ANALYSIS OF INCIDENCE OF AIR QUALITY ON HUMAN HEALTH. A CASE STUDY ON THE RELATIONSHIP BETWEEN POLLUTANT CONCENTRATIONS AND RESPIRATORY DISEASES IN KENNEDY, BOGOTÁ

Nidia Isabel Molina-Gómez[1,2], Dayam Soret Calderón-Rivera[1], Ronal Sierra-Parada[1], José Luis Díaz-Arévalo[3], P. Amparo López-Jiménez[2]

[1.]Department of Environmental Engineering, Universidad Santo Tomás, Bogotá, Colombia
[2.]Hydraulic and Environmental Engineering Department, Universitat Politècnica de València, Valencia, Spain
[3.]Department of Civil and Agricultural Engineering, Universidad Nacional de Colombia, Bogotá, Colombia

ABSTRACT

Thousands of deaths associated with air pollution each year could be prevented by forecasting the behavior of factors that pose risks to people's health and their geographical distribution. Proximity to pollution sources, degree of urbanization, and population density are some of the factors whose spatial distribution enables the identification of possible influence on the presence of respiratory diseases (RD). Currently, Bogotá is among the cities with the poorest air quality in Latin America. Specifically, the locality of Kennedy is one of the zones in the city with the highest recorded concentration levels of local pollutants over the last 10 years. From 2009 – 2016, there were 8619 deaths associated with respiratory and cardiovascular diseases in the locality. Given these characteristics, this study set out to identify and analyze the areas in which the primary socio-economic and environmental conditions contribute to the presence of symptoms associated with RD. To this end, information collected in field by performing georeferenced surveys was analyzed through geostatistical and machine learning tools which carried out cluster and pattern analyses. Random forests and AdaBoost were applied to establish hotspots where RD could occur, given the conjugation of predictor variables in the micro-territory. It was found that random forests outperformed AdaBoost with 0.63 AUC. In particular, this study's approach applies to densely populated municipalities with high levels of air pollution. In using these tools, Municipalities can anticipate environmental health situations and reduce the cost of respiratory disease treatments.

Keywords: geostatistics, machine learning, sustainable development, air quality, hot spots

INTRODUCTION

In developing countries, air pollution is among the environmental problems of greatest concern. It is a risk factor for populations' health, which can affect different age groups more severely. Furthermore, growing urbanization has increased urban density and populations' proximity to pollution sources. Therefore, it is essential to analyze the impact of atmospheric pollutants on a population's health. In this vein, epidemiological studies and predictive modeling which employ machine learning (ML) techniques have been carried out.

Epidemiological studies consist of designing experimental or observational studies. Experimental studies are randomized and quasi-experimental trials, in which the researcher has a certain degree of control over the variables. Observational studies include cohort, case-control, cross-sectional and ecological studies (Kestenbaum 2019). In cohort studies, individuals are classified in sub-groups, according to exposure to a potential cause of sickness, in which the entire evolution of the cohort is monitored (Lazcano-Ponce et al. 2000). In case-control studies, a comparison is made between the groups in which the event occurs, and those in which it does not. Cross-sectional studies analyze the frequency of a health event with respect to the exposure level of the analyzed individuals or group in a given moment (Hernández and Velasco-Mondragón 2000). Ecological, correlational and exploratory, studies focus on studying groups with an analysis of geographic areas or different time periods, and are useful in evaluating multiple exposure levels (Borja-Aburto 2000).

ML techniques were also used to identify the influence of physical and chemical factors in the population's health. ML can process large volumes of data, as well as linear and non-linear relationships (Ivanov 2018). ML can also perform classification and regression tasks through decision trees, artificial neural networks (ANN), support vector machines (SVM) or through ensemble methods, such as random forests (RF) or adaptive boosting (AdaBoost-AdB). From a data set, ML is able to identify data patterns and predict their behavior (Kuhn and Johnson 2013). ANN were applied to determine the influence of physical and chemical stressors in hospital admissions for respiratory and cardiac diseases (Kassomenos et al. 2011; Polezer et al. 2018). Generalized boosting models were applied in the same manner for exposure periods before, during and after forest fires (Reid et al. 2016). Furthermore, Bayesian kernel regression was used to estimate the function of response doses and to identify the combination of pollutants responsible for adverse health effects (Bobb et al. 2015). Moreover, statistical tools such as generalized linear regression, multiple linear regression, logistic regression and ML techniques (RF, SVM, and ANN) were used to forecast atmospheric pollutant levels that may generate a public health risk (Huang et al. 2018; Ivanov et al. 2018; Kami 2019; Pandey et al. 2013; Weizhen et al. 2014; Zhan et al. 2017).

