks, 2019).

2.3 Selection of Predictor Variables

Different statistical techniques were applied to the dataset to explore the relationship among air quality and meteorological variables, and thus, define which of these variables are suitable to be used as predictors in subsequent regression models. The methods applied were the following:

2.3.1 Spearman coefficient

A preliminary analysis of the concentration histograms (Fig. S2) shows that pollutants presented a non-normal distribution, and even the transformation of data by using a base 10 logarithm did not guarantee a normal distribution. For this reason, the Spearman coefficient was used to evaluate the strength and direction of relationships between two sets of non-normal distributed variables. Its value ranges from -1 (a perfect negative correlation between variables) to $+1$ (a perfect positive correlation) (Schober *et al.*, 2018).

Spearman coefficients were estimated for all variables in the design dataset, to establish the relationships among meteorology and pollutant concentrations. In this study, variables that exhibited coefficients greater than ± 0.3 in relation to O_3 and $\text{PM}_{2.5}$ concentrations were selected as predictor variables for the subsequent regression models. Correlation coefficients lower than ± 0.3 were considered weak or negligible correlations.

2.3.2 Principal component analysis (PCA)

PCA is a technique for reducing the dimensionality of large datasets, increasing interpretability, and minimizing the loss of information (Jolliffe and Cadima, 2016). In addition to reducing data dimension, PCA has been applied in other air quality studies for identification of inter-correlations within pollutants and meteorological variables. This technique could support the recognition of the most dominant parameters and mechanisms affecting pollutants concentration (Voukantsis *et al.*, 2011; Binaku and Schmeling, 2017; Sharma *et al.*, 2017).

A PCA analysis was applied to the centered (mean = 0) and scaled (standard deviation = 1) data set, for obtaining the different principal components (PCs) with their respective values of variable loadings, eigenvalues, and the represented variance of the original dataset. The number of PCs to retain was defined using Kaiser's rule, according to which PCs with eigenvalues lower than one should be discarded since they do not add significant information (Wilks, 2019).

Only the variables with a loading greater than ± 0.3 in the retained PCs, related to O_3 and $\text{PM}_{2.5}$, were considered for interpretation and selected as candidates to become the predictor variables,

as suggested by Binaku and Schmeling (2017). Since smaller values of loadings represent insignificant relationships or are affected by noise in the data.

2.3.3 Use of all meteorological variables

For comparison purposes, a third alternative was implemented using meteorological variables as predictors. The measurement of these variables has better geographic coverage and are presented as an option in areas where only meteorological variables are monitored.

2.3.4 Use of all available variables

Finally, the fourth alternative represents the scenario in which all variables are available to be used, which include both air pollutants and meteorological variables. This scenario is proposed as a complement to the variables pre-selected using Spearman correlations and PCA. These methodologies might not identify nonlinear representative interactions between variables. Therefore, by using all possible predictor variables, it is possible to have a point of comparison to judge the performance of the models developed with limited features.

2.4 Statistical Forecasting Models

The first step consisted of performing several sensitivity analyses to determine the optimal parameter for the non-linear regression techniques. Then, using the appropriate parameters, the models were trained and validated using a 10-fold cross-validation procedure on the design data (October 1, 2019, to December 31, 2019). This procedure aimed to improve accuracy and avoid over-fitting of the models, as suggested by Weizhen *et al.* (2014). At the end of this process, the model that presented the lowest RMSE on the validation set was selected as the best model for each regression technique. The regression techniques that were implemented and further details regarding its structure definitions are given below.

2.4.1 Multiple linear regression (MLR)

MLR was used to determine the correlation between one dependent variable and two or more independent variables having a cause-effect relationship. Then, a linear equation that represents those relations is formulated to make predictions (Uyanik and Güler, 2013). This regression technique has been used in different air-quality-related studies (Vlachogianni *et al.*, 2011). This technique was applied to have a baseline comparison with the other non-linear methods tested.

2.4.2 Support vector regression (SVR)

SVR is a machine learning algorithm specially developed for pattern recognition and classification (Luna *et al.*, 2014). The model consists of several support vectors and non-linear model coefficients, requiring parameters of insensitivity zone (ϵ) and penalty (c). Both parameters determine a trade-off between the training error and the model complexity (Kecman, 2001).

SVR was applied using a radial basis function (RBF) kernel. In order to define the values of ϵ and c parameters, a grid search was made, varying ϵ from 0 to 1 (in 0.01 increment) and c from 0 to 100 (0.1 increments). The best parameters were chosen based on which data pair offered the lowest RMSE value on the cross-validation procedure.

2.4.3 Artificial neural networks (ANNs)

ANNs are a machine learning algorithm which simulates the function of a biological nervous system. This method is useful in a wide variety of applications, including the control of complex non-linear systems, optimization, system identification, and pattern recognition (Jiang *et al.*, 2017). The basic structure of an ANN consists of an input layer, hidden layers, and an output layer. Every layer is composed of a certain number of neurons.

A feedforward backpropagation neural network with a linear activation function was developed for this study. The problem was limited to a single hidden layer and the number of neurons within the hidden layer was selected through a sensitivity analysis, in which different models were developed varying the number of neurons from 2 to 10. Based on RMSE values, the elbow method was used to define the optimal number of neurons.

2.5 Models Application and Evaluation Criteria

All the combinations of predictor variables and regression models, that were selected after the training and validation stage, were implemented to forecast O_3 and $PM_{2.5}$ hourly concentrations for a whole month (January 1, 2020, to January 30, 2020). The results obtained when applying the different models were compared against ground measurement with different performance measures. For this purpose, the following measures were used: R coefficients, MB, RMSE, NRMSE and MAE. The definitions of these metrics can be found in equations S1 to S5.

The best models for forecasting O_3 and $PM_{2.5}$ concentrations were used to forecast an additional month (May 1, 2020, to May 31, 2020) of hourly concentrations, to test the ability of the models in a different period. This exercise could indicate if the regression techniques applied are suitable for forecasting O_3 and $PM_{2.5}$ concentrations in Manizales. Furthermore, the results could indicate if the procedures undertook for selecting features as input parameters were useful.

