Nature Environment and Pollution Technology
An International Quarterly Scientific Journal

p-ISSN: 0972-6268
(Print copies up to 2016)
e-ISSN: 2395-3454

Vol. 23 | No. 2 | pp. 711-724 | 2024

Original Research Paper

https://doi.org/10.46488/NEPT.2024.v23i02.009

Open Access Journal

Application of Random Forest in a Predictive Model of PM_{10} Particles in Mexico City

Alfredo Ricardo Zárate Valencia*† and Antonio Alfonso Rodríguez Rosales**
*Doctorado en Ciencias Ambientales, Universidad Autónoma de Guerrero, 39070, México
**Instituto de Ciencias Aplicadas y Tecnología, Universidad Nacional Autónoma de México, 04510, México
†Corresponding author: Alfredo Ricardo Zárate Valencia; azaratev@hotmail.com

Nat. Env. & Poll. Tech.
Website: www.neptjournal.com

Received: 09-08-2023
Revised: 14-10-2023
Accepted: 16-10-2023

Key Words:
Predictive model
Air pollution
Random forest
Monitoring
PM_{10} particles

ABSTRACT

Over time, predictive models tend to become more accurate but also more complex, thus achieving better predictive accuracy. When the data is improved by increasing its quantity and availability, the models are also better, which implies that the data must be processed to filter and adapt it for initial analysis and then modeling. This work aims to apply the Random Forest model to predict PM_{10} particles. For this purpose, data were obtained from environmental monitoring stations in Mexico City, which operates 29 stations of which 12 belong to the State of Mexico. The pollutants analyzed were CO carbon monoxide, NO nitrogen oxide, and PM_{10} particulate matter equal to or less than 10 $\mu g.m^{-3}$, NOx nitrogen oxide, NO_2 nitrogen dioxide, SO_2 sulfur dioxide, O_3 ozone, and $PM_{2.5}$ particulate matter equal to or less than 2.5 $\mu g.m^{-3}$. The result was that when calculating the certainty of our model, we have a value of 80.40% when calculating the deviation from the mean, using 15 reference variables.

INTRODUCTION

Air pollution is a global problem. The World Health Organization estimates that 90% of people breathe polluted air (WHO 2019). Consequently, around seven million deaths are attributed to air pollution (WHO 2014). Several organizations are focusing on evaluating pollution indicators, considering the characteristics of each place (Perevochtchikova 2013).

One of the primary air pollutants is ozone (O_3), a gas of three oxygen atoms in the upper atmosphere and at the surface level. At the surface level, the latter becomes hazardous to people's health (Liu et al. 2018). For example, industrial and vehicle emissions are considered the primary precursor sources of ozone (EPA 2021). It should be noted that wind is a significant factor in pollutant dispersion; its speed and direction are linked to pollutant concentration, and the more wind, the less pollutant concentration (Biancofiore et al. 2015).

Poor air quality is a public health problem for large cities with high population concentrations, and these air pollutant emissions are generated by motor vehicles and factories (Perevochtchikova 2009). Our case study was conducted in Mexico City, Latin America's largest city. For this reason, the Federal and Local governments have taken care to have monitoring stations and record in databases all records of the primary pollutants on average per hour, through the guidelines and supervision of the INECC (National Institute of Ecology and Climate Change) with the operation of the DMA (Directorate of Atmospheric Monitoring) of Mexico City. Under this premise, the data generated can be used for monitoring and data science analysis to make inferences about the behavior of pollutants, which is the case here.

The data recorded in different industries and applications such as monitoring (in just one year, the DMA is more than 6 million records), which is why they have to be analyzed to look for behaviors and opportunities to improve processes and inferences, for which it is necessary to have computational tools of Data Science; otherwise, it would be almost impossible (Provost & Fawcett 2013). Big Data is about analyzing vast amounts of data, and we can divide it into two parts: the technology (Hadoop, Spark, etc.) and the platform architecture (Gonzalez Diaz 2017). With Big Data tools, any analysis would take a long time, and the relevance would be adequate.

The application of Data Science in environmental and specifically air quality monitoring has been worked on, e.g.,

Kurt and Oktay applied Airpol, a real-time forecasting system using PHP programming language, MySQL database, and MATLAB for the inferential statistics part (Kurt & Oktay 2010). It acquires data from different sources and stores them in a single central database. In retrospect, systems for predicting pollutant behaviors were treated as black boxes, where the system was not understood. It was challenging to understand the relationships of the variables and thus not know why things happened (Mohan & Kandya 2011); nowadays, analysts need to understand these relationships.

Sertel et al. (2012) integrated remote sensing, geostatistical, and spatial analysis methods. They looked at the relationship between transport, land use, and air quality (Sertel et al. 2012), i.e., they studied data from different sources to improve their results. Essentially, four types of models are used to predict $PM_{2.5}$ concentrations: regression, artificial intelligence, time series, and chemical transport (Zhou et al. 2014). Zhou et al. (2014) used EMD (Empirical Mode Decomposition), which was initially developed to study ocean waves but is now also applied in nature and social science studies.

Zhang & Yuan (2015) implemented a distributed random forest algorithm using Spark to create a predictive air quality model using actual meteorological data in Beijing (Zhang & Yuan 2015). In the same year, Anaya Díaz (2015) took meteorological data and an air quality index collected over four years in Valledupar, Colombia (Anaya Díaz 2015), using clustering techniques to estimate air quality with the application of data mining techniques. Hsieh et al. proposed an entropy-minimizing model to suggest locations for new monitoring stations using air quality data in Beijing by constructing i-layer graphs that reflect temporal correlations. In contrast, the data connections remain identical (Hsieh et al. 2015).

Data mining allows to analysis of air quality using analytical methods when scientific methods do not exist; in that sense, Soh et al. propose a predictive system for air quality using ST-DNN (Shape-Tailored Deep Neural Networks) to predict PM2.5 48 hours in advance (Soh et al. 2016). Also, Li et al. (2016) applied to predict air quality deep learning approach, applying a regional data treatment as a spatiotemporal process that considers spatial and temporal correlations of data to predict the air quality of all monitoring stations simultaneously with seasonal stability (Li et al. 2016), which recognizes and applies seasonal behavior in analysis and predictions.

Wang et al. (2017) categorized the main pollutant forecasting models as deterministic, statistical, and hybrid models (Wang et al. 2017). They proposed a hybrid model based on the ELM (Extreme Learning Machines) model optimized by the DE (Differential Evolution) algorithm, managing to forecast random, irregular, and non-stationary data series. Alongside model improvements, Bellinger et al. recognize that advances in technology and lower prices for computing power in computers allow for measuring and storing different variables related to the environment (Bellinger et al. 2017), coupled with data from resources such as social networks, give a new perspective to environmental health analysis. Also, in 2017, Mahajan set out to develop a stable model with a real-time implementation (Mahajan et al. 2017), for which they used an ARIMA (Autoregressive Integrated Moving Average) model and an NNAR (Neural Network Autoregression) model to predict PM2.5 levels in four regions of Taiwan with a total of 557 monitoring stations. In a more recent publication, Soh et al. propose this time to analyze the spatiotemporal patterns of particulate matter (PM) in Taiwan by developing a PM probability map with its patterns per day. It is a method that uses dynamic time warping and analyses the temporal similarity between multiple stations and their performance (Soh et al. 2017).

