Machine Learning Techniques for PM_{10} Levels Forecast in Bogotá

Nicolás Mejía Martínez, Laura Melissa Montes, Ivan Mura, Juan Felipe Franco
Universidad de los Andes, Bogotá, Colombia
{n.mejia10, lm.montes10, i.mura, jffranco}@uniandes.edu.co

***Abstract*—Air quality in Bogotá, Colombia, has become of increasing concern. Especially, the levels of PM_{10} are alarming, because of their relation to health risks. A forecast system for PM_{10} levels is beneficial for developing preventive policies of environmental authorities. This paper proposes different forecasting models of particulate matter obtained with three machine learning techniques. A dataset from 8 air quality monitoring stations including PM_{10} and environmental measurements was constructed. Three selection methods of relevant variables for prediction were assessed: selecting variables with the assistance of an expert group, and using two automatic selection methods. Having three sets of potential variables to use as an input, three different forecasting methods were implemented: logistic regression, classification trees and random forest. Finally, a validation and comparison of results are made, to conclude about the best forecast model to be implemented for the city.**

***Keywords*–Air quality forecast, PM_{10}, predictive models, logistic regression, classification and regression trees, random forest.**

I. Introduction

Air quality has become a major public health concern around the globe. Annually, approximately 6.5 million deaths are linked to air pollution [1], which accounts to 10% of global deaths [2]. Besides the implementation of polluter regulations, several national and local governments develop air quality monitoring and reporting systems directed toward the population. Immediate alerts result ineffective, as the population exposure to air contamination is occurring at the moment of the forewarning. Preventive warnings, on the other hand, can mitigate the negative health effects of pollutants by calling for adopting precautions, according to the predicted pollution level. Hence, air pollution forecasting has gained attention in recent years. However, it usually results expensive and requires a high level of technical knowledge. Developing countries lack the economic resources for implementing elaborate forecasting software for air pollution. Statistical and machine learning methodologies arise then as practical solutions for developing effective predictive models of pollutants. Logistic regression, classification and regression trees and random forest are statistical and machine learning techniques that fit the requirements; additionally they work with incomplete databases, which is a recurring characteristic in Latin American cities datasets of air pollution. Additionally, the possibility of elaboration of forecast models in open source software decreases significantly the costs of preventive policies of air pollution to local governments.

Particulate matter (PM) is in Latin America the pollutant that most affect population [3]. The most dangerous particles are those with a diameter of 10 microns or less (PM_{10}). Therefore, PM_{10} is considered in this work for elaborating a forecast model that can be used to raise preventive alarms. We evaluate a forecasting system of PM_{10} elaborated with three data mining techniques: logistic regression, classification trees (CART) and random forest (RF) for the city of Bogotá. We evaluate the input data required for our prediction models, according to several selection criteria. Then, we implement the techniques for the predictions, according to a proposed methodology. Afterwards, the results are analyzed in order to identify the best forecast method and the strengths and weaknesses of each method. The predictions of the three data mining techniques are finally evaluated to compare the different approaches and draw conclusions on forecast system that could be implemented.

II. Data sources and variables

The Monitoring Network of Air Quality in Bogotá, Colombia (known by its Spanish acronym RMCAB), established in 1998 and administrated by the Secretary of Environment of Bogotá, allows hourly recollection of pollutants concentrations and weather conditions. The network consists of 13 automated stations. However, due to equipment failures and technical problems, only 8 sites were chosen for this study. The RMCAB database provided data on PM_{10} concentrations and meteorological values for each station in the network. Wind speed, wind direction, precipitation and temperature were chosen as meteorological input variables for the prediction models. Daily 24-hour average, maximum, minimum and median data were calculated based on the hourly available data. Researches conducted at the department of Civil and Environmental Engineering at Universidad de los Andes demonstrated a relationship between physical atmospheric properties and PM_{10} concentration. Thus, atmospheric variables are taken into account when forecasting PM_{10} levels. These variables are: the precipitable water, the height of the planet boundary layer, the differential of height of the superficial inversion layer and the differential of height of the stable layer.