3 RESULTS AND DISCUSSION

3.1 Statistics Summary of Ground Measurements for the Period of Analysis

Table S2 shows the statistics summary of hourly average measured concentration for the four pollutants and some meteorological variables during the study period (October 01, 2019, to January 30, 2020) in Manizales. Overall, the daily mean concentration of CO , O_3 , $PM_{2.5}$, and SO_2 are $839.3 \mu g m^{-3}$, $15.2 \mu g m^{-3}$, $12.6 \mu g m^{-3}$ and $4.3 \mu g m^{-3}$, respectively.

Fig. 2 shows the measured hourly concentration time series plot, boxplot, and hourly average of pollutants during the study period. The pollutant concentrations showed a constant pattern around the whole period, with concentrations ranging from $399\text{--}1618 \mu g m^{-3}$, $4\text{--}37 \mu g m^{-3}$, $4\text{--}21 \mu g m^{-3}$ and $2\text{--}8 \mu g m^{-3}$ for CO , O_3 , $PM_{2.5}$ and SO_2 respectively, according to the 5th and 95th percentile. The variability of concentrations could be explained by local effects, influenced by urban emission patterns and daily meteorological forcing (Cuesta *et al.*, 2020).

Daily patterns showed similarities of diurnal variability for CO , $PM_{2.5}$, and SO_2 which generally exhibited a flat “W” shape, influenced by the traffic rush-hour, and suggesting that in Manizales these pollutant concentrations are influenced mainly by vehicular emissions. These patterns are concordant with the local emission inventory of the city (CORPOCALDAS and UNAL, 2019).

On the contrary, the diurnal variability of O_3 exhibited a flat “U” shape. This pattern is caused by the solar radiation intensity that catalyzes photochemical reactions from hydrocarbons and nitrogen oxides precursor in the air (Khedairia and Khadir, 2012). The night peak (observed around 3 AM) can be influenced by factors such as horizontal transport of cold air from the mountain to the city. Furthermore, as the night progresses the atmosphere becomes more stable and contributes to O_3 accumulation. Finally, in the nighttime and early morning, human activities are reduced. Consequently, the emission of NO_x drops, stabilizing the chemistry of the atmosphere generating no net production or destruction of O_3 (Yusoff *et al.*, 2019).

On the other hand, the meteorological parameters exhibited a low variability during the period of analysis with small standard deviations. The local weather is characterized by two sessional periods of high (Mar. to May. and Sep. to Nov.) and low (Dec. to Feb. and Jun. to Aug.) rainfall (CORPOCALDAS and UNAL, 2020).

3.2 Selected Predictor Variables and Relationships Discussion

3.2.1 Spearman coefficient

Results from the Spearman coefficient are presented in Table 1(a). The best features to forecast $PM_{2.5}$ are CO and SO_2 concentrations; and the best features to forecast O_3 are SO_2 , CO , T , RH , SR , ws , wd_i , and wd_j (see Table 2).

$PM_{2.5}$ concentrations presented a direct relationship with pollutants CO and SO_2 with coefficients higher than +0.3. This suggests the effect of emissions from on-road sources near the monitoring station. Furthermore, $PM_{2.5}$ did not exhibit a significant correlation with any meteorological variable, as coefficients were lower than ± 0.3 . This indicates that $PM_{2.5}$ dynamics are mostly affected by direct emissions and not by meteorological drivers.

O_3 concentrations presented an indirect relationship with other criteria pollutants such as CO

Fig. 2. Measured hourly concentration time series plot, boxplot and hourly average variation of four criteria pollutants in Manizales during study period. The central box represents values within quartiles 25th and 75th, the vertical line extends from the 10th and 90th quartile, the middle solid line represents median, and outliers are plotted as dots.

Table 1. Spearman coefficients among criteria air pollutants and meteorology, and principal component analysis results.

	(a) Spearman correlations				(b) Principal component analysis (PCA)			
	O ₃	PM _{2.5}	SO ₂	CO	PC1	PC2	PC3	PC4
O ₃ (µg m ⁻³)	1.00	-0.28	-0.82	-0.46	0.44	-0.02	0.01	-0.08
PM _{2.5} (µg m ⁻³)	-0.28	1.00	0.42	0.52	-0.13	0.49	-0.18	0.23
SO ₂ (µg m ⁻³)	-0.82	0.42	1.00	0.54	-0.34	0.37	-0.04	-0.02
CO (µg m ⁻³)	-0.46	0.52	0.54	1.00	-0.21	0.51	-0.31	0.06
T (°C)	0.61	0.18	-0.37	0.03	0.37	0.32	-0.03	-0.02
RH (%)	-0.45	-0.07	0.15	-0.02	-0.31	-0.38	-0.15	0.21
SR (W m ⁻²)	0.65	-0.01	-0.41	-0.18	0.36	0.13	-0.01	0.16
Ppt (mm)	0.07	-0.18	-0.18	-0.08	-0.01	-0.27	-0.86	-0.20
ws (m s ⁻¹)	0.39	-0.04	-0.26	-0.11	0.26	0.17	-0.14	-0.68
wd_i (---)	-0.66	0.11	0.56	0.24	-0.36	0.08	0.11	-0.33
wd_j (---)	-0.41	0.12	0.32	0.06	-0.27	-0.02	0.27	-0.51
Eigenvalue	—	—	—	—	4.61	1.91	1.02	0.88
%Variance	—	—	—	—	41.9%	17.3%	9.3%	8.0%
Cumulative%	—	—	—	—	41.9%	59.3%	68.6%	76.6%

Table 2. Summary of predictor variables selected by each method and optimization results.

Pollutant	Method	Predictor variables ^a											SVR optimization		ANNs optimization	
		O ₃	PM _{2.5}	SO ₂	CO	T	RH	SR	Ppt	ws	wd_i	wd_j	ε	c	# Neurons	
O ₃	Spearman coefficients			X	X	X	X	X		X	X	X	0.15	1.4	5	
	PCA			X		X	X	X			X		0.19	2	4	
	Meteorology					X	X	X	X	X	X	X	0.35	1.2	4	
	All available variables	X		X	X	X	X	X	X	X	X	X	0.12	1.2	5	
PM _{2.5}	Spearman coefficients			X	X								0.77	1.5	3 (1 ^b)	
	PCA			X	X	X	X						0.52	1.2	3 (1 ^b)	
	Meteorology					X	X	X	X	X	X	X	0.6	1.5	4 (1 ^b)	
	All available variables	X		X	X	X	X	X	X	X	X	X	0.45	4	5 (1 ^b)	

^a X symbol correspond to selected variable as a predictor.