Zhu (2018) highlighted the importance of regional meteorological conditions as essential data to be considered in a predictive model for pollutants such as O_3, as they found that in Chicago, the concentration of that pollutant is more sensitive to air temperature, wind speed, and direction, relative humidity, incoming solar radiation and cloudiness (Zhu et al. 2018). Gao also conducts model studies to predict the behavior of O_3 (most studies are based on PM). They use a few climate parameters as predictors, and the application of the Monte Carlo method to study the uncertainty of the ANN (Artificial Neural Networks) model is highlighted (Gao et al. 2018). In a study, Amado and de la Cruz (2018) built predictive models that relate the values of a prototype that makes sensor readings and relates them to the air quality index (Amado & De la Cruz 2018), together with Bayesian models, allowing them to obtain an accuracy of up to 99 % in their tests.

In an article published by Rybarczyk & Zalakeviciute (2018), they infer that forecasting is less accurate than estimation and justify the use of more demanding methods, such as Deep Learning, to predict hours or even days in advance the concentration of a pollutant (Rybarczyk & Zalakeviciute 2018). Based on the above, it seems that data analysis for predicting pollutant concentrations has been applied sparingly. Still, one also has to consider the application of the model to be developed and the experience of the implementers. In the same period, when evaluating various models, Roy et al. (2018) suggest that the Multivariate Adaptive Regression Splines (MARS) model has a better description of the data set and a higher prediction

compared to the Random Forest (RF) and the Classification and Regression Tree (CRT) (Roy et al. 2018).

Regarding accuracy again, Zhang et al. (2019) noted that until 2019, pollutant concentration prediction methods needed to effectively use existing Big Data to extract the temporal and statistical characteristics of the data (Zhang et al. 2019), resulting in limited model accuracy. Another article highlights that it is highly recommended to compare different models, such as ANN, PLS (Partial Least Squares), RFR, and MLR, when choosing which one would be the most suitable (Wei et al. 2019) and obtain the best model by averaging the results of different models. Having contributed to Big Data, more and more governments are publishing their data on their portals to stimulate its use in applications that serve their citizens (Buenadicha et al. 2019). Commercial industries are also embracing Big Data by using innovative techniques that have extracted demographic, socioeconomic, and consumer behavioral data for forecasting and analysis (Tan et al. 2020). Lim and his team presented a complementary option, mobile sampling, that improves the spatial granularity of Land Use Regression (LUR) models by deploying low-cost sensors that could improve and modernize how air pollution is measured (Lim et al. 2019). Also, Yuchi et al. (2019) showed that when measuring indoor $PM_{2.5}$ concentrations, MLR (multiple linear regression) and RFR (random forest regression) models obtain similar results in a heavily polluted environment by using a small number of variables (Yuchi et al. 2019).

We will now look at two proposals for predicting pollutant concentrations using Air Quality Indices (AQI) (which use a combination of measured pollutants for their integration). The first is presented by Lino-Ramirez et al., which is a real-time system that monitors environmental variables from several points and makes a prediction of the behavior of those variables (Lino-Ramirez et al. 2019). It was applied in Guanajuato, Mexico, and predicts air quality according to an established traffic light. In the second case, three learning models were applied to predict the $PM_{2.5}$ air level using data from the CPCB (Central Pollution Control Board) using more than three years of data resulting in an air quality index for the Delhi NCR region (Sihag et al. 2019). Camí Núñez highlights in his article that applied measurements such as the two previous ones are used to trigger different anti-pollution protocols and should be able to forecast pollutant concentrations sometime in advance to be timely (Camí Núñez 2020).

Next, we want to point out that, over time, predictive models tend to become more accurate but also more complex, indeed, based on the experience of previous model developments. Zhou et al. (2020), for example, in his model made measurements at different seasons in four cities in the Yangtze River Delta with five prevalent forecasting tools, including SFGM (Seasonal Fractional-order Grey Model), SGM (Seasonal Grey Model), LSSVM (Least Squares Model) and LSSVM (Least Squares Squared Model), and LSSVM (Least Squares Squared Model), LSSVM (Least Squares Support Vector Machine), SARIMA (Seasonal Auto-regress Integrated Moving Average) and BPNN (Back Propagation Neural Network), with a considerable improvement of the prediction accuracy of seasonal air quality changes (Zhou et al. 2020). By increasing and thus enriching the data available for modeling, Pinto et al. (2020) proposed introducing data on traffic patterns (speed, intensity) and emissions generated by vehicle emission models (Pinto et al. 2020).

Another study on Delhi NCR done by Yadava & Agarwal (2020) to predict the level of $PM_{2.5}$ in the air applied three models: LSTM (Long Short-Term Memory), Auto-regression, and SVM (Vector Machine) to test which of the models is the most suitable, using the data from the CPCB (Central Pollution Control Board) website (Yadava & Agarwal 2020).

Already Harbola et al. (2021) introduced the Air Quality Temporal Analyser (AQTA), which is a system with visual analysis of air quality data that allows the use of visualization techniques that visually display anomalies and correlations of correlated trends through an interactive, non-directed search (Harbola et al. 2021).

Neural networks rely on data, which generally comes from ground-based monitoring stations, so their coverage is limited to the number of stations installed. To overcome this limitation, Lightstone et al., through a deep neural network (DNN), combine spatial Kriging interpolation with additional local source variables by interpolating the measured $PM_{2.5}$ concentrations across locations without monitoring stations installed (Lightstone et al. 2021).

MATERIALS AND METHODS

For our research, which is based on developing a predictive model, we have followed the steps suggested by some authors, such as (Herman et al. 2013), who propose four stages for model development (acquire, prepare, analyze, and act) and we complement it with what is suggested by (Provost & Fawcett 2013), who propose six stages (understanding the business, understanding the data, data preparation, modeling, evaluation and model development).

Fig. 1 shows the methodology we followed for our research. In step 1 of Data Access and Understanding, we seek to know the most reliable sources of data and understand how they were recorded and the formats of

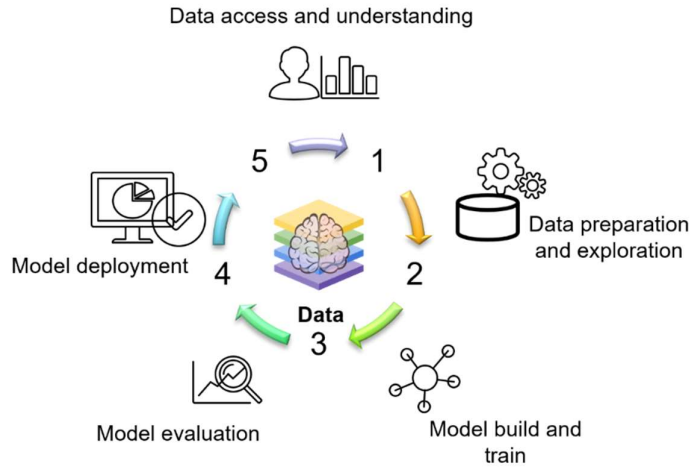

Fig. 1: Methodology followed for the research.

the data provided. Step 2 aims to review the records, eliminate incomplete or erroneous records, set nulls to a value if applicable, and prepare the data in a suitable format for modeling. In step 3, we separate data to train, review, evaluate, and apply the model (Random Forest in our case).

The 4th step allows us to evaluate the model against the actual data left for that case and to see the degree of certainty of our model.

Important note: A discussion will occur at each stage since data were analyzed at each step and further data were generated.