Two additional categorical variables are elaborated for a general model that includes the complete dataset of all stations. The first variable, station, is the number of the station where

978-1-5386-9459-6/18/$31.00 ©2018 IEEE

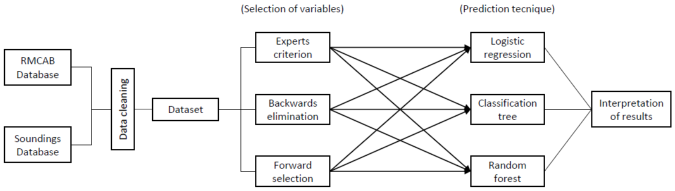

Fig. 1. Summary of the methodology used for constructing the forecasting models.

the data was collected. The second variable, station type, is an ordered categorical variable constructed from the historical average of PM_{10}: the stations are divided into 4 groups according to their PM_{10} concentration: high, medium, low and very low.

A. Construction of the dataset

After obtaining the measurements, data diagnosis and consolidation was performed in order to obtain a valid dataset for the construction of the models.

The original RMCAB database shows invalid values, so a data tidying process was performed. General rules described by Bernal and Melo [4] were applied for PM_{10} and environmental values. Even though there is data available from 1998 to 2014, the time window chosen for this study is from 2009 on wards, due to very low density of available data in previous years.

Since the time series of PM_{10} concentrations and meteorological values are not complete, we considered inadequate a prediction of the exact concentration value of PM_{10} based on time-series analysis. Autoregressive models, integrated models, moving average models and its combinations (e.g. AR, MA, ARIMA) could not be applied to the dataset as they depend on the complete time series. Therefore, based on the limit concentrations of PM_{10} established by the World Health Organization (WHO) and Colombian national government, we decided to divide the data into categories and make the prediction model based upon them. Random Forest [5], Classification and Regression Trees [6] and Logistic Regression [7] are methods that can all work with incomplete time series dataset.

We proposed 3 data ranges for our prediction model: the first range being PM_{10} concentrations under 50 $\mu g/m^3$ (concentration limit established by the WHO [8]), the second between 50 and 100 $\mu g/m^3$ and the third above 100 $\mu g/m^3$ (concentration limit as per the Colombian national law [9]).

Data distribution shows that every station only has 2 significant groups, while the third one is not densely represented. Therefore, we decided that the prediction categories would be composed of only two categories, that differ according to the data distribution. Additionally, a general model for all stations is proposed, evaluating its effectiveness in both 50 and 100 $\mu g/m^3$ ranges.

TABLE I
REGRESSION RESULTS BY SELECTION OF EXPERTS, BACKWARDS ELIMINATION AND FORWARD SELECTION METHOD. ODDS REPORTED. ALL OF THE ENVIRONMENTAL AND ATMOSPHERIC ARE VALUES FROM THE DAY BEFORE THE FORECASTING.

Variables	**Model**		
	Experts	**Backwards**	**Forward**
Mean PM_{10}	0.905	0.925	0.922
Mean Wind Speed	0.398	0.597	0.634
Max Wind Speed	1.301	-	-
Mean Wind Direction	0.998	-	-
Mean Temperature	0.825	1.162	-
Max Temperature	1.136	-	-
Max Precipitation	1.576	1.110	1.102
PBLH	-	1.000	-
Max PM_{10}	1.003	-	-
Monday	0.022	0.022	0.023
Tuesday	0.027	0.027	0.028
Wednesday	0.037	0.037	0.038
Thursday	0.039	0.041	0.041
Friday	0.036	0.037	0.037
Saturday	0.125	0.125	0.125
February	1.559	-	-
May	2.072	-	-
June	3.673	2.370	2.248
July	2.839	1.784	1.641
August	2.303	-	-
September	2.202	-	-
October	1.666	-	-
December	-	-	0.639
Median PM_{10}	-	0.982	0.984
Min Temperature	-	0.785	0.837

III. RELATED WORK

Air quality forecast is a research topic subject to intense research, given the increasing awareness of the adverse effect that the pollutants have on the environment and the population. Academic investigation presents distinct approaches to forecasting, from classical approaches based on time series or linear regression to methodologies based on machine learning and data mining.