^b An additional model was developed using only one neuron in the hidden layer to use as a reference.

and SO₂, with coefficients lower than −0.45. These pollutants, as well as NO, are commonly related to vehicular combustion. High values of CO and SO₂ represent high levels of NO which react to consume O₃ (Voukantsis *et al.*, 2011). Other authors have observed similar patterns, in which the primary pollutants such as CO intervene in the photochemical reactions of precursors, decreasing the total amount of O₃ (Sharma *et al.*, 2017).

On the other hand, O₃ showed a direct relationship with air temperature (0.61), solar radiation (0.65), and wind speed (0.39). Solar radiation catalyzes the photochemical reactions, and the air temperature enhances the production of biological volatile organic compounds (Leung *et al.*, 2018). These meteorological variables were also related to the increase of the mixing layer, and consequently, the increase in the dispersion of precursors to be transformed in O₃ by photochemical reactions (Cuesta *et al.*, 2020). Finally, the negative effect of relative humidity and components of wind direction with O₃ stands out, as these parameters are related to wind patterns, cloud cover, and rainfall.

3.2.2 Principal component analysis (PCA)

According to the PCA results (Table 1(b)), the best input variables to forecast PM_{2.5} are SO₂, CO, T, and RH, whereas the best features to forecast O₃ are SO₂, T, RH, SR, and wd_i (see Table 2). It is noteworthy that through PCA, two new features are suggested for forecasting PM_{2.5}, compared to the Spearman correlation approach (T and RH). In the case of O₃, some features were removed (CO, ws, Ppt, wd_j). This result suggests that different significant relationships can be found depending on the statistical approach selected.

A total of 10 PCs were obtained during PCA, from which the first four ones (PC1–PC4) were retained for interpretation due to having eigenvalues greater or close to one, to avoid loss of significant information. Particularly, these four PCs represent 76.6% of the total variation of the original data. Table 1(b) shows the PCs that were retained and the loadings of the original variables within each PC. As in the Spearman correlations approach, the loadings greater than ±0.3 are highlighted in boldface, since they represent the most significant parameters and interactions within each PC.

The first PC (PC1) corresponds to 41.9% of the total data variations. This PC consists mainly of positive contributions of O₃, T, and SR, and negative contributions of SO₂, RH, and the wd_i. PC1 clearly illustrates the influence of O₃ formation mechanisms; a higher downward solar radiation promotes the formation of O₃. Similarly, greater temperatures enhance O₃ formation, by promoting the propagation of the radicals involved in the formation reactions (Voukantsis *et al.*, 2011; Sharma *et al.*, 2017). The interactions among O₃ and SO₂ become an important factor in data variation due to the O₃ consumption by liquid-phase oxidation reactions with SO₂ (Abdul-Wahab *et al.*, 2005).

The second PC (PC2) represents 17.3% of the total data variation. It is composed primarily of positive contributions of PM_{2.5}, CO, SO₂, and temperature; and negative contributions of relative humidity. PC2 illustrates the processes that boost the sink and consumption of SO₂ and PM_{2.5} in

the atmosphere. The relative humidity is a factor related to precipitation which decreases the concentration of $\text{PM}_{2.5}$ by scavenging effects. Moreover, relative humidity is associated with the humidity of particles, the consequent growth of the mass, and faster deposition (Leung *et al.*, 2018).

The positive correlation exhibited between $\text{PM}_{2.5}$ and temperature can be attributed to the fact that higher temperatures enhance the formation of secondary aerosols, due to faster SO_2 and volatile organic compounds (VOC) oxidations reactions (Yang *et al.*, 2017). Finally, the positive correlation among $\text{PM}_{2.5}$, CO, and SO_2 could be attributed to the fact that these pollutants are mostly emitted by a common source, which for the case of Manizales corresponds to the internal combustion vehicles.

For the third and four PCs (PC3 and PC4) which correspond to 9.3% and 8.0% of the data variance respectively, no relationships between O_3 and $\text{PM}_{2.5}$ with other variables were displayed, so these PCs did not add significant information for the scope of this study. To sum up, only PC1 and PC2 convey key information about different relationships between the variables of interest.

3.2.3 Only meteorological variables

In this scenario, T, RH, SR, Ppt, ws, wd_i, and wd_j are used as predictor variables for forecasting O_3 and $\text{PM}_{2.5}$ hourly concentrations. This exercise is implemented as a complement to the methods described previously (Spearman and PCA); in order to assess the possibility for developing air pollutant forecasting models based exclusively on meteorological predictors.

3.2.4 All available variables

$\text{PM}_{2.5}$, SO_2 , CO, T, RH, SR, Ppt, ws, wd_i, and wd_j were used as predictors to forecast O_3 concentrations, whereas O_3 , SO_2 , CO, T, RH, SR, Ppt, ws, wd_i, and wd_j were used as predictors to forecast $\text{PM}_{2.5}$ concentrations. This method is implemented to provide a comparison to the models developed using limited predictor variables, in order to identify the accuracy.

3.3 Development of Forecasting Models

Table 2 presents the optimized parameters of the non-linear regression techniques for each set of predictor variables for O_3 and $\text{PM}_{2.5}$. For SVR, ϵ values obtained ranged from 0.12 to 0.77, and c values range from 1.2 to 4.0. On the other hand, for ANNs, the number of neurons in the hidden layer ranges from 3 to 5. In the ANNs tuning, a larger number of neurons involves more computational time and, sometimes, a tradeoff between computing time and accuracy is required. Furthermore, a larger number of neurons not necessarily will be better, models can be over or under fitted which always needs to be checked (Madhilarasan and Deepa, 2017). Table S3 shows the RMSE, on the training and validation set, for the best models obtained during the cross-validation procedure.

3.4 Evaluation of the Models and Comparison

Table 3 shows the performance statistics obtained on the test set, using the different regression techniques and sets of predictor variables. In general, it was found that all the models developed to predict O_3 concentrations present a satisfactory performance since all of them exhibit R values between 0.84 and 0.94, evidencing a strong linear relationship between the observed and modeled values. It was also found that the SVR model with only meteorological predictors, tends to slightly overestimate the predicted values (MB = 0.16). On other hand, all other models show a tendency to underestimate O_3 concentrations with MB between -0.48 and -3.08. Concerning RMSE, the models present values that vary between 4.63 and 6.35 $\mu\text{g m}^{-3}$, which corresponds to normalized values of 0.40–0.54; these values reveal a good ability of the models to predict ozone hourly concentrations.