The above referenced studies forecasted pollutants' behavior and health effects from information recorded in a database. However, there is still a lack of forecasting of possible respiratory disease (RD) hotspots based on variables' spatial distribution and behavior. The spatial distribution of risk factors and their interactions in territories increase interest in knowing future spatial scenarios of possible health effects. The use of ML and spatial zoning of these factors facilitate forecasting variables' behavior by identifying hotspots with a territorial approach, which is more specific than national or capital city focuses. These scenarios are essential for decision-makers so that they are able to implement measures to mitigate costs related to the treatment of morbidity and mortality.

With nearly 8.3 million inhabitants and located along the plateau of the eastern range of the Colombian Andes at an elevation of 2600 meters above sea level (m.a.s.l), Bogotá is one of the most populated cities in Latin America and one of the cities with the highest recorded levels of atmospheric pollutants, which represent a risk factor for its population. 21.5% of medical consultations performed for the productive age population (15 – 65 years old) are related to air pollution (García-Ubaque et al. 2011). Furthermore, changes in NO_2, SO_2 and $PM_{2.5}$ concentration levels in Bogotá were correlated with statistically significant effects regarding changes in emergency room visits due to RD by children younger than fifteen years old, while changes in SO_2, PM_{10} and $PM_{2.5}$ were related to changes in emergency room visits for circulatory system diseases in adults older than sixty years (Rodríguez-Villamizar et al. 2018).

In addition to the above, several studies have been developed in Bogotá aimed at forecasting its air quality. A combined linear regression model was used to predict air quality (Westerlund et al. 2014). Moreover, ANN were applied to predict PM_{10} and $PM_{2.5}$ concentration levels (Franceschi et al. 2018). As a result of the data analyzed, it was determined that Kennedy is one of the zones in the city with the highest air pollution levels, which also happens to be one of the most densely populated localities in Bogotá.

Among the references consulted, there was no study that combined geostatistical tools and ML to identify specific zones in which cases of RD may occur due to air quality. Therefore, an ecological case study was conducted by applying these tools in a specific analysis of Kennedy, setting out to determine the influence of meteorological variables and atmospheric pollutants on the population's health, establishing not only the most relevant variables, but also the zones of greatest interest in geographic spaces based on the locality's characteristics. Applying geostatistical tools and ML in an environmental health study on an atmospheric component is innovative. Furthermore, the local scale of the analysis is emphasized, in addition to data collection in the field being one of the inputs to feed the ML model. This is the first study of this nature developed in one of the most populated zones of a city such as Bogotá. The approach created in this study can be applied to different territories, particularly densely populated areas with high air pollution levels. Different municipalities can anticipate environmental health situations and reduce the cost of RD treatments by applying these tools.

MATERIALS AND METHODS

This study consists of three sequentially phases which are described below (see Fig. 1), namely: study area analysis, exploratory analysis of air quality and RD, and the forecasting model with ML techniques and geographical information system (GIS).

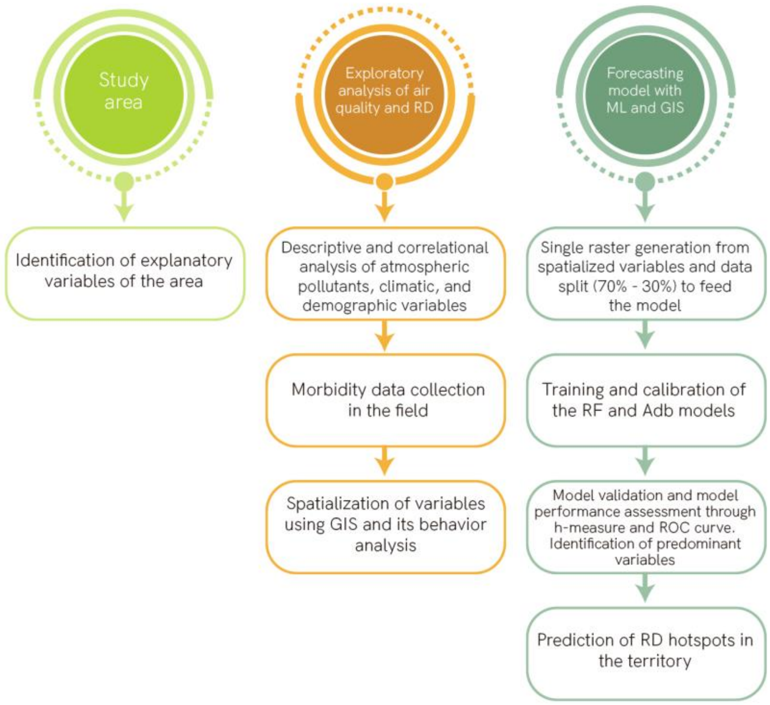


Fig. 1 Methodological framework to identify and forecast RD hotspots.