Data Access and Understanding

In Mexico, the INECC (National Institute of Ecology and Climate Change) dictates how data from pollutant concentration monitoring should be recorded in a database. All state or municipal governments in the republic that wish to monitor air quality for official purposes must adhere to these guidelines. For example, for each pollutant being monitored, the hourly average is recorded for each hour, starting from hour 0 to hour 23 of each day. Flags are added to indicate whether the recorded data is correct or inconsistent and is therefore invalidated.

In Mexico City and the metropolitan area, the official body that records pollutant concentrations and indicates air quality is the DMA (Atmospheric Monitoring Directorate), which depends on the SEDEMA (Ministry of the Environment) of the Mexico City Government, and operates 29 monitoring stations, 12 of which belong to the State of Mexico.

Fig. 2 shows Mexico City framed by a line and the surrounding municipalities of the State of Mexico, which make up the Metropolitan Area. Each point refers to the locations of the monitoring stations.

The names of the monitoring stations are listed in Table 1. The records accessed are grouped by monitoring station, date and hourly average and by each pollutant, with the inclusion of flags as shown in Table 2. For example, the first record shows that it belongs to station 243, on 25 December 2020, to the hourly average 01 of the pollutant NO, which has the status of validated with a 1, and its record value of NO (nitric oxide), is 0.001 ppm.

Table 1: Some air quality monitoring stations in the CDMX metropolitan area.

Station ID	Short_ name	Name	Municipality	Entity
242	AJM	Ajusco Medio	Tlalpan	CDMX
243	ATI	Atizapán	Atizapán de Zaragoza	Estado de México
300	BJU	Benito Juarez	Benito Juárez	CDMX
244	CAM	Camarones	Azcapotzalco	CDMX
245	CCA	Centro de Ciencias de la Atmósfera	Coyoacán	CDMX
246	CHO	Chalco	Chalco	Estado de México
248	CUA	Cuajimalpa	Cuajimalpa de Morelos	CDMX
249	CUT	Cuautitlán	Tepotzotlán	Estado de México
250	FAC	FES Acatlán	Naucalpan de Juárez	Estado de México
431	FAR	FES Aragón	Nezahualcóyotl	Estado de México

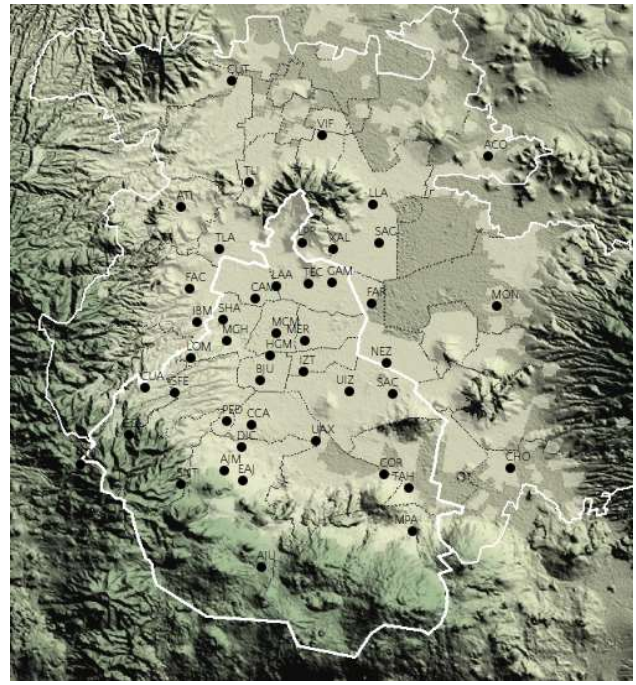

Fig. 2: Location of monitoring stations for reference only taken from the site http://www.aire.cdmx.gob.mx/.

Table 2 Example of pollutant records as submitted.

Station ID	Date	Time	Parameter	Validate	Value_act
243	2020-12-25	1	NO	1	0.001
243	2020-12-26	1	NO	1	0.001
243	2020-12-31	1	NO	1	0
243	2020-01-02	1	NO_2	0	\0
243	2020-01-03	1	NO_2	1	0.003

In the field "Parameter," the pollutant is recorded; in "Validate," a 1 if the data is correct and 0 if it is not, and in "Value_act," the concentration data is recorded on that date and at that time in each station.

Data Preparation and Exploration

Because the analysis of the data requires that all pollutants be found in the same record, it was necessary to "pivot" the records for each pollutant (CO carbon monoxide, NO nitrogen oxide, PM_{10} particulate matter equal to or less than 10 $\mu g.m^{-3}$, NOx nitrogen oxide, NO_2 nitrogen dioxide, SO_2 sulfur dioxide, O_3 ozone, and $PM_{2.5}$ particulate matter equal to or less than 2.5 $\mu g.m^{-3}$), so that each was a field in the table and is shown in Table 3.

The INECC has established specific pollutants (referred to as "Criteria"; see Table 4 for more information) for the Mexican territory. It is precisely those pollutants that we must measure, according to INECC parameters, to

Table 3: Pollutant records converted to fields.

FECHA	HORA	CO	NO	PM_{10}	NOx	NO_2	SO_2	O_3	$PM_{2.5}$
03/01/2020	8	0.4	0.001	5	0.008	0.007	0.001	0.032	2
03/01/2020	9	0.5	0.004	9	0.014	0.01	0.001	0.03	4
03/01/2020	10	0.7	0.012	8	0.03	0.018	0.001	0.025	1
03/01/2020	11	0.7	0.012	17	0.032	0.019	0.001	0.029	4
03/01/2020	12	0.8	0.015	13	0.039	0.024	0.001	0.032	3
03/01/2020	13	0.6	0.005	39	0.018	0.013	0.001	0.044	21
03/01/2020	14	0.6	0.003	17	0.014	0.011	0.001	0.048	8
03/01/2020	15	0.6	0.003	25	0.014	0.011	0.001	0.054	9

Table 4: Criteria pollutants and their allowable concentrations.

Pollutant	Concentration		Exposure time
	[ppm]	[µg.m-3]	
Ozone (O_3) NOM-020-SSA1-1993	0.11	216	1 h
	0.06		8 h
Carbon monoxide (CO) NOM-021-SSA1-1993	11	12.595	8 h
Lead (Pb) NOM-026-SSa1-1993	n/a	1.5	Quarterly
Sulfur dioxide (SO_2) NOM-022-SSA1-1993	0.13	341	24 h
Nitrogen dioxide (NO_2) NOM-023-SSA1-1993	0.21	395	1 h
Total Suspended Particles (TSP) NOM-025-SSA1-1993	n/a	120	24 h
		50	Annual
PM_{10}	n/a	65	24 h
		15	Annual
$PM_{2.5}$	n/a	1.5	24 h
			Annual

determine whether the air quality we breathe is satisfactory or not.

Below, weather data is aggregated by date and time for temperature, wind speed and direction, precipitation, humidity, and pressure, which were obtained from the Mexican National Meteorological Service.

Seasonal Analysis of PM_{10}

Our first analysis looks for any seasonality of the PM_{10} variable. It contrasts it with some climatic variables that might be related, such as wind speed, precipitation, humidity, and temperature. When we performed this analysis, we found inconsistent data at the stations Benito Juarez, Cuautitlán, Gustavo A. Madero, and Iztacalco. Since missing values for one or more fields affect the model, these stations have been removed from our study.

For this purpose, we have averaged the monthly values of the mentioned variables using the open-source programming language R (CRAN 2022), as the basis for all our research. In each case, if appropriate, we will show snippets of the code we have used. The computer equipment is a 7th-generation Intel Core i7 laptop with 16 GB of RAM.