Fig. 3(a) shows the Taylor diagram that summarizes the performance statistics obtained in the O_3 regression models. This graph shows the models with greater performance. The best results are obtained when using ANNs, and the best model was obtained when using ANNs and all the available variables as predictors. This model presents the highest R (0.94), a normalized standard deviation close to the original data, and the lowest RMSE (4.63).

Although the best model performance was achieved when using all variables as predictors, the

Table 3. Evaluation and comparison criteria on the test set 1 (Jan. 1 to Jan. 30, 2020) of O₃ and PM_{2.5} for each set of predictor variables and regression technique.

O ₃ ON TEST SET 1, JAN. 2020 (Number of observations = 621)							
Spearman predictors				PCA predictors			
Variable	MLR	SVR	ANNs	Variable	MLR	SVR	ANNs
R	0.91	0.88	0.93	R	0.90	0.84	0.91
MB	−3.05	−1.10	−2.34	MB	−3.08	−0.48	−1.62
RMSE	5.72	5.79	4.96	RMSE	6.01	6.35	5.18
NRMSE	0.49	0.49	0.42	NRMSE	0.51	0.54	0.44
MAE	4.74	4.52	3.92	MAE	4.97	4.88	4.11
Meteorological predictors				All variables			
Variable	MLR	SVR	ANNs	Variable	MLR	SVR	ANNs
R	0.87	0.86	0.90	R	0.91	0.93	0.94
MB	−1.52	0.16	−1.15	MB	−3.02	−1.17	−2.49
RMSE	5.99	5.94	5.38	RMSE	5.66	4.64	4.63
NRMSE	0.51	0.51	0.46	NRMSE	0.48	0.40	0.40
MAE	4.78	4.62	4.20	MAE	4.69	3.78	3.80
PM _{2.5} ON TEST SET 1, JAN. 2020 (Number of observations = 621)							
Spearman predictors				PCA predictors			
Variable	MLR	SVR	ANNs	Variable	MLR	SVR	ANNs
R	0.51	0.48	0.48	R	0.40	0.45	0.36
MB	−0.99	−0.65	−0.67	MB	−0.25	−0.01	1.04
RMSE	4.87	4.85	4.89	RMSE	5.09	5.00	6.23
NRMSE	0.89	0.89	0.89	NRMSE	0.93	0.91	1.14
MAE	3.52	3.60	3.66	MAE	3.79	3.71	4.68
Meteorological predictors				All variables			
Variable	MLR	SVR	ANNs	Variable	MLR	SVR	ANNs
R	0.16	0.27	0.24	R	0.41	0.49	0.43
MB	−1.76	−0.87	−0.36	MB	−0.34	−0.01	0.49
RMSE	5.90	5.66	6.39	RMSE	5.14	5.06	5.45
NRMSE	1.08	1.03	1.17	NRMSE	0.94	0.92	0.99
MAE	4.53	4.34	4.94	MAE	3.88	3.86	4.19

performance obtained when using the features selected according to the Spearman and PCA methods are similar as suggested by the values of R and NRMSE (ANNs-Spearman: 0.93–0.42, ANNs-PCA: 0.91–0.44, ANNs-All variables: 0.94–0.40). This shows the applicability of the methods for selecting features, as it is possible to obtain simpler models, that require fewer variables without significantly affecting the forecasting accuracy.

The models developed from exclusively meteorological variables as predictors to forecast O₃, present slightly lower performance, but still satisfactory, in contrast with those obtained when using all variables as predictors, as shown by the values of R and NRMSE (ANNs-Meteorological variables: 0.90–0.46, ANNs-All variables: 0.94–0.40). Therefore, using only meteorological variables as predictors is also a good choice to forecast O₃ hourly concentrations if other pollutants cannot be measured. This is noteworthy, as meteorological stations are more extensively used compared to air quality stations. Therefore, it would be possible to establish a statistical forecasting model to indicate O₃ concentrations in areas with few information of pollutants. The model could be developed (trained and validated) from a reduced number of O₃ and meteorological measurements in the area of interest, through a monitoring campaign, and continue to be operated using only the continuous measurement of meteorological variables.

The fact that adequate estimations of O₃ concentrations could be made even when the values of the pollutants that interact in chemical reactions to form or consume O₃ are unknown, suggests that the main drivers of O₃ concentrations in the city of Manizales are variations of the different meteorological variables, with a minimum impact in variations of pollutant concentrations of precursors. Similar results were obtained by Zeng *et al.* (2017), which identified that T and RH are

Fig. 3. (a) O₃ and (b) PM_{2.5} Taylor diagram of performance on test data for each predictor variable method. Predictor variables for each scenario are summarized in Table 2.

the dominant factors for O₃ variability in the surface layer of the atmosphere. In addition, it was previously discussed in this study that solar radiation, and indirectly temperature, affect the rates of O₃ production by enhancing the photochemical formation of this pollutant.

On the other hand, for the PM_{2.5} forecasting models, using only meteorological variables as predictors is not an adequate strategy, as all developed models exhibited low R values ranging between 0.16 and 0.27. Moreover, these models present high RMSE values (5.90–6.39), corresponding to NRMSE values from 1.03 to 1.17, which indicate a low prediction ability. This

suggests that $\text{PM}_{2.5}$ concentrations are not strongly influenced by meteorological drivers. Instead, its variations are mainly dependent on emission patterns, indicating that $\text{PM}_{2.5}$ concentrations in the area are caused mainly by primary emissions.

Alternatively, the other $\text{PM}_{2.5}$ developed models present R values that vary between 0.36 and 0.51, showing a medium-strength linear correlation between observed and modeled values. They also have lower deviations with MB ranging between -0.99 and 1.04 .