STUDY AREA

The study area is Kennedy, located in southwest Bogotá, Colombia. It covers an area of 38.6 km^2, of which, 1.9% is green space and 10% is protected. Kennedy is located in a transition zone between the eastern plateau and mountains. It is a flat zone which borders five Bogota localities (Puente Aranda, Fontibon, Bosa, Ciudad Bolívar and Tunjuelito), and is where some of the city's main industrial, mining, and commercial activities are located. It also borders Mosquera, Cundinamarca. Wind in Kennedy predominately comes from southwest Bogotá at speeds of 2.2 to 2.5 m/s. The average temperature is 14.6 ± 0.4 ºC, with its highest recorded values in 2016 (14.9 ± 0.8 ºC). Moderate thermal inversions are common, primarily in the dry months. With respect to precipitation, low values have been recorded in this area of the city, with cumulative averages for the period 2009 – 2016 between 483 and 1018 mm, with a multi-annual average precipitation of nearly 767 mm (SDA 2017).

This locality is made up of twelve zoning planning units (ZPU), which act as territorial units for urban development planning at the zonal level: Américas, Bavaria, Calandaima, Carvajal, Corabastos, Castilla, Gran Britalia, Kennedy Central, Las Margaritas, Patio Bonito, Tintal Norte and Timiza. Four of the above are for residential urban land use,[1] three are for residential use in incomplete urbanization zones,[2] two are in the developing stages,[3] one is the urban center of the locality,[4]

[1] The use changes are occurring in predominately residential sectors with an increase of unplanned territorial occupancy.
[2] Strata 1 & 2 non-consolidated peripheral residential sectors with deficiencies in their infrastructure, accessibility, equipment, and public space.
[3] Under-developed sectors with large unoccupied lots.
[4] Consolidated sectors that have urban centers with the dominating residential use having been displaced for uses that encourage economic activities.

and two are allocated for public use[5] (SDP 2018). There are units registered for industrial development activities in the Américas, Carvajal and Bavaria ZPUs (Galindo 2013), and approximately 47.2% of households are located near industries (DANE 2018).

Roadways with high vehicular traffic cross the locality, such as Avenida Boyacá, Avenida Ciudad de Cali, and Avenida de las Américas (see Fig. 2), on which light, cargo and public transportation vehicles represent the main traffic. Predominantly type 1 and 2 roadways cross and border the locality, as part of the arterial road system. These roads support traffic flows caused by the inter-urban transport of goods and people. Due to their length and characteristics they support traffic caused by mass public transportation and connect to the local road network. Type 1 roads are 60 m wide, while type 2 are 40 m wide. Furthermore, there are local roads within the study area with widths ranging from 4 to 22 m that facilitate entry and local traffic caused mainly by individual transport vehicles. The locality is also bordered by the Autopista Sur highway and Calle 13, whose road accesses to the city were used by an average of 5300 – 6600 trucks in 2017, with a daily average of 12,000 different types of vehicles going to the CORABASTOS supply center, according to government entities in Bogotá.

The predominant buildings in the locality are made of concrete, cinder blocks and bricks, whose heights mainly range from 1 to 5 floors; 67% correspond to buildings up to 3 floors (8.1 m) high, with 18.5% of buildings having 4 and 5 floors (up to 13.5 m), and some buildings are taller than 37 m (DANE 2018).

Kennedy has an estimated 1,208,980 inhabitants according to the population census (SDP 2018). In accordance with a 2017 analysis of its population structure, 53.7% of its population are adults, while the early childhood and adolescent population groups have a smaller representation. Furthermore, the overall population rate in the labor market in Bogotá is approximately 60.8%[6] (SDP 2018).

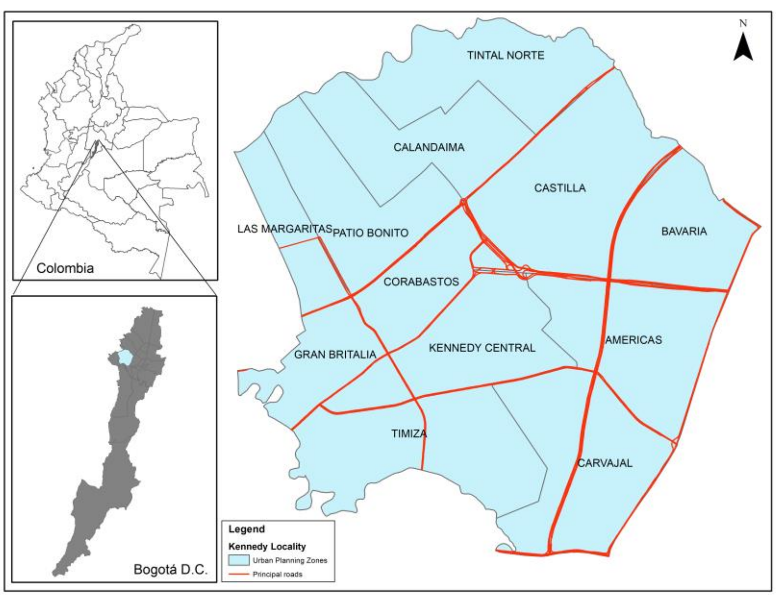


Fig. 2. Locality of Kennedy and its location in Bogotá, Colombia. The twelve ZPUs that make up the internal distribution of Kennedy along with its main roadways are displayed.