Fig. 3 aimed to see if there was a relationship between PM_{10} and some meteorological variables affecting its concentration or measurement, such as wind speed, precipitation, humidity, and temperature (Dung et al. 2019). According to the monthly averages plotted, we can observe that PM_{10} has its highest measurements at the beginning and end of the year and its lowest point in August. However, we

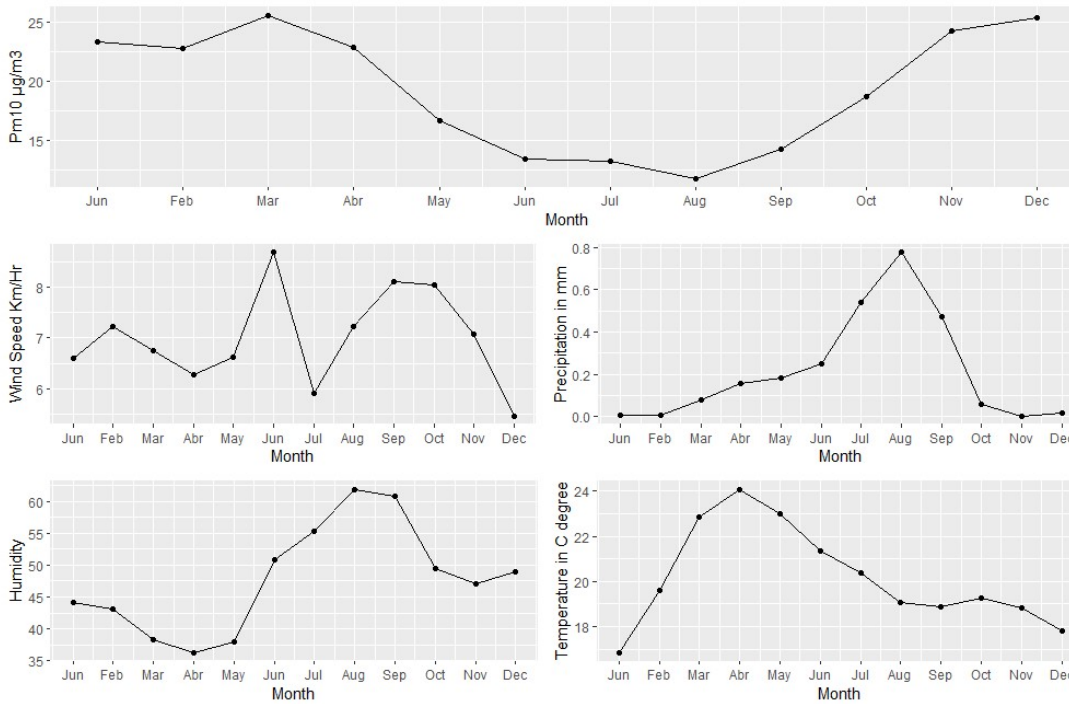

Fig. 3: Monthly comparison of 2020 PM_{10} and some meteorological indicators.

cannot observe a similar behavior derived from the monthly averages, for example, for velocity, which has the highest values in June, September, and October, very different compared to particulate matter. The cases of temperature and humidity also show no relationship with PM_{10}. An inverse relationship can be seen in the case of precipitation, which

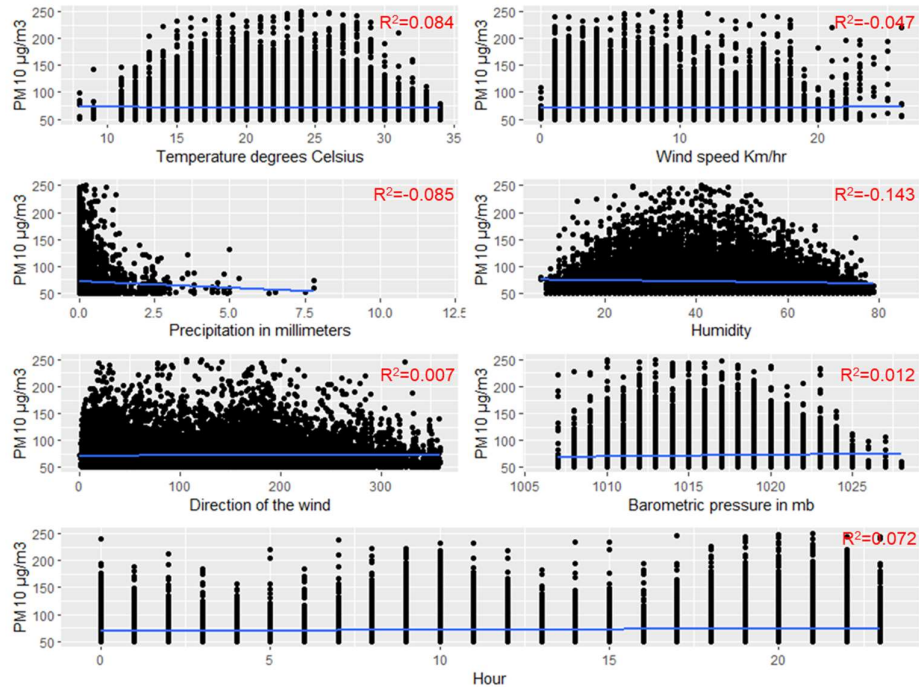

Fig. 4: Correlation of PM_{10} with meteorological variables.

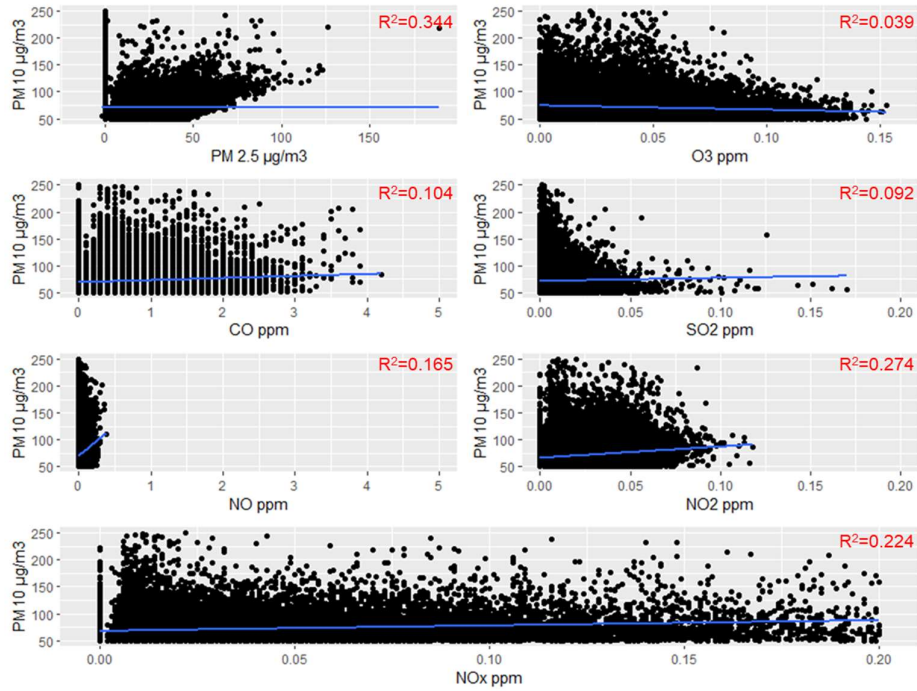

Fig. 5: Correlation of PM_{10} and other pollutants.

has its highest values in August, in contrast to PM_{10}, and its lowest values at the beginning and end of the year.