The best performance statistics for $\text{PM}_{2.5}$ are obtained using the selected predictors from Spearman correlations, with the three regression techniques providing similar results. Overall, the lowest RMSE (4.85) was obtained with the SVR-Spearman model. The predictor variables, in this case, are only pollutants concentrations of SO_2 and CO . This situation shows that the variations in $\text{PM}_{2.5}$ are highly related to SO_2 and CO variations, as explained by the Spearman correlations and the PCA analysis; these are primary pollutants and are directly related with a common source: vehicle emissions. These results show that the methods for selecting features to use as predictors (Spearman and PCA), in the case of O_3 for example, are also useful for developing $\text{PM}_{2.5}$ forecasting models. As it is possible to identify the most relevant predictor variables and exclude features that could create noise in the data or lead to a more complex model with a reduced ability of generalization.

In fact, it is striking that the performance of predicting $\text{PM}_{2.5}$ decreases when adding more variables, such as the case of the models using all the variables as predictors. These results are caused by the major complexity of the models, leading to overtraining. Therefore, these models presented a better performance on training sets but lost the ability to generalize in different situations on the test data. As previously discussed, $\text{PM}_{2.5}$ concentrations have stronger associations with emission patterns throughout the day and do not exhibit a strong sensitivity to the variation of meteorological conditions.

Additionally, when observing Fig. 3(b), by comparing regression techniques, the worst performances are obtained when using ANNs. This could be caused by the risk minimization principle applied by ANNs. As discussed by Mogollón-Sotelo *et al.* (2021), that causes ANNs to better fit the training data. However, the training data has noise, then affecting the generalization ability of the models.

The lower results obtained when trying to predict $\text{PM}_{2.5}$ compared to O_3 might be associated with the complex pattern of this pollutant, which makes it necessary to include other types of predictors in the models to improve the estimates (e.g., Concentration of PM_{10} and NO_2 , PBL, temporal and spatial disaggregated emission inventories).

Figs. S3 and S4 present the scatterplots that relate the observed concentrations and the forecasted values of O_3 and $\text{PM}_{2.5}$, respectively. In the case of O_3 , the data present a good linear relationship throughout the range of studied concentrations, a fact that is also supported by the statistics summarized in Table 3.

In contrast, when observing the $\text{PM}_{2.5}$ scatterplots, the linear relationship is not so clear. It is also striking that all models presented difficulties in representing the peaks of maximum concentration since, in general terms, these models are not capable of predicting concentrations greater than $20 \mu\text{g m}^{-3}$. The episodes in which concentrations exceeded $20 \mu\text{g m}^{-3}$ can be attributed to specific situations, which are not derived from variations in the meteorology or changes in the concentration of CO and SO_2 . For instance, higher values of $\text{PM}_{2.5}$ could be attributed to unusual scenarios in which emissions increase, such as an activity related to the industries near the air quality monitoring station or a possible natural ash emission from the nearby volcano. Finally, it is noteworthy that the ANNs-Spearman models for $\text{PM}_{2.5}$ exhibit predictions with negative concentration values, which have no physical meaning; these values can be explained due to overtraining of the neural networks.

Fig. 4(a) shows the hourly average profile of O_3 during the test period for the observed and modeled values. Similarly, Fig. 5 presents the time series of hourly concentrations for the period from January 6 to January 13. The figures showed that all the models accurately represent maximum concentration peaks around noon and early-morning.

Similarly, Fig. 4(b) shows the hourly average profile of $\text{PM}_{2.5}$ for the observed and modeled values, and Fig. 6 presents the hourly time series, again for the period from January 6 to 13. Figures showed that the pattern of $\text{PM}_{2.5}$ is well represented in most of the models, except for the one developed using ANNs-PCA, in which the decrease in concentrations during non-peak

Fig. 4. Hourly average variation of (a) O_3 and (b) $PM_{2.5}$ for the different models and sets of predictor variables. Predictor variables for each scenario are summarized in Table 2.

vehicle hours is largely underestimated. The one developed using MLR–Meteorology, which does not follow the desired profile during any period of the day.

In addition, Fig. 6 showed that none of the models can forecast the maximum concentration peaks, and the negative values estimated through the ANNs models are once again evident. For the sake of comparison, the ANN technique was also developed using only one neuron in the hidden layer, (denoted as ANN(1) in Fig. 6). This design aimed to evaluate the possible overtraining effect on the neural networks when using a larger number of neurons. It was found that the performance achieved when using a single neuron in the hidden layer is superior to that of the neural network with a higher number of neurons, and it is similar to the one obtained in the models developed with the MLR technique. To sum up, for the case of $PM_{2.5}$, the simplest models offer the best predictions.

Other studies have developed statistical forecast models for hourly $PM_{2.5}$ concentrations. For instance, Francheschi *et al.* (2018) used ANNs to forecast $PM_{2.5}$ concentrations at two points in the city of Bogota, Colombia, obtaining performance statistics such as an R of 0.51–0.68 and RMSE

Fig. 5. O_3 time series for predicted and observed values ranging from January 6 to January 13, 2020. Predictor variables for each scenario are summarized in Table 2.

of $5.79\text{--}7.87\ \mu\text{g m}^{-3}$. Similarly, Sekar *et al.* (2015a), developed different forecast models for the city of Delhi, India, obtaining performance statistics such as R of $0.47\text{--}0.89$ and RMSE of $56.20\text{--}103.48\ \mu\text{g m}^{-3}$. Compared to these results, the models obtained through the present study failed to obtain R as high as these studies, however, the RMSE values were improved. On the other hand, Sekar *et al.* (2015b) also studied the forecast of hourly concentrations of O_3 in Delhi, India, obtaining performance statistics such as an R of $0.51\text{--}0.82$ and RMSE of $18.31\text{--}25.51\ \mu\text{g m}^{-3}$. In this case, the performance statistics obtained in this study were superior, indicating models with enhanced predictive capabilities.

Fig. 6. $\text{PM}_{2.5}$ time series for predicted and observed values ranging from January 6 to January 13, 2020. Predictor variables for each scenario are summarized in Table 2.

It should be noted that the comparisons made with the models developed in other regions are merely indicative, since different predictors, statistical models, and different procedures were used in each study.

3.5 Case Study: Evaluation of the Best Models in a Different Period

To assess the generalization ability of those models that presented the best performance on test set 1 (Jan 1 to Jan 30, 2020), a new test was defined and named as set 2 (May 1 to May 31, 2020). Those sets are defined over a different chronological period, with different meteorological patterns and emission scenarios. For instance, test set 2 was dominated by low emission patterns

Table 4. Evaluation and comparison criteria on the test set 2 (May 1 to May 31, 2020) of O₃ and PM_{2.5} for best configuration obtained on test set 1 (Jan 1 to Jan 30, 2020).