[5] Large areas allocated to produce urban and metropolitan equipment.

[6] The working age population is 12 years and older in the urban zone, which for Kennedy corresponds to 1,019,894 people.

Air quality and its effects on health

Bogotá has an Air Quality Monitoring Network (AQMN), which is comprised of thirteen monitoring stations in the city's urban area, two of which are located in Kennedy (see Fig. 3). The first (Kennedy station) is situated in a residential zone and monitors PM_{10}, $PM_{2.5}$, NO, NO_2, CO, SO_2, as well as meteorological variables including humidity, barometric pressure, solar radiation, temperature, precipitation, wind speed and direction. The second (Carvajal station) is located in a residential zone with a presence of industrial activity. It is an industrial-traffic station that monitors PM_{10}, $PM_{2.5}$, NO, NO_2, NOx, O_3, CO, SO_2 and meteorological variables such as precipitation, temperature, wind direction and speed.

This study analyzed data from the two stations mentioned above, as well as the Tunal, Puente Aranda and Centro de Alto Rendimiento stations (see Fig. 3), located in Kennedy's influence area, which are the traffic, industrial and background stations, respectively. These automated stations monitor the same parameters mentioned above as the stations in Kennedy. Furthermore, the Mosquera-Sena manual station is located in the Bogotá Savanna, and measures PM_{10}, SO_2 and NO_2. The Centro de Alto Rendimiento station is located in a zone with a low concentration of pollutants, where winds from all directions converge, and has historically recorded low pollution levels. The Tunal and Puente Aranda stations are located in zones with high traffic and industrial activities. The Tunal station receives winds from the south, while the winds that hit the Puente Aranda station come from the west and northwest (SDA 2017).

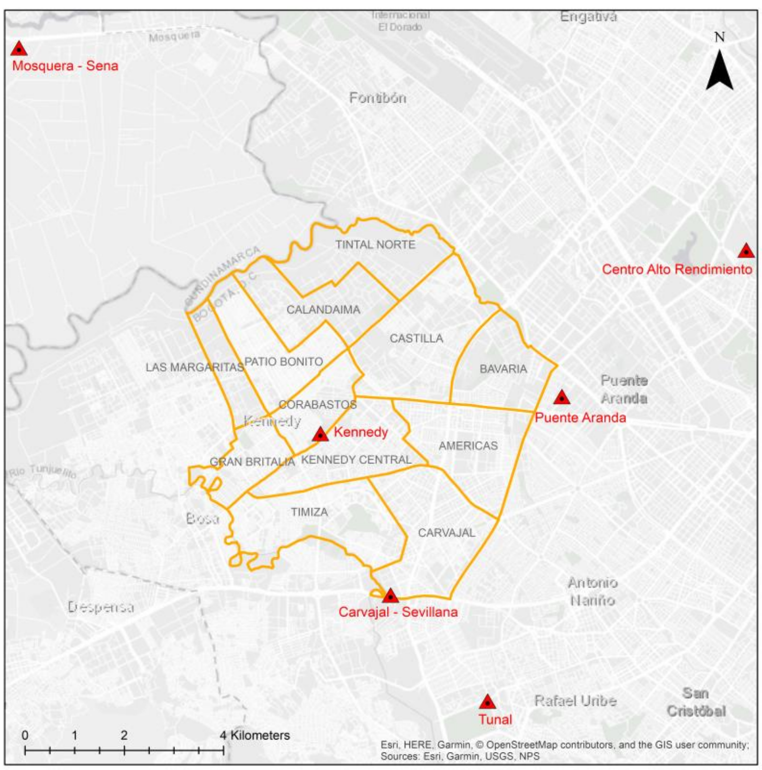


Fig. 3. Location of the monitoring stations in the study site and the area of influence.

In 2015, the Carvajal and Kennedy stations had the highest concentrations of PM_{10} and $PM_{2.5}$ in the city. In the first quarter of 2019, environmental emergencies were declared in areas monitored by these two stations. These emergencies occurred as the result of a variation in meteorological conditions and the intensification of the temperature inversion phenomena during the dry season, which generally occurs in the city with moderate effects and breaks atmospheric stability between 7:00 – 9:00 a.m. From 2011 – 2015, neither station met the national regulation standard and recorded yearly concentration averages that were among the highest for stations that monitor PM_{10} and $PM_{2.5}$ in the country, which met the temporal coverage criterion of 75% (IDEAM 2016). NO_2, O_3, CO and SO_2 pollutants did not exceed the limits established in the regulation. However, SO_2 did have higher concentrations in the monitoring stations situated in the locality. The presence of different types of industrial activities in the city and in neighboring municipalities, as well as road and traffic conditions with different types of vehicles, all contribute to the increased concentration of atmospheric pollutants in Kennedy. It is important to note that in Bogotá, prevailing winds from the northeast and southeast displace particulate matter towards the west (Ramírez et al. 2018).