In this first exploration, the observations present certain behaviors that allow us to verify that precipitation has a specific correlation with PM_{10}. However, we performed a correlation analysis and plotted the results in Figs. 4 & 5 for a more accurate analysis.

Correlation of Independent Variables

To carry out the analysis, we calculated the covariance and correlation of the variables involved using a linear regression model using the Pearson method and the coefficient of determination R^2.

Fig. 4 shows the correlations between PM_{10} and meteorological variables such as temperature, wind speed, precipitation, humidity, wind direction, barometric pressure, and time of day (time is plotted in this item for the practicality of graphical presentation).

The highest correlation is found in unfavorable humidity, i.e., the more humidity, the less PM_{10} with an R2 value of -0.143; the next most important are precipitation with -0.085 and temperature with 0.084. The others have even lower values.

The exact process was performed in Fig. 5 for the other pollutants for which data are available, such as $PM_{2.5}$, O_3, CO, SO_2, NO, NO_2, and NOx.

The highest coefficients are $PM_{2.5}$ with 0.344, NO_2 with 0.274, NOx with 0.224, NO with 0.165, CO with 0.104, SO_2 with 0.092, and O_3 with 0.039.

Although there is no conclusive data at this exploration stage, there is a correlation of meteorological variables in which humidity is inverse and the highest. In the case of pollutant variables, the highest value is $PM_{2.5}$.

There is no single indirect variable that is decisive. However, the set of all of them will define the behavior of PM10, and their modeling will be subject to this characteristic of our study data.

Model Construction and Training

Based on the experience of other research, such as Deshmukh's (Deshmukh & Gulhane 2016), we think that to develop a well-founded predictive model, a cluster analysis is necessary, which allows us to group similar objects, in our case, monitoring stations to see if some of them have similar behaviors or rather, similar measurements and group them so that perhaps, we have not one but several models according to the number of clusters found.

In our research, the function fviz_nbclust belonging to the R language package "factoextra" (cran.r-project.org 2022) was used to determine the optimal number of clusters. The k-means clustering minimizer aims to minimize the total variation within clusters or the total sum of squares within the cluster. That value is required to be the minimum. This function also plots the results using the so-called "elbow" method, using the "silhouette" method that calculates the average of the observations for different values of k. The optimal number of k is the highest over the range of k.

In Fig. 6, the highest number is found in cluster 2 (the elbow), indicating 2 clusters of monitoring stations. To

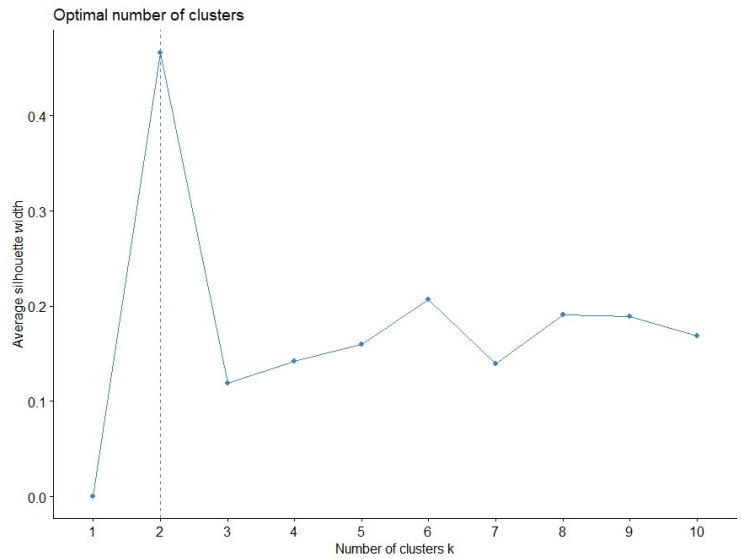

Fig. 6: Optimal number of clusters.

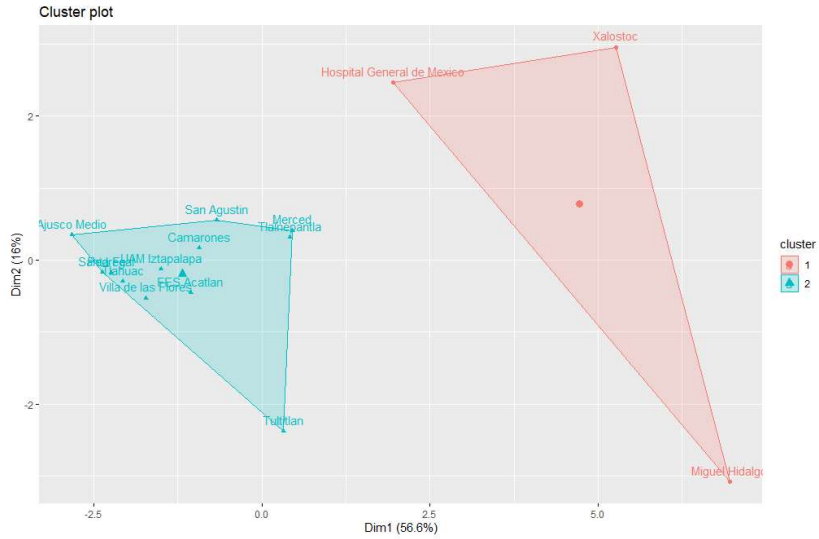

Fig. 7: Grouping of monitoring stations.

locate which stations belong to which block (clustering) by Euclidean distances, we use the function kmeans (df, centers = 2, nstart = 25), also included in the "factoextra" package, to which we parameterize the number of centroids in 2 (the value obtained above) and the other value of nstart indicates the number of random initial partitions, a minimum value of 25 is suggested. This function uses, by default, the Hartigan and Wong algorithm.

The "factoextra" function fviz_cluster(model_k1, data = df) plots the clustering results as shown in Fig. 7.

Dim1 and Dim2 shown on the axes are the new variables derived from the calculations derived from the principal component analysis process; % indicates the data's variability, i.e., Dim1 represents 50.2% variability and Dim2 14.6%. Three of the 16 stations are from Cluster 1: Hospital General de México, Xalostoc, and Miguel Hidalgo, and the other 13 are grouped in Cluster 2.

The Method Selected for Our Model

A suitable method for classification and prediction is decision trees, but they have a different level of predictive accuracy than other methods and are not very robust (James et al. 2017). By aggregating multiple trees, methods such as bagging and random forest significantly improve their performance.

In our case, we have decided to use the random forest method as we consider it meets our requirements of having an excellent predictive performance and excellent libraries in R. For its implementation, we have chosen the "ranger" package, which is a speedy implementation of random forest (Alvear 2018), in this site is also proposed a method for its application of which we follow some steps.

We have also based the application of random forest on what is proposed by the website "Decision trees, random forest, gradient boosting and C5.0" (Joaquín Amat 2020), which we found to be correct and very similar to other sites proposed by other data science experts.

The first part is to find the number of trees needed. However, this is not a critical parameter; it can improve the resulting model, especially regarding the computational resources used.

Using the validation with Out-of-Bag error (root mean squared error) in Fig. 8, we can see that the number of suggested trees is 381. It is important to note that this process can take up to 5 hours using our laptop.

Now, we implement the k-cross-validation (root mean squared error) to have another parameter regarding the number of trees we get.

When k-cross validation is used, the number of trees suggested is 391. As can be seen, the values in both cases are very similar, and for our case, we rounded the total number of trees to be used to 400.