ON TEST SET 2, May 2020 (Number of observations = 721)			
O ₃ with all variables as predictors		PM _{2.5} with Spearman predictors	
Variable	ANNs	Variable	SVR
R	0.90	R	0.28
MB	−1.39	MB	2.33
RMSE	4.41	RMSE	4.91
NRMSE	0.47	NRMSE	4.66
MAE	3.54	MAE	3.91

due to the pandemic scenario where mobility restrictions were adopted. Moreover, May was wetter (83.4%) than January (77.1%) and have the lowest solar radiation (269 W m^{−2} compared to 364.8 W m^{−2}).

Table 4 shows the performance statistics obtained on test set 2. For the O₃ case (ANNs-All variables), predictions present a satisfactory performance with an R value of 0.9, similar to the performance obtained for test set 1. This configuration underestimates O₃ concentrations with MB value of −1.39, and the RMSE (4.41) and NRMSE (0.47) ratified the good ability of the model to predict O₃ hourly concentration even when the scenario is different. These results suggest that ANNs model with all variables as predictors are suitable to provide initial concentration estimates and even predict or fulfill O₃ hourly concentrations on short periods of interest.

The inclusion of all variables in the model improves the prediction accuracy due to the compensation on variables by the different phenomena of the atmosphere. For instance, Abdul-Wahab and Al-Alawi (2002) exposed that meteorology has a contribution in ANNs model in a range of 33.1 to 40.6%, and the inclusion of pollutants such as SO₂ and CO can improve the model performance between 6 to 10 percent due to it is participation in the photochemical formation of O₃. Other authors such as Zhen *et al.* (2017) include physical phenomena such as stratospheric/tropospheric exchanges, dominated by T and RH as a variable to improve the model performance.

For the case of PM_{2.5} (SVM-Spearman predictor), predictions present unsatisfactory performance with an R value of 0.28 and high RMSE (4.91), corresponding to a normalized value of 4.66. Moreover, the configuration selected overestimate PM_{2.5} concentrations with MB value of 2.33. These results indicated the low predictive ability of the model. Similar overestimated concentrations with SVM models were obtained by Mogollón-Sotelo *et al.* (2021) in Bogota, Colombia, where SVM errors were related to the quality of data, such as the quality of training and validation data and its relationship with testing data. Moreover, these models have a limited ability to represent the sudden changes, mainly affected by emission patterns and generating an accumulative error in forecast values of the model.

The main explanation for the lower results of the model on test set 2 could be attributed to the fact that PM_{2.5} concentrations in the study area are mainly emitted by primary emissions. These emissions change drastically for test set 2, due to the mobility restrictions adopted as a response to the pandemic scenario, causing a reduction of emissions (Corpocaldas and UNAL, 2020).

4 CONCLUSIONS

Results obtained show that the linear and nonlinear models represent the patterns of hourly O₃ concentrations, offering the best performance when using the ANNs regression technique in O₃ forecasting. The best model for O₃ prediction involved the use of all available variables. Models with variables selected by Spearman coefficients and PCA offer similar performance. Hence, these techniques identified the key predictor variables, reducing the number of variables monitored, without compromising forecasting accuracy. On the other hand, by using only meteorological variables as predictors it is possible to obtain an adequate estimate of O₃ concentrations with R values higher than 0.8. These results suggest that, for the case of Manizales, variation in O₃ concentrations is mostly dominated by meteorological variables such as temperature and solar radiation, and is less sensitive to variations of the measured pollutants (PM_{2.5}, SO₂, and CO). This

suggests that it could be possible to develop a forecasting model for O₃ concentrations in areas where only meteorological stations are present.

On the other hand, forecasting PM_{2.5} with meteorological and primary pollutant data was less successful compared to O₃. However, the models developed here show correlation coefficients similar to those obtained in other studies and better RMSE values. The models developed using ANNs were less effective for PM_{2.5} forecasting, which is associated with a possible overtraining of the neural network reducing the models generalization capacity. In contrast, the models developed using MLR and SVR, show better performances.

Regarding the selection of predictor variables, the use of only meteorological variables does not represent the variation of hourly concentrations of PM_{2.5}, because this pollutant is associated with on-road sources emissions, and thus emission patterns dominated the variations of PM_{2.5} concentration in Manizales. The best PM_{2.5} model used only SO₂ and CO concentrations in Manizales as predictors, as indicated by the Spearman correlation analysis. This suggests emissions from common vehicular sources. Additional meteorological variables did not improve model forecasting of PM_{2.5}. Finally, none of the models could predict concentrations higher than 20 µg m⁻³; hence, the peak PM_{2.5} concentrations were underestimated.

The models developed have potential for making statistical forecasts of O₃, demonstrating strong associations to meteorology, which can vary according to the study area. In addition, the use of hourly data, instead of daily averages, allows studying the changes in these relationships as a function of the period of the day, and its corresponding changes in emission patterns.

The methodology developed indicates that it is possible to select a subset of variables that are suitable for regression models by means of a preliminary exploratory analysis, such as the computation of the Spearman coefficients or through principal component analysis. The methodology proposed in this study might serve to other urban regions for implementing air quality forecasting models, not only as a tool for a better comprehension of pollutant dynamics, but also for contributing with air quality management.

DISCLAIMER

The authors declare no conflict of interest and no competing financial relationships that could influence data and results reported in this paper.

ACKNOWLEDGMENTS

The authors acknowledge the financial support given by Corporación Autónoma Regional de Caldas (CORPOCALDAS) and Universidad Nacional de Colombia Sede Manizales through Convenio Interadministrativo de Asociación 114-2018 and 176-2020.