EXPLORATORY ANALYSIS

Work began on an exploratory analysis of the variables, which in terms of air quality, could impact the population's health in Kennedy. A descriptive analysis was conducted of the spatial distribution of atmospheric pollutants and meteorological variables for the period 2009 – 2017, as well as descriptive analyses of individual records from health care providers (RIPS, as per its Spanish acronym) in Kennedy, reported by the District Health Secretariat (DHS) for the same period, for diseases associated with air quality, in accordance with Version 10 of the International Classification of Diseases (ICD). Furthermore, a bivariate Pearson correlation of the continuous variables (pollutants, climatology, and demographic variables) was conducted for 2016. The data on atmospheric pollutants and meteorological variables was determined based on a weighted average calculated by the ArcGis 10.5.1 software, using information from the AQMN stations (Carvajal, Kennedy, Puente Aranda, Centro de Alto Rendimiento and Tunal), as well as the Mosquera-Sena station in Cundinamarca, which are in the locality and its boarding zones (see Fig. 3).

The spatial distribution of PM_{10}, $PM_{2.5}$, CO, NO_X, and SO_2 pollutants, as well as the precipitation and temperature meteorological variables were analyzed via the deterministic method for interpolation called inverse distance weighting (IDW) interpolation. This is a univariate interpolation method, which is useful in evaluating small study areas. To generate a predictive surface, the value taken by an unknown point is influenced more by nearby sampled data, than by data from areas further away (Ly et al. 2011). This method does not consider spatial groupings and has better results when the sampled data comes from irregularly spaced locations (Li and Heap 2014). This is the case of Bogotá, which has approximately one station every 23 km^2. Given that the information for this study comes from irregularly distributed monitoring points (see Fig. 3), this study contemplated examining the influence of the data recorded at the stations concerning its surrounding areas. The spatial behavioral analysis of the data was performed via the natural grouping data classification method proposed by Jenks (Jenks 1967).

FORECASTING MODEL WITH ML AND GEOGRAPHICAL INFORMATION SYSTEM (GIS)

Medical consultation records from RIPS do not contain information on the spatial location of health care service users. Consequently, using data provided by the DHS to spatially identify the zones of the locality with possible RD due to the presence of atmospheric pollutants was not possible. As such, the decision was made to develop the field work by conducting a survey on health perception, identifying individuals from households in the locality who have been diagnosed with a RD[7] in 2016. To this end, a georeferenced primary source data collection instrument was applied, which considered socio-demographic variables and the surveyed person's perception of their health condition. Through a structured questionnaire, the survey developed which consisted of thirty-one questions, was conducted with households in the locality. This instrument was applied in twelve ZPUs in 2017 in accordance with the sample size established by the study.

The required sample was established to conduct the surveys based on the 2016 number of inhabitants in each ZPU (SDP 2018). The equation for finite populations was used with the following criteria: error: 4%; confidence level: 96%; and positive and negative variables: 50%. This equation was applied to the general population of the locality yielding a result of 656, which was distributed in accordance with the population proportion of each ZPU. During the development of the study, the sample size grew to 912 surveys, thus expanding its spatial coverage (see Fig. 4).

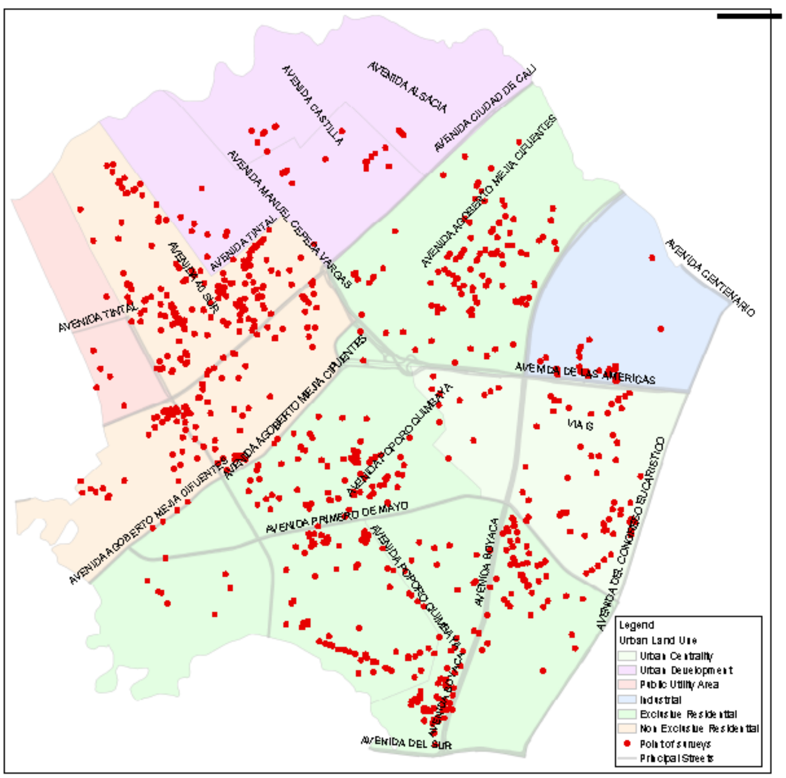


Fig. 4. Field data collection points in the Kennedy ZPUs, primary roadways, and land use typologies