Another important parameter is "mtry," which is the number of variables or predictors randomly sampled as candidates in each split. As in the tree number process, we validated them against "oob" and k-cross with the following results.

Fig. 10 shows that the value for mtry is 10 after the values of oob train rmse stabilize.

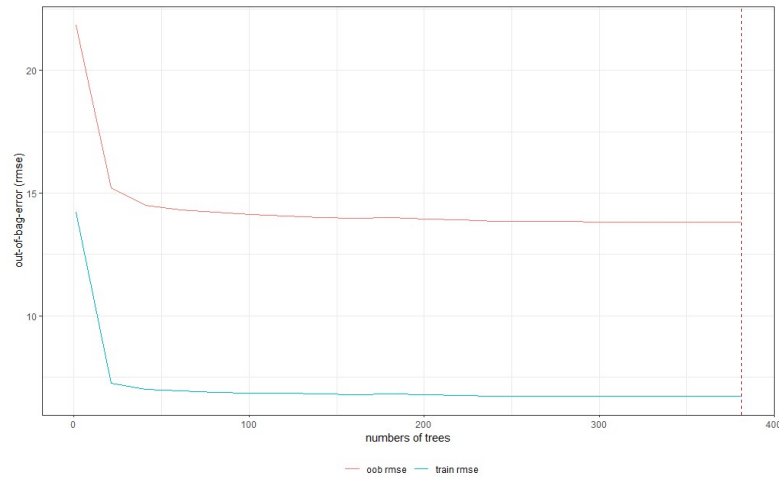

Fig. 8: Evolution of out-of-bag error vs. number of trees.

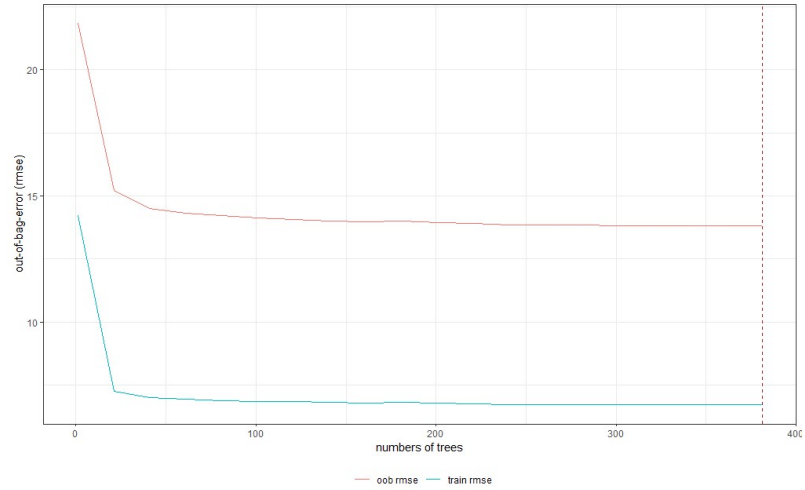

Fig. 9: Validation using kcross-validation (root mean squared error).

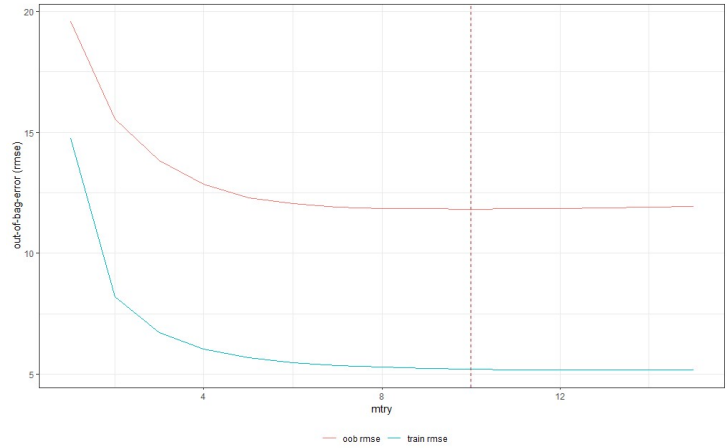

Fig. 10: Validation using out-of-bag error and mtry.

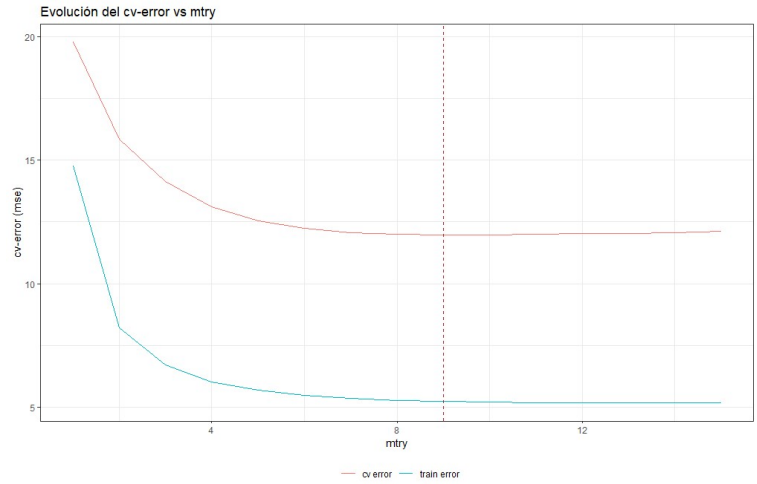

Fig. 11: Validation using kcross-validation and mtry.

When cv-error is used, mtry has a value of 9, as shown in Fig. 11, which infers that mtry will be between 9 and 10 in the model to be developed.

Grid Search

Understanding the individual behavior of the parameters is important to develop a good model. Still, it is better to analyze all of them as a whole because each one interacts with the others. So, it is better to perform a grid search or random search analysis, which we will do next.

The parameters when interacting with each other after applying the grid search are as follows (record 1): num trees 600, mtry is 9, max depth is 20, and oob is 11.9.

Now, we do a grid search based on cross-validation.

The results are consistent as they are num three 600, mtry is 9, max depth is 20, and with an average estimator of 12.1, which are the best parameters for the model and its final training.

RESULTS AND DISCUSSION

To obtain the final model, we use the existing parameters, train the model with test data, and then apply the model with the test data to validate the results and estimate how close the estimated data is to the predicted data.

Based on the model developed, what would be the expected results, what is its level of certainty, and is it feasible to operate in real-time?

In the model with the optimized parameters where we use training data against another sample for the test, we have, on average, a difference of 9.67 $\mu g.m^{-3}$ from the actual PM_{10} value.

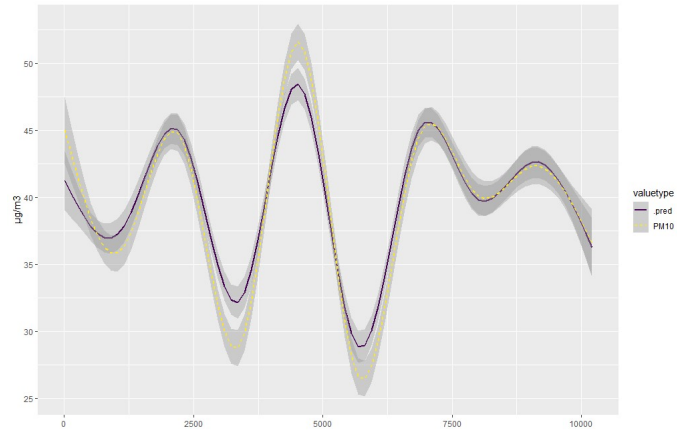

Fig. 12: Trend lines of predicted vs. actual PM_{10} values.