For example Karatzas, Pappadopulus and Slini [10] use a linear regression model in order to forecast the mean concentration of Ozone in the city of Athens. For the prediction they consider an array of environmental variables that according to literature play a role in the concentration of this pollutant. The article concludes that the model does not achieve an acceptable predictive power, because its simplicity fails to capture the behavior of the pollutant dynamics.

On the other hand there are studies like the ones developed by Cortes [11], Quiones [12] or Mejía [13] that before the prediction, complete the dataset with diverse imputation methodologies in order to use dynamic models, based on time series, to forecast the mean amount of particulate matter in a specific station of the network in Bogotá.

Finally, studies like the one proposed by Siwek and Osowoki [14] make use of data mining techniques, Random Forest, Support Vector Machine and Neural Networks in order to forecast the contamination by particulate matter, ozone, carbon dioxide and sulfur dioxide in Warsaw. In order to find the best input for the model the authors in [14] use two automatic selection methods: a genetic algorithm and stepwise regression. The article concludes that a structured selection of the predictors, prior to the construction of the model, contributes appreciably to enhance the predictive power of the models.

IV. Implementation of forecasting techniques

We designed a methodology that includes the evaluation of the dataset within logistic regression, classification trees and random forest, and its subsequent analysis. Such evaluation is completed with the three sets of variables chosen by the experts assisted criterion, the backwards elimination and the forward selection methods, respectively. The methodology is described in Figure 1.

A general prediction model, designed to predict PM$_{10}$ concentration levels for all the stations, is chosen as an example for the methodology implementation. This model is constructed with the complete dataset, and is implemented for the 100 $\mu g/m^3$ range. We generate a new variable in the dataset in order to convert our PM$_{10}$ continuous variable into a dichotomous one, as follows:

$$\text{PM}_{10}\text{level} = \begin{cases} 1|A & \text{if PM}_{10} \leq 100 \\ 0|B & \text{if PM}_{10} > 100 \end{cases} \tag{1}$$

Prior to analysis all data were normalized in order to guarantee stability in the algorithms. Additionally, we performed 10 fold cross validation for all three algorithms in order to tune the parameters of each one and estimate the most appropriate model [15]. As already mentioned, given the amount of variables present in the database, the process regarding the selection of variables to be included in the model was performed by 3 different methods: based on the criteria of Experts, Backwards Elimination and Forward Selection.

A. Logistic Regression Model

At first we estimate a logistic regression model using the variables selected by experts. As a first step we performed a validation process consisting in a residual analysis, an influential observation analysis and a goodness of fit test. We delete the observations that modify the fit of the model, and then we re-estimate the model with the remaining data. Both models full and reduced- were compared, concluding that the reduced model has better explanation power and better fits the data. Therefore, a reduced model is selected for working from now on. Then we proceeded to estimate the coefficients of the logistic model to quantify the effects of independent variables in our response one.

We perform the two automatic selection methods mentioned above (backwards elimination and forward selection) for the 100 $\mu g/m^3$ partition. We have huge similarities in the variables used in all three models, since most of the variables selected by the automatic methods were proposed in the initial experts model. We conclude that the selected variables used for the study not only affect the pollution from an environmental point of view but also from a statistical one, giving to the models and the study case more robustness.

Instead of the raw coefficients, we choose to analyze the odds ratio (OR) referring to the variables for obtaining proper interpretation of the coefficients. The resulting odds of the maximum likelihood estimation of the models are shown in Table I. Only OR that are statistically significant with an α of 0.05 are presented.

Variables like the mean PM$_{10}$ levels, the mean wind speed, mean temperature, max precipitation of the day before, the day of the week and the month of the year score significant coefficients across all three models, indicating the considerable effect these variables have on the PM$_{10}$ levels of any given day.

B. Classification and Regression Tree

Two classification trees were obtained: one for the experts criterion model and one for the automatic selection models (the same tree was obtained for forward selection and backwards elimination). A graphical representation of both trees is shown in Figures 2 and 3 respectively. Both CART models considered several variables for tree splitting. Accuracy values of the complete model reached 95.18% and 95.23% respectively. However, correct prediction of the category above 100 $\mu g/m^3$ is significantly lower, between 32.80% and 35.64% . This indicates that the model does not capture accurately the highest ranges of PM$_{10}$ values. This result is not adequate, as the efficiency of prediction for the higher category is critical for the quality of the model.