SUPPLEMENTARY MATERIAL

Supplementary material for this article can be found in the online version at <https://doi.org/10.4209/aaqr.200471>

REFERENCES

- Abdul-Wahab, S.A., Al-Alawi, S.M. (2002). Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. *Environ. Modell. Software* 17, 219–228. [https://doi.org/10.1016/S1364-8152\(01\)00077-9](https://doi.org/10.1016/S1364-8152(01)00077-9)
- Abdul-Wahab, S.A., Bakheit, C.S., Al-Alawi, S.M. (2005). Principal component and multiple regression analysis in modelling of ground-level ozone and factors affecting its concentrations. *Environ. Modell. Software* 20, 1263–1271. <https://doi.org/10.1016/j.envsoft.2004.09.001>
- Binaku, K., Schmeling, M. (2017). Multivariate statistical analyses of air pollutants and meteorology in Chicago during summers 2010-2012. *Air Qual. Atmos. Health* 10, 1227–1236. <https://doi.org/10.1007/s11869-017-0507-7>

- Corporación Autónoma Regional de Caldas (CORPOCALDAS) and Universidad Nacional de Colombia Sede Manizales (UNAL) (2019). Actualización del inventario de emisiones atmosféricas en la ciudad de Manizales año base 2017. In Spanish. Manizales. http://cdiac.manizales.unal.edu.co/inicio/publicaciones/Boletín_inventario_de_emisiones_atmosféricas_Manizales_año_base_2017.pdf (accessed 29 February 2020).
- Corporación Autónoma Regional de Caldas (CORPOCALDAS) and Universidad Nacional de Colombia Sede Manizales (UNAL) (2020). Boletín calidad del aire Manizales - año 2020. SIMAC. In Spanish. Manizales. http://cdiac.manizales.unal.edu.co/inicio/publicaciones/Bolet%C3%ADn_CA_Manizales_a%C3%B1o_2020.pdf (accessed 29 March 2021).
- Cuesta, A.P., Wahl, M., Acosta, J.G., García Reynoso, J.A., Aristizábal, B.H. (2020). Mixing layer height and slope wind oscillation: Factors that control ambient air SO₂ in a tropical mountain city. *Sustain. Cities Soc.* 52, 101852. <https://doi.org/10.1016/j.scs.2019.101852>
- Departamento Nacional de Estadísticas (DANE) (2019). Resultados Censo Nacional de Población y Vivienda 2018 Manizales, Caldas. In Spanish. www.dane.gov.co (accessed 1 February 2020).
- Dimitriou, K. (2016). Upgrading the estimation of daily PM₁₀ concentrations utilizing prediction variables reflecting atmospheric processes. *Aerosol Air Qual. Res.* 16, 2245–2254. <https://doi.org/10.4209/aaqr.2016.05.0214>
- Franceschi, F., Cobo, M., Figueredo, M. (2018). Discovering relationships and forecasting PM₁₀ and PM_{2.5} concentrations in Bogotá Colombia, using Artificial Neural Networks, Principal Component Analysis, and k-means clustering. *Atmos. Pollut. Res.* 9, 912–922. <https://doi.org/10.1016/j.apr.2018.02.006>
- González, C.M., Ynoue, R.Y., Vara-Vela, A., Rojas, N.Y., Aristizábal, B.H. (2018). High-resolution air quality modeling in a medium-sized city in the tropical Andes: Assessment of local and global emissions in understanding ozone and PM₁₀ dynamics. *Atmos. Pollut. Res.* 9, 934–948. <https://doi.org/10.1016/j.apr.2018.03.003>
- Guo, Q., He, Z., Li, S., Li, X., Meng, J., Hou, Z., Liu, J., Chen, Y. (2020). Air pollution forecasting using artificial and wavelet neural networks with meteorological conditions. *Aerosol Air Qual. Res.* 20, 1429–1439. <https://doi.org/10.4209/aaqr.2020.03.0097>
- Hoshyaripour, G., Brasseur, G., Andrade, M.F., Gavidia-Calderón, M., Bouarar, I., Ynoue, R.Y. (2016). Prediction of ground-level ozone concentration in São Paulo, Brazil: Deterministic versus statistic models. *Atmos. Environ.* 145, 365–375. <https://doi.org/10.1016/j.atmosenv.2016.09.061>
- Jia, M., Cheng, X., Zhao, T., Yin, C., Zhang, X., Wu, X., Wang, L., Zhang, R. (2019). Regional air quality forecast using a machine learning method and the WRF Model over the Yangtze River Delta, East China. *Aerosol Air Qual. Res.* 19, 1602–1613. <https://doi.org/10.4209/aaqr.2019.05.0275>
- Jiang, Y., Yang, C., Na, J., Li, G., Li, Y., Zhong, J. (2017). A brief review of neural networks based learning and control and their applications for robots. *Complexity*. 2017, 1895897. <https://doi.org/10.1155/2017/1895897>
- Jolliffe, I.T., Cadima, J. (2016). Principal component analysis: A review and recent developments. *Phil. Trans. R. Soc. A* 374, 20150202. <https://doi.org/10.1098/rsta.2015.0202>
- Karimian, H., Li, Q., Wu, C., Qi, Y., Mo, Y., Chen, G., Zhang, X., Sachdeva, S. (2019). Evaluation of different machine learning approaches to forecasting PM_{2.5} mass concentrations. *Aerosol Air Qual. Res.* 19, 1400–1410. <https://doi.org/10.4209/aaqr.2018.12.0450>
- Kecman, V. (2001). *Learning and soft computing: Support Vector Machines, Neuronal Networks, and Fuzzy Logic Models*, MIT Press, London.
- Khedairia, S., Khadir, M.T. (2012). Impact of clustered meteorological parameters on air pollutants concentrations in the region of Annaba, Algeria. *Atmos. Res.* 113, 89–101. <https://doi.org/10.1016/j.atmosres.2012.05.002>
- Kisi, O., Parmar, K.S., Soni, K., Demir, V. (2017). Modeling of air pollutants using least square support vector regression, multivariate adaptive regression spline, and M5 model tree models. *Air Qual. Atmos. Health* 10, 873–883. <https://doi.org/10.1007/s11869-017-0477-9>
- Leung, D.M., Tai, A.P.K., Mickley, L.J., Moch, J.M., Van Donkelaar, A., Shen, L., Martin, R.V. (2018). Synoptic meteorological modes of variability for fine particulate matter (PM_{2.5}) air quality in major metropolitan regions of China. *Atmos. Chem. Phys.* 18, 6733–6748. <https://doi.org/10.5194/acp-18-6733-2018>
- Luna, A.S., Paredes, M.L.L., de Oliveira, G.C.G., Corrêa, S.M. (2014). Prediction of ozone concentration