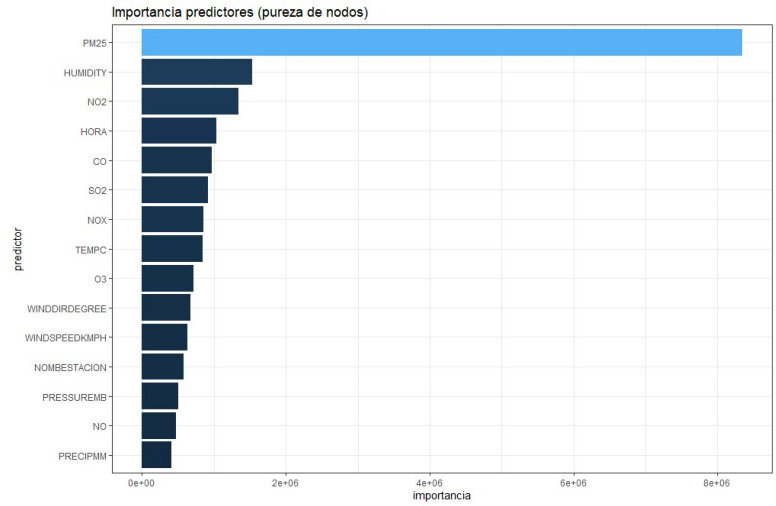

Fig. 13: Importance of predictors.

When calculating the certainty of our model, we have a value of 80.40 % certainty by calculating the deviation from the mean. This calculation is obtained by calculating the differences of the absolute differences between the calculated and actual PM_{10} values of each element, then dividing that value by each actual value multiplied by 100; the mean is then calculated and subtracted from 100%, as shown in the code above.

However, whether a model with a certainty of 80.4% is acceptable depends on what we are looking for with the model. For example, our model clearly shows trends very close to the actual, as shown in Fig. 12.

The number of predictors to be analyzed is 15, and having a certainty level of 80.4% is quite acceptable in light of the results obtained in models with even fewer variables to be analyzed.

Fig. 13 shows the predictors in the degree of importance from highest to lowest, with $PM_{2.5}$ standing out as the most important predictor, followed by relative humidity, which is a known correlation.

PM_{10} and $PM_{2.5}$ have a well-studied correlation, which is verified in our analysis. Notably, in fourth place, the importance of time is found, which we assume has to do with the peak hours of vehicular traffic, although, in Fig. 4, no such relationship is apparent.

CONCLUSIONS

It is feasible to apply Random Forest to model the behavior of air pollutants, in our case, to predict PM_{10} particles based on 15 predictors, including pollutants and climate measurements. These complex models involve a lot of database work before modeling.

A fundamental first step is selecting valid data from one or more tables in different records of the pollutant database and processing it to be analyzed in a model. Generally, there are not wrong models but incorrect or badly processed data. In this sense, we have found records where, for example, $PM_{2.5}$ values are almost 0 and PM_{10} values above 20 $\mu g.m^{-3}$, a fact that caught our attention because of the known correlation between these pollutants, which should be between 40% and 60% and impacts on the accuracy of the model. Then clustering, measuring, and selecting the parameters to be used in the model is also crucial.

As a recommendation, which we hope to make in a forthcoming study, it is fascinating to apply an analysis of the data now and model it using neural networks and thus be able to contrast these results with those obtained in our study.

REFERENCES

Alvear, J.O. 2018. Decision Trees and Random Forest. Retrieved from https://bookdown.org/content/2031/ensambladores-random-forest-parte-ii.html (accessed date 24 July 2022).

Amado, T.M. and De la Cruz, J.C. 2018. Development of Machine Learning-based Predictive Models for Air Quality Monitoring and Characterization. Proceedings of TENCON 2018. IEEE Region 10 Conference, pp. 668-672. https://doi.org/10.1109/TENCON.2018.8650518

Anaya Díaz, J.J. 2015. Prototype System for Estimating the Behavior of the Air Quality Index Using Computational Learning Techniques. National University of Colombia, Bogotá.

Pinto, J.A., Kumar, P., Alonso, M.F., Andreão, W.L., Pedruzzi, R., dos Santos, F.S. and de Almeida Albuquerque, T.T. 2020. Traffic data in air quality modeling: A review of key variables, improvements in results, open problems and challenges in current research. Atmos. Pollut. Res., 11(3): 454-468. https://doi.org/10.1016/j.apr.2019.11.018

Bellinger, C., Shazan, M., Zaïane, O. and Osornio-Vargas, A. 2017. A systematic review of data mining and machine learning for air pollution

epidemiology. BMC Public Health, 17: 907. https://doi.org/10.1186/s12889-017-4914-3

Biancofiore, F., Verdecchia, M., Di Carlo, P., Tomassetti, B., Aruffo, E., Busilacchio, M. and Colangeli, C. 2015. Analysis of surface ozone using a recurrent neural network. Sci. Total Environ., 514: 379-387. https://doi.org/10.1016/j.scitotenv.2015.01.106

Buenadicha, C., Galdon, G., Hermosilla, M.P., Loewe, D. and Pombo, C. 2019. Why it matters and how to make fair use of data in a digital world IDB. Ethical Data Management.

Camí Núñez, V., 2020 Estimation of air quality in Galicia using machine-learning techniques. Open University of Catalonia, Catalonia.

CRAN. 2022 Factoextra: Extract and Visualize the Results of Multivariate Data Analyses. [En línea] Retrieved from https://cran.r-project.org/web/packages/factoextra/readme/README.html (accessed date 24 July 2022).

Deshmukh, M.A. and Gulhane, R.A. 2016. Importance of clustering in data mining. Int. J. Sci. Eng. Res., 7(12): 54.

Dung, N.A., Son, D.H., Hanh, N.T.D. and Tri, D.Q. 2019. Effect of meteorological factors on PM10 concentration in Hanoi, Vietnam. J. Geosci. Environ. Protect., 7(11): 138.

EPA. 2021. Environmental Protection Agency. Retrieved from https://www.epa.gov/ground-level-ozone-pollution/ground-level-ozone-basics (accessed date October 15, 2021).

Gao, M., Yin, L. and Ning, J. 2018. Artificial neural network model for ozone concentration estimation and Monte Carlo analysis. Atmos. Environ., 184: 129-139.

Gonzalez Diaz, I. 2017. Big Data for CEOs and Marketing Directors: How to master Big Data Analytics in 5 weeks for managers. Independently Published. https://doi.org/10.1016/j.atmosenv.2018.03.027

Harbola, S., Koch, S., Ertl, T. and Coors, V. 2021. Air quality temporal analyzer: Interactive temporal analyses with visual predictive assessments. The Eurograph. Assoc., 12: 45-50. https://doi.org/10.2312/envirvis.20211083

Herman, M., Rivera, S., Mills, S., Sullivan, J., Guerra, P., Cosmas, A., Farris, D. and Kohlwey, M. 2013. The Field Guide to Data Science.

Hsieh, H.P., Lin, S.D. and Zheng, Y. 2015. Inferring air quality for station location recommendation based on urban big data. Knowl. Discov. Data Mining, 16: 437-446. https://doi.org/10.1145/2783258.2783344

Joaquín Amat, R. 2020. Decision Trees, Random Forest, Gradient Boosting, and C5.0. Available at https://rpubs.com/Joaquin_AR/255596 (Accessed date 18 September 2022).