Design and development of the forecasting model

Health risks arise from a combination of socio-economic factors, environmental conditions, habitat, and individual behavior. Geospatial and ML tools were applied to identify the areas of greatest interest related to the population's respiratory health, in contrast with the presence of pollution sources, pollutant distribution, and the exposed population. The ArcGis 10.5.1

[7] According to the ICD, RD range from rhinopharyngitis, known as the common cold, to respiratory disorders in diseases classified elsewhere (J00-J99).

software and the open source software R 3.5.2 (CRAN 2018) were used for this purpose. Information was entered on atmospheric pollutants (PM_{10}, $PM_{2.5}$, CO, NOx, SO_2), meteorological variables, precipitation and temperature previously determined by the IDW method for 2016, as well as data on population, population density, households' proximity to roadways (type 1:T1 and type 2:T2), and land use typology. According to Salam et al. (2008) and Li et al. (2011), the proximity of households, located between 100 m – 1000 m to local and main roads, may increase the risk of RD. This study considered households' distance to primary and secondary roads (T1 and T2, respectively).

Geocoded information was also entered and arranged in a GIS of the set of categorical responses from the individuals surveyed to the question, "In the last year (2016), have you or any of members of your household been diagnosed by a doctor with a respiratory disease or infection such as asthma, pneumonia or severe lung disease?" A single raster was created with the resulting information, in which the analyzed variables (continuous and categorical) were overlaid and then used as input information for the R software. This process provided information on the twelve explanatory variables and the categorical answer for each point in the locality. RF and the AdB algorithm were the ML tools used, both of which improve the accuracy of single decision tree classifiers by combining trees grown (Breiman 2001). These tools maintain a bias-variance trade off trough bagging or boosting methods. It is important to note that ANN and SVM are also useful in classification tasks. However, collinearity of variables is a condition that limits the accuracy and generalization capacity of ANN (Kuhn and Johnson 2016). Furthermore, the proximity between classes in the geographical space limits the accuracy of SVM. RF and AdB perform better in those aspects and facilitate the analysis of variables distributed in space, which is useful in the integrated and spatial analysis of possible health risk factors.

RF are one of the most accurate bagging methods. RF are a consistent classifier in collecting tree-structured classifiers $\{h(x,(\Xi)k), k = 1,...\}$, in which $\{(\Xi)k\}$ are independent random vectors identically distributed, with each tree issuing a single vote for the most popular class in the x input (Breiman 2001). For categorical predictions, the voting process selects the class with the most votes (Kuhn and Johnson 2016). RF can handle large numbers of features (Ivanov et al. 2018) and identify the most important variables for the model. The precision of RF depends on the strength of the individual classifiers and the dependence measure between them (Breiman 2001).

The model used 70% of the data for training and 30% for testing. The partition was performed by randomly considering the proportionality between affirmative and negative responses. A forecast was created of the areas with the strongest confluence of affirmative responses to the possibility of RD cases by a majority vote, resulting in the classification that determined the most influential variables in the model and the distribution of response data according to the conjugate of the predictor variables in the classification with RF. The model calibration included an iteration of 300 – 1500 trees, with every 100 trees establishing the best combination with the number of variables, according to the accuracy results and the Kappa index.

Subsequently, the AdB algorithm, which has no random elements and uses decision trees as the model base, was applied to create a strong classifier (an ensemble of trees) built from weak classifiers by successively reweighing them (Breiman 2001). AdB is one of the most widely-used boosting methods in which each classifier focuses on the data that was erroneously classified by its predecessor, in order to adapt the algorithm and generate better results with each iteration and reduce the generalization error (Schapire and Freund 2012; Rokach and Maimon 2015). In this method, each constructed tree depends on its predecessor's trees and the prediction come from the most frequent selected class. The samples that are incorrectly classified in the iteration are given more weight than the samples correctly classified. Therefore, samples that are difficult to classify are given greater weight until Adb identifies the best model (Kuhn and Johnson 2016). In this study, the same input parameters for the RF model were used.