C. Random Forest

Random Forest (RF) was later executed, expecting better performance values than CART methodology to forecast 24-h PM$_{10}$ levels. Our first approach to RF determined models with default cutoff values for categories 1/k, where k is the number of classes, but with 1000 built trees. As suggested in Yang et al. [16], a default value of trees (500) can be insufficient in some cases. The number of variables used in each tree was selected using 10 fold cross validation.

The estimated models present an accuracy of 84.11% (experts), 95.22% (Forward selection) and 95.32% (Backwards elimination). Performance measures present similar values across experts, backwards elimination and forward selection sets of variables. However, as Logistic Regression and CART models, prediction of the highest 24-h PM$_{10}$ mean concentration range is not ideal it only reaches 50%. Even though this percentage is better than the two previous methods, it still presents a weak prediction of the most dangerous pollutant values.

D. Comparison of performance between general models for all stations

With the resulting models of logistic regression, classification tree and random forest methodologies, we proceeded to perform a comparison between them. We considered 3 different quality measures to compare the performance of the models: accuracy, sensitivity and specificity. Performance

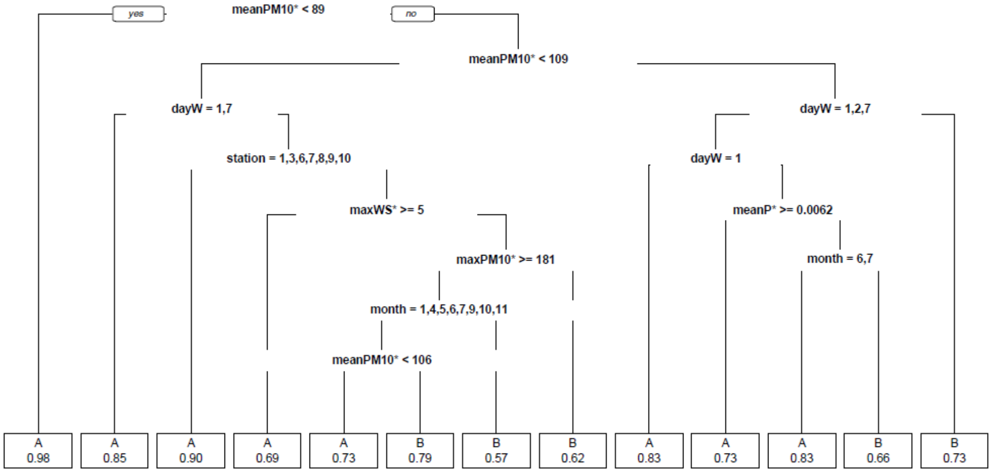

Fig. 2. Classification tree for the general model of all stations with automatic selection methods (cutting point of 100 $\mu g/m^3$). Values below the winning class in the final nodes indicate the percentage of the data in the node that belong to the winning class. All of the environmental and atmospheric values are from the day before forecasting.

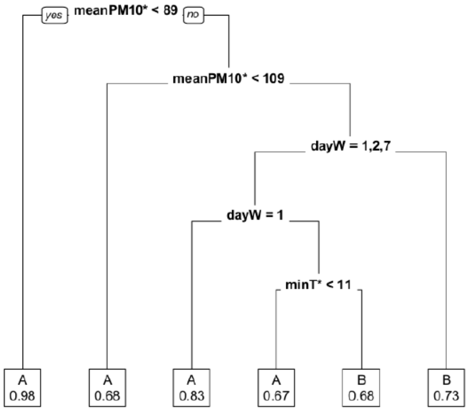

Fig. 3. Classification tree for the general model of all stations with automatic selection methods (cutting point of 100 $\mu g/m^3$). Values below the winning class in the final nodes indicate the percentage of the data in the node that belong to the winning class. All of the environmental and atmospheric are values from the day before the forecasting.