Kurt, A. and Oktay, A.B. 2010. Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks. Expert Syst. with Appl., 37: 7986-7992. https://doi.org/10.1016/j.eswa.2010.05.093

Lightstone, S., Gross, B., Moshary, F. and Castillo, P. 2021. Development and assessment of spatially continuous predictive algorithms for fine particulate matter in New York State. Atmosphere, 12(3): 315. https://doi.org/10.3390/atmos12030315

Lim, C.C., Kim, H., Vilcassim, M.R., Thurston, G.D., Gordon, T., Chen, L.C. and Kim, S.Y. 2019. Mapping urban air quality using mobile sampling with low-cost sensors and machine learning in Seoul, South Korea. Environ. Int., 131: 105022. https://doi.org/10.1016/j.envint.2019.105022

Lino-Ramírez, C., Bautista-Sánchez, R. and Bombela-Jiménez, S.P. 2019. Use of a real-time system for the prediction of air pollution. Res. Comp. Sci., 148: 441-453.

Liu, H., Liu, S., Xue, B., Lv, Z., Meng, Z., Yang, X. and He, K. 2018. Ground-level ozone pollution and its health impacts in China. Atmos. Environ., 173: 223-230. https://doi.org/10.1016/j.atmosenv.2017.11.014

Li, X., Peng, L., Hu, Y., Shao, J. and Chi, T. 2016. Deep learning architecture for air quality predictions. Environ. Sci. Pollut. Res., 23: 22408-22417.

Mahajan, S., Liu, H.M., Tsai, T.C. and Chen, L.J. 2018. Improving the accuracy and efficiency of PM2. 5 forecast service using cluster-based hybrid neural network model. IEEE Access, 6: 19193-19204. https://doi.org/10.1109/ACCESS.2018.2820164

Mohan, M. and Kandya, A. 2011. An evaluation and comparison of the various statistical and deterministic techniques for forecasting the concentration of criteria air pollutants. Int. J. Environ. Pollut., 44(1-4): 96-105. https://doi.org/10.1504/IJEP.2011.038407

Perevochtchikova, M. 2009. The current situation of environmental monitoring systems in the metropolitan area. Estud. Demogr. Urban., 24: 513-547.

Perevochtchikova, M. 2013. Ambient impact assessment of the importance of environmental indicators. Manag. Pub. Policy, 22: 283-312.

Provost, F. and Fawcett, T. 2013. Data Science for Business. O'Reilly, Sebastopol, CA.

Roy, S. S., Pratyush, C. and Barna, C. 2018. Predicting ozone layer concentration using multivariate adaptive regression splines, random forest, and classification and regression tree. Inorg. Chem. Commun., 144: 109895.

Rybarczyk, Y. and Zalakeviciute, R. 2018. Machine learning approaches for outdoor air quality modeling: A systematic review. Appl. Sci., 8(25): 2570. https://doi.org/10.3390/app8122570.

Sertel, E., Demirel, H. and Kaya, S. 2012. Predictive mapping of air pollutants: A spatial approach. Inorg. Chem. Commun., 17.

Sihag, P., Kumar, V., Afghan, F. R., Pandhiani, S. M. and 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.

Soh, P. W., Chang, J. W. and Huang, J. W. 2018. Adaptive deep learning-based air quality prediction model using the most relevant spatial-temporal relations. IEEE Access, 6: 38186-38199. https://doi.org/10.1109/ACCESS.2018.2849820.

Soh, P.W., Chen, K.H., Huang, J.W. and Chu, H.J. 2017. Spatial-temporal pattern analysis and prediction of air quality in Taiwan. Inorg. Chem. Commun., 11: 1-6. https://doi.org/10.1109/UMEDIA.2017.8074094.

Tan, M., Hatef, E., Taghipour, D., Vyas, K., Kharrazi, H., Gottlieb, L. and Weiner, J. 2020. Including social and behavioral determinants in predictive models: trends, challenges, and opportunities. JMIR Med. Inform., 8(9): e18084. https://doi.org/10.2196/18084.

Wang, D., Wei, S., Luo, H., Yue, C. and Grunder, O. 2017. A novel hybrid model for air quality index forecasting based on a two-phase decomposition technique and modified extreme learning machine. Sci. Total Environ., 580: 719-733. https://doi.org/10.1016/j.scitotenv.2016.12.018.

Wei, W., Ramalho, O., Malingre, L., Sivanantham, S., Little, J.C. and Mandin, C. 2019. Machine learning and statistical models for predicting indoor air quality. Indoor Air, 29(5): 704-726. https://doi.org/10.1111/ina.12580.

World Health Organization (WHO). 2014. Million Premature Death Annually Linked to Air Pollution. Retrieved from https://www.who.int/news/item/25-03-2014-7-million-premature-deaths-annually-linked-to-air-pollution (Access date 20 October 2021).

World Health Organization (WHO). 2019. Out of 10 People Worldwide Breathe Polluted Air, But Moe Countries Are Taking Action. Retrieved from https://www.who.int/es/news/item/02-05-2018-9-out-of-10-people-worldwide-breathe-polluted-air-but-more-countries-are-taking-action (Accessed date 20 October 2021).

Yadav, S. 2020. Predictive Model for Analyzing PM 2.5 Level Of Air. In Proceedings of the Int. Conf. Innov. Comp. Commun., 15: 295. https://dx.doi.org/10.2139/ssrn.3562955.

Yuchi, W., Gombojav, E., Boldbaatar, B., Galsuren, J., Enkhmaa, S., Beejin, B. and Allen, R. W. 2019. Evaluation of random forest regression and multiple linear regression for predicting indoor fine particulate matter concentrations in a highly polluted city. Environ. Pollut., 245: 746-753. https://doi.org/10.1016/j.envpol.2018.11.034.

Zhang, C. and Yuan, D. 2015. Fast Fine-Grained Air Quality Index Level Prediction Using Random Forest Algorithm on Cluster Computing of Spark. IEEE, pp. 929-934. https://doi.org/10.1109/UIC-ATC-ScalCom-CBDCom-IoP.2015.177.

Zhang, Y., Wang, Y., Gao, M., Ma, Q., Zhao, J., Zhang, R. and Huang, L. 2019. A predictive data feature exploration-based air quality prediction approach. IEEE Access, 7, pp 30732-30743. https://doi.org/10.1109/ACCESS.2019.2897754.

Zhou, Q., Jiang, H., Wang, J. and Zhou, J. 2014. A hybrid model for PM2.5 forecasting based on ensemble empirical mode decomposition and a general regression neural network. Sci. Total Environ., 496: 264-274. https://doi.org/10.1016/j.scitotenv.2014.07.051.

Zhou, W., Wu, X., Ding, S. and Cheng, Y. 2020. Predictive analysis of the air quality indicators in the Yangtze River Delta in China: An application of a novel seasonal grey model. Sci. Total Environ., 748: 141428. https://doi.org/10.1016/j.scitotenv.2020.141428.

Zhu, D., Cai, C., Yang, T. and Zhou, X. 2018. A machine learning approach for air quality prediction: Model regularization and optimization. Big Data Cogn. Comput., 2(1): 5. https://doi.org/10.3390/bdcc2010005.

ORCID DETAILS OF THE AUTHORS

Alfredo Ricardo Zárate Valencia: https://orcid.org/0000-0002-9584-4593
Antonio Alfonso Rodríguez Rosales: https://orcid.org/0000-0002-2889-075X