By using the *Mean Decrease Accuracy* tool, the variables with the greatest influence on the classification error were determined for each model. Subsequently, forecasts were made with 100% of the spatial behavior data according to possible RD cases in the locality. The H-measure and the classification error from the receiver operating characteristic (ROC) curve were used as the performance indicators. The H-measure is a measurement of the loss from erroneous classification contingent on the relative proportion of the objects belonging to each class (Hand 2009). The ROC curve enables a comparison of the accuracy and precision of the representing model for each threshold value. This curve is a plot showing all the sensitivity and specificity pairs resulting from the continuous variation of cutoff over the entire range of observed results (Altman and Bland 1994). Furthermore, as a function of sensitivity and specificity metrics, the area under the ROC curve (AUC) is insensitive to disparities in class proportions. A perfect model separates the two classes with sensitivity and specificity values of 100% (Kuhn and Johnson 2016). Therefore, sensitivity and the specificity metrics of the diagnostic test, as well as the AUC closest to 1, in the 0.5 – 1.0 interval, represents greater accuracy than the discriminant test (Del Valle 2017). This area establishes the probability that a random person with the disease has a higher measurement value than a random person without the disease (Altman and Bland 1994).

The "Geographic data analysis and modeling" raster and the "Bindings for the 'Geospatial' Data Abstraction Library" rgdal were the packages used in the R software to read and process the raster images. "Breiman and Cutler's random forests for classification and regression" were used to design and develop the RF model. The "C_lassification _A_nd _RE_gression _T_raining" caret was used to determine the most optimal model parameters. The "Visualizing the performance of scoring classifiers" ROCR was used to calculate the AUC and display the ROC curve. These packages made it possible to adjust the spatial information to the databases adapted for statistical and predictive processing. A computer with Core i5 8th generation processor, 8Gb RAM and 1Tb hard disk was used.

RESULTS

The locality of Kennedy is characterized by its location between primary roadways and the diversity of economic activities carried out in the same. It has gone through different changes as it is one of the most densely populated localities in Bogotá. By using the IDW method for the period 2009 – 2017, a decreasing trend was found in the concentration of different pollutants with values ranging from: (78.72 – 53.11 ug/m^3) for PM_{10}; (35.09 – 24.32 ug/m^3) for $PM_{2.5}$; (1.2 – 0.73 ug/m^3) for CO; (64.15 – 40.46 ppm) for NO_x; and (8.69 – 2.77 ppb) for SO_2. The largest values mainly occurred in 2009, which also had the largest number of consultations associated with RD.

The PM_{10} and $PM_{2.5}$ values surpassed those established by WHO guidelines (PM_{10}= 20 μg/m^3; $PM_{2.5}$= 10 μg/m^3). According to the WHO (2006), these are the lowest levels that demonstrate, with more than 95% confidence, that total cardiopulmonary and lung cancer mortality increases in response to prolonged exposure to $PM_{2.5}$. However, values were recorded in 2017 that were close to WHO guideline values according to which, the risk of premature mortality is reduced by 6% compared to the severe level; (PM_{10} =50 µg/m^3; $PM_{2.5}$= 25 µg/m^3) (WHO 2006). The SO_2, CO and NOx values indicate a reduction of pollutants. SO_2 did not surpass standards (30 ppb) set by the Environmental Protection Agency (EPA). In the case of CO, there is no yearly standard, yet the values recorded at the monitoring stations did not, at any time, exceed the Colombian standard (5000 ug/m^3), nor the EPA standard (9000 ug/m^3) for 8 hours of exposure. NO_X, which is an unregulated pollutant and ozone precursor, decreased by approximately 37% compared to the analyzed periods.

In the period covering 2009 – 2017, after the common cold, chronic obstructive pulmonary disease (COPD), acute bronchitis, and unspecified asthma were the most common RDs for which different patients went to consultations. Consultations ranged from: (1493 – 5744) for COPD; (2294 – 4736) for acute bronchitis; and (1380 – 2573) for unspecified asthma.

Correlation Analysis

A matrix was created by applying Pearson's method, which demonstrates high correlation between the 2016 climatic variables and atmospheric pollutants analyzed; precipitation and temperature have an inverse relationship (see Fig. 5). Moreover, in Fig. 5, the relationships with no significance are marked with an X, that is, their parameter *p-value* is greater than 0.05.