TABLE II
PERFORMANCE MEASURES OF THE MODELS WITH A CUTTING POINT OF 100 $\mu g/m^3$

Method	Model	Performance Measures		
		Accuracy	Sensitivity	Specificity
Logistic Regression	Experts	95.15%	98.59%	43.57%
	Backwards	95.11%	98.57%	43.03%
	Forward	95.08%	98.53%	43.12%
CART	Experts	95.29%	99.05%	32.80%
	Backwards	95.18%	98.91%	33.12%
	Forward	95.18%	98.91%	33.12%
Random Forest	Experts	84.11%	86.70%	81%
	Backwards	95.32%	89.39%	48.70%
	Forward	95.22%	84.40%	49.81%

measures for general forecast models with the 100 $\mu g/m^3$ category cutting point are shown in Table II.

We consider accuracy and specificity as being critical measures. The objective of the forecast model is to provide alerts about high PM$_{10}$ concentration levels, therefore high specificity is to be preferred over high sensibility. In other words, a model that predicts with high certainty the highest class of PM$_{10}$ concentrations is selected over a model that accurately forecasts the lowest class of PM$_{10}$.

Overall, we can see that all models have similar accuracy percentages, therefore we proceed to compare them based on specificity. The method that presents the worst performance is CART, with a specificity of approximately 30%. Logistic regression presents a specificity of 43%, and Random Forest is the method that overall performs better. On the other hand, we notice that the differences between the three evaluated models are minimal, being the greatest a difference of 0.50% percentage points the expert model obtains the best performance.

E. Random Forest Modified

Prediction of the highest PM$_{10}$ level or specificity- does not achieve desirable values in the models evaluated before (excluding the Random Forest method evaluating the experts model, for which specificity reaches 50% in the best scenario). Therefore we propose a method that improves the specificity of the model, using the Random Forest methodology.

We implemented an unbalanced proportion of votes to cutoff high and low concentration classes. Specifically, we generated an iterative model that evaluates the change in all the performance measures when the proportion of votes changes. Every iteration incremented the proportion of votes needed for the first class to be chosen by 2%. For example, in the first iteration every prediction needed more than 50% of the votes to assign the value into the first category. In the second

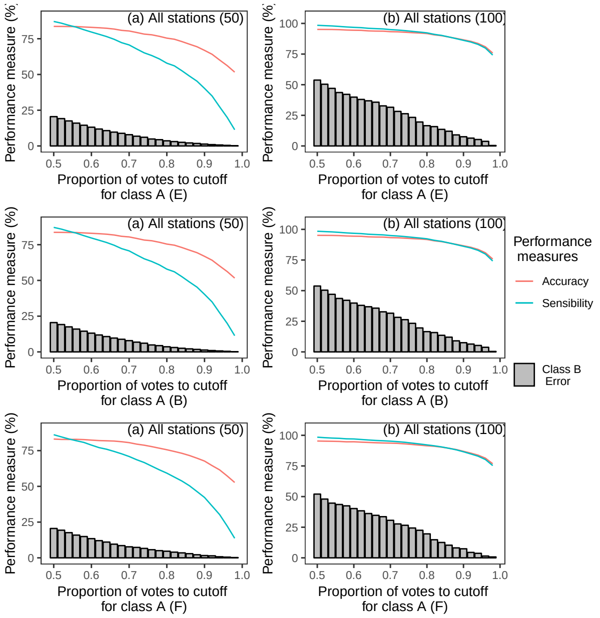

Fig. 4. Behaviour of performance measures according to the proportion of votes to cutoff for class A with experts criterion (E), backwards elimination (B) and forward selection (F) models

iteration, every prediction needed more than 52% of the votes, and so on. The results are shown in Figure 4.

As the cut point for the lowest class increases, the percentage of prediction error for the highest class (class B in Figure 4) decreases. The decrease of the prediction error for the highest class is analogous to the increase in specificity. The decreasing rate of class B error is more significant in these models, while the decrease of accuracy occurs in a much smaller rate. For example, in the experts criterion model with a 100 $\mu g/m^3$ cutting point, accuracy of PM_{10} prediction decreases 20% between the first and the last iteration (from 95% to 75%). Meanwhile, class-B error decreases from 53% to 0.3% between the first and last iterations, obtaining an almost perfect prediction of PM_{10} values above 100 $\mu g/m^3$. Therefore, this method is useful for adjusting the ideal values of the performance measures of each forecasting model. The optimal proportion between performance values cannot be determined by us, it should be established by the user of the forecasting model according to its needs.