- in tropospheric levels using artificial neural networks and support vector machine at Rio de Janeiro, Brazil. *Atmos. Environ.* 98, 98–104. <https://doi.org/10.1016/j.atmosenv.2014.08.060>
- Mehdipour, V., Stevenson, D.S., Memarianfard, M., Sihag, P. (2018). Comparing different methods for statistical modeling of particulate matter in Tehran, Iran. *Air Qual. Atmos. Health* 11, 1155–1165. <https://doi.org/10.1007/s11869-018-0615-z>
- Mogollón-Sotelo, C., Casallas, A., Vidal, S., Celis, N., Ferro, C., Belalcazar, L. (2021). A support vector machine model to forecast ground-level PM_{2.5} in a highly polluted city with a complex terrain. *Air Qual. Atmos. Health* 14, 399–409. <https://doi.org/10.1007/s11869-020-00945-0>
- Schober, P., Boer, C., Schwarte, L. (2018). Correlation coefficients: Appropriate use and interpretation. *Anesthesia Analgesia* 126, 1763–1768. <https://doi.org/10.1213/ANE.0000000000002864>
- Sekar, C., Gurjar, B.R., Ojha, C.S.P., Goyal, M.K. (2015a). Potential assessment of neural network and decision tree algorithms for forecasting ambient PM_{2.5} and co concentrations: Case study. *J. Hazard. Toxic Radioact. Waste* 20, A5015001 [https://doi.org/10.1061/\(ASCE\)HZ.2153-5515.0000276](https://doi.org/10.1061/(ASCE)HZ.2153-5515.0000276)
- Sekar, C., Ojha, C.S.P., Gurjar, B.R., Goyal, M.K. (2015b). Modeling and prediction of hourly ambient ozone (O₃) and oxides of nitrogen (NO_x) concentrations using artificial neural network and decision tree algorithms for an urban intersection in India. *J. Hazard. Toxic Radioact. Waste* 20, A4015001 [https://doi.org/10.1061/\(ASCE\)HZ.2153-5515.0000270](https://doi.org/10.1061/(ASCE)HZ.2153-5515.0000270)
- Sharma, A., Mandal, T.K., Sharma, S.K., Shukla, D.K., Singh, S. (2017). Relationships of surface ozone with its precursors, particulate matter and meteorology over Delhi. *J. Atmos. Chem.* 74, 451–474. <https://doi.org/10.1007/s10874-016-9351-7>
- Sihag, P., Kumar, V., Afghan, F.R., Pandhiani, S.M., Keshavarzi, A. (2019). Predictive modeling of PM_{2.5} using soft computing techniques: Case study—Faridabad, Haryana, India. *Air Qual. Atmos. Health* 12, 1511–1520. <https://doi.org/10.1007/s11869-019-00755-z>
- Uyanık, G.K., Güler, N. (2013). A study on multiple linear regression analysis. *Procedia Soc. Behav. Sci.* 106, 234–240. <https://doi.org/10.1016/j.sbspro.2013.12.027>
- Vlachogianni, A., Kassomenos, P., Karppinen, A., Karakitsios, S., Kukkonen, J. (2011). Evaluation of a multiple regression model for the forecasting of the concentrations of NO_x and PM₁₀ in Athens and Helsinki. *Sci. Total Environ.* 409, 1559–1571. <https://doi.org/10.1016/j.scitotenv.2010.12.040>
- Voukantsis, D., Karatzas, K., Kukkonen, J., Räsänen, T., Karppinen, A., Kolehmainen, M. (2011). Intercomparison of air quality data using principal component analysis, and forecasting of PM₁₀ and PM_{2.5} concentrations using artificial neural networks, in Thessaloniki and Helsinki. *Sci. Total Environ.* 409, 1266–1276. <https://doi.org/10.1016/j.scitotenv.2010.12.039>
- Weizhen, H., Zhengqiang, L., Yuhuan, Z., Hua, X., Ying, Z., Kaitao, L., Yan, M. (2014). Using support vector regression to predict PM₁₀ and PM_{2.5}. *IOP Conf. Ser.: Earth Environ. Sci.* 17, 012268. <https://doi.org/10.1088/1755-1315/17/1/012268>
- Wilks, D.S. (2019). *Statistical methods in the atmospheric sciences*, Chapter 13 - Principal Component (EOF) Analysis (Fourth edition), Elsevier. <https://doi.org/10.1016/B978-0-12-815823-4.00001-8>
- World Meteorological organization (WMO) (2020). The gaps in the Global Basic Observing Network (GBON). https://library.wmo.int/index.php?lvl=notice_display&id=21771#.YFtxGa9KhPa (accessed March 24 2020).
- Yang, Q., Yuan, Q., Li, T., Shen, H., Zhang, L. (2017). The relationships between PM_{2.5} and meteorological factors in China: Seasonal and regional variations. *Int. J. Environ. Res. Public Health* 14, 1510. <https://doi.org/10.3390/ijerph14121510>
- Yusoff, M.F., Latif, M.T., Juneng, L., Khan, M F., Ahamad, F., Chung, J.X., Mohtar, A.A.A. (2019). Spatio-temporal assessment of nocturnal surface ozone in Malaysia. *Atmos. Environ.* 207, 105–116. <https://doi.org/10.1016/j.atmosenv.2019.03.023>
- Zeng, G., Morgenstern, O., Shiona, H., Thomas, A.J., Querel, R.R., Nichol, S.E. (2017). Attribution of recent ozone changes in the Southern Hemisphere mid-latitudes using statistical analysis and chemistry–climate model simulations. *Atmos. Chem. Phys.* 17, 10495–10513. <https://doi.org/10.5194/acp-17-10495-2017>
- Zeng, Y., Jaffe, D.A., Qiao, X., Miao, Y., Tang, Y. (2020). Prediction of potential high PM_{2.5} concentrations in Chengdu, China. *Aerosol Air Qual. Res.* 20, 956–965. <https://doi.org/10.4209/aaqr.2019.11.0586>

Zhang, H., Zhang, W., Palazoglu, A., Sun, W. (2012). Prediction of ozone levels using a hidden Markov model (HMM) with Gamma distribution. *Atmos. Environ.* 62, 64–73. <https://doi.org/10.1016/j.atmosenv.2012.08.008>