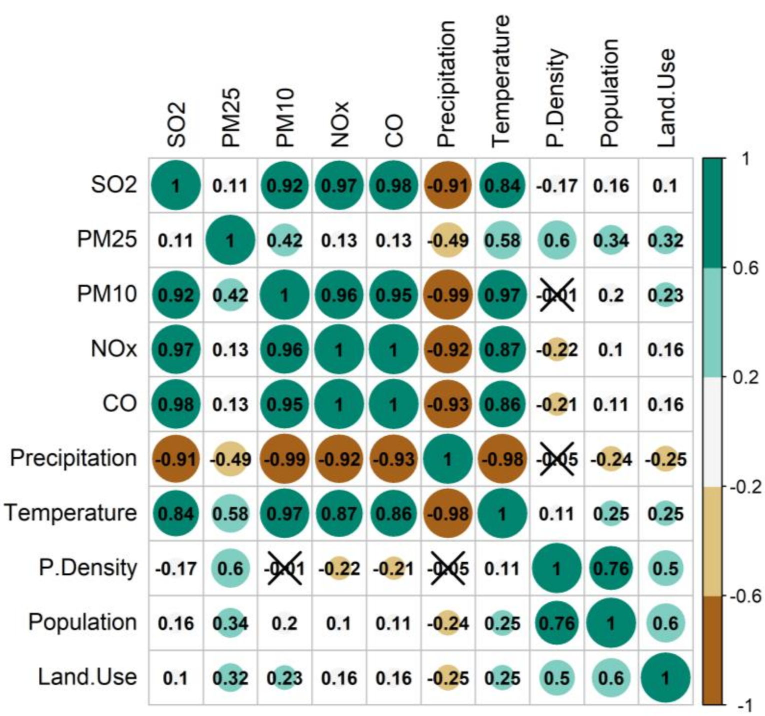


Fig. 5. Correlation matrix of pollutants, meteorological and demographic variables in 2016. The color scale on the sidebar shows the degree of positive (0 – 1) or negative (0 – -1) correlation between the variables.

Areas of interest for possible cases of respiratory disease

In 2016, the largest concentration of pollutants analyzed were found in the Carvajal and Timiza ZPUs (see Fig. 6). However, the highest concentrations of $PM_{2.5}$ were in the areas of the Corabastos, Kennedy Central, Carvajal, Patio Bonito, Calandaima, Margaritas, Gran Britalia and Timiza ZPUs in the center and western zones of the locality. Temperature had a constant behavior (14.98°C), with its lowest values found in the eastern zone of the Bavaria ZPU (14.7°C), which has larger precipitation values (769.3 mm) with respect to the rest of the study area (712.42 mm on average). The smallest precipitation values were found in the Carvajal and Timiza zones, with 620.6 mm and 637.2 mm, respectively.

In total, 912 surveys were conducted in the twelve ZPUs that make up the locality. The Patio Bonito, Carvajal, Kennedy Central and Castilla ZPUs had the largest number of affirmative responses to the questions asked in the field work (see Fig. 6); 21.4% of the individuals surveyed indicated that a member of their household was diagnosed with a RD, of which 51.8% corresponded to the working age population to 60 years old, 23.6% were young people between 5 – 14 years old, and 14.3% were people older than 60. Furthermore, it was found that 49.2% of respondents had lived in the study area for more than ten years. These indicative figures are comparable with those reported by DHS in 2016, given that in Kennedy nearly 27% of RD cases in children under 14 years of age were attended to in emergency rooms, and a prevalence of wheezing was reported in 12.6% of adults over 60 years old (SDS 2019).

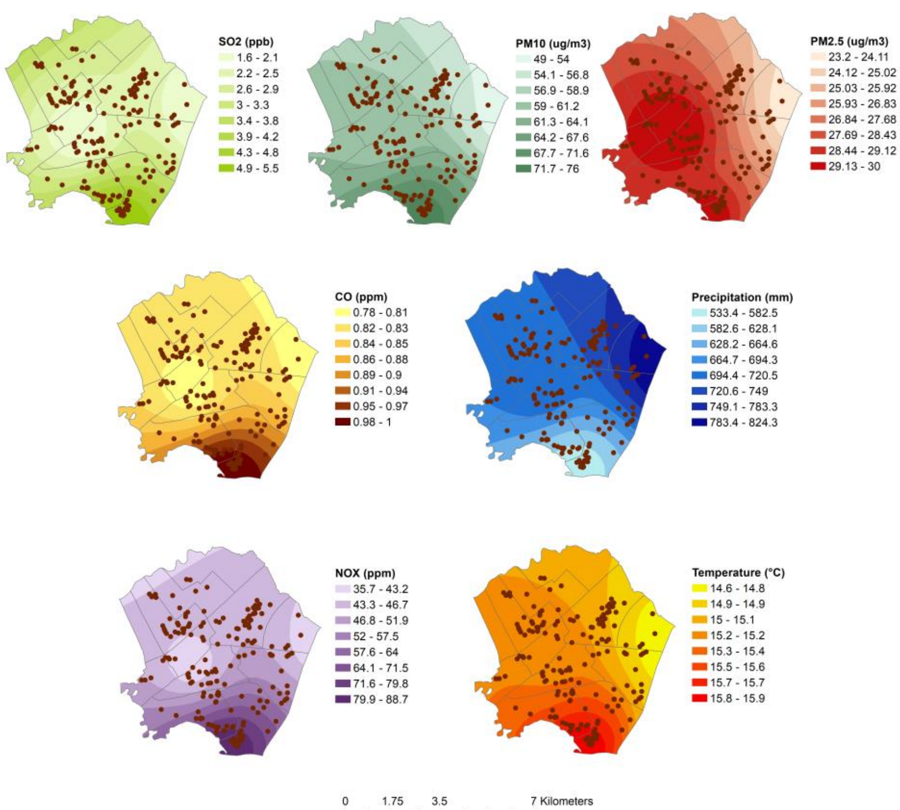