This approach provides a useful tool for parameterizing the model to obtain the optimal results for the users, especially with models that have low specificity values.

V. Interpreting the Results

After estimating the models, our goal is now to interpret them. Besides the evaluation of the effectiveness of the model in terms of its performance, an analysis of the significance of the input models in all three forecasting models is performed to study the dynamics of PM_{10} and the variables of the day before the prediction.

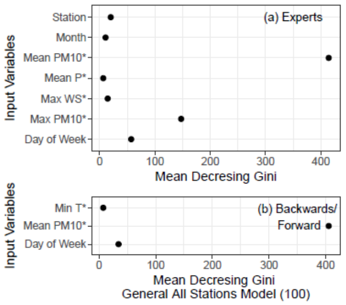

Fig. 5. Variable importance for the general CART model (cutting point 100 $\mu g/m^3$), for (a) Experts criterion and (b) Backwards elimination and Forward selection. Variables with an * are values from the day before the forecasting.

A. Logistic Regression

Table I contains the results of the logistic regression model over the data with the three variable selection techniques. The odds ratio (OR) will be 1 if the variable has no effect over the probability, > 1 if a change in the variable increases the probability and < 1 if a change in the variable decreases the probability of occurrence.

In the models, variables such as Maximum Wind Speed, Maximum Precipitation, the set of dummies representing Month, Precipitable Water (PWAT) for which an increase in value turns out in a higher probability to have a PM_{10} concentration below 100 $\mu g/m^3$.

On the other hand, there are variables such as Mean Temperature, the set of dummies representing Day of the week, the mean amount of PM_{10} on the day before, which appear to have a negative effect over the concentration of the pollutant. When these variable increase the day before, is more likely that the predicted day will have a PM_{10} concentration above 100 $\mu g/m^3$.

Finally, variables like the maximum amount of PM_{10}, the planetary layer boundary height (PBHL) do not have a statistical effect over whether or not the concentration of PM_{10} on a given day is below 100 $\mu g/m^3$.

B. Classification Tress

In the expert criterion classification tree, we have 12 partitions of 7 variables: mean PM_{10}, maximum PM_{10}, mean of wind speed, mean of precipitation, day of week, month and station. Backwards elimination and forward selection classification tree, on the other hand, presents 5 partitions of 3 variables: mean PM_{10}, minimum temperature and day of week. Figure 5 shows a graphical representation of the variable importance for both models. The mean of PM_{10} concentrations of the previous day plays an important role in the construction of the models. In addition, meteorological values present different splitting criterion for the data, so as categorical values (e.g., day of the week, month and station). Still, PM_{10} concentration is the most important variable for elaborating our forecasting models.

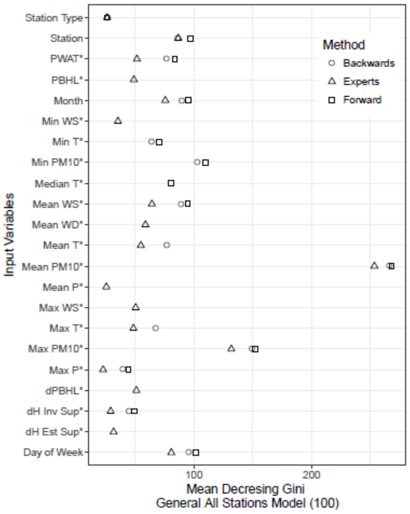

Fig. 6. Variable importance for the general RF model (cutting point 100 $\mu g/m^3$), for (a) Experts criterion and (b) Backwards elimination and forward selection. Variables with an * are values from the day before the forecasting.

C. *Random Forest*

Figure 6 shows the importance of each variable in the general model with the RF method. Interpretations for the variables can be performed globally, as variations between methods are not significant. The mean, maximum and minimum PM_{10} contribute largely in the estimation of the final models. Such relation can be related to a high auto-correlation for PM_{10} concentration series. Station number, station type, day of week and month obtain high importance in RF models; their importance is ranked second after PM_{10} concentrations characteristics of the day before.

Meteorological variables describe approximately 23%, 25% and 20% of the model for backwards elimination, forward selection and expert models, respectively. On the other hand, atmospheric variables contributed no more than 15% in estimating PM_{10} levels. As sounding is performed uniquely in the morning to measure the atmospheric variables, their values do not provide any information about the daily variations. Atmospheric values are however a viable complement to meteorological data to predict mean PM_{10} levels. Usefulness of both meteorological and atmospheric data for PM_{10} concentrations forecasting is confirmed by RF methodology.

VI. Conclusions and future work

The forecasts developed with each one of the methodologies described helped to understand the impact of climate and atmospheric variables on pollution levels. Moreover, they provide useful tool in the development of preventive strategies aimed to the mitigation of the dangerous effects of pollutants on the population.

Regarding the selection of the input variables, we found that choosing them with expert knowledge provides the best results, compared with automatic selection methods. All the three forecasting methods used in this paper provide satisfactory results, presenting good values according to some of the performance measures analyzed. The levels of PM_{10} were correctly predicted with accuracy between 70% and 90%. However, some models present a weak specificity, mostly with values under 40%.

Of the three methods, the models developed using the Random Forest approach were generally the best ones in terms of accuracy and specificity. These models have the advantage of reducing the variance of the forecast, and diminishing the probability of over fitting. Additionally, we propose the implementation of a Random Forest methodology with modified voting values, which provides a tool to improve the specificity of the models, without sacrificing the good measures of accuracy and sensitivity. We recommend the use of Random Forest given its versatility and superior performance, but all of the methods can be adapted to fit the user needs.

As for the dynamics of air pollution, we found that the particular characteristics of the stations, each one having its own significant variables, play a pivotal role in the behavior of the contamination.

In this paper we provide three basic methodologies for developing forecasting models. In a further research, we plan to execute a comparison between them and other statistical methodologies. Also given the gaps present in the original RMCAB dataset, different interpolation methods can be used to complete it. Another perspective for the analysis can be to quantify the impact that the pollution has in the life quality of the population exposed to it, in order to identify the monetary cost in health services for the city.

References

[1] Cozzi, L. et al.Energy and air pollution. 2016

[2] Hsu, A et al. Epi. environmental performance index. Global Metrics for the Environment, pages 1213, 2016.

[3] World Health Organization. Fact sheet ambient (outdoor) air quality and health, 2016.

[4] Bernal, L and Melo, N. Air Quality and the Impact of Regulatory Actions in Bogotá. B.s. thesis, Universidad de los Andes, Bogotá, Colombia, 2016.

[5] Breiman, L. Random forests. Machine learning, 45(1):532, 2001.

[6] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. Classification and regression trees. CRC press, 1984.

[7] German Rodriguez. Logit Models for Binary Data. http://data.princeton.edu/wws509/notes/c3.pdf, September 2007.

[8] Chan, M. and Danzon, M. Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide. 2010.

[9] Ministerio de Ambiente Vivienda y Desarrollo Territorial. Resolucin nmero 610 de 2010. Calidad del aire. 2005.

[10] Karatzas, K., Papadopoulos, A., & Slini, T. Regression Analysis and Urban Air Quality Forecasting: An Application for the city of Athens. Global Nest, 153-162. 2002.

[11] Cortes, S. Análisis del comportamiento de la concentración de material particulado menor a 10 micras en la localidad de Puente Aranda a partir de un modelo de regresión dinámica. Bogotá. 2010

[12] Quinones, L.Predicción de contaminación por PM10 en las estaciones de Kennedy y Carvajal . Bogotá. 2015.

[13] Mejía, N. Pronostico y Prevención de la contaminación por PM_{10} en la red de monitoreo de calidad de aire de Bogot (RMCAB)

[14] Siwek, K., & Osowski, S. Data Mining Methods for Prediction of Air Pollution. International Journal of Applied Math and Computer Sciences, 467-478. 2016.

[15] Mucherino, A., Papajorgji, P., and Pardalos,P. Data mining in agriculture, volume 34. Springer Science & Business Media, 2009.

[16] Yang, R. et. al. Comparison of boosted regression tree and random forest models for mapping topsoil organic carbon concentration in an alpine ecosystem. Ecological Indicators, 60:870878, 2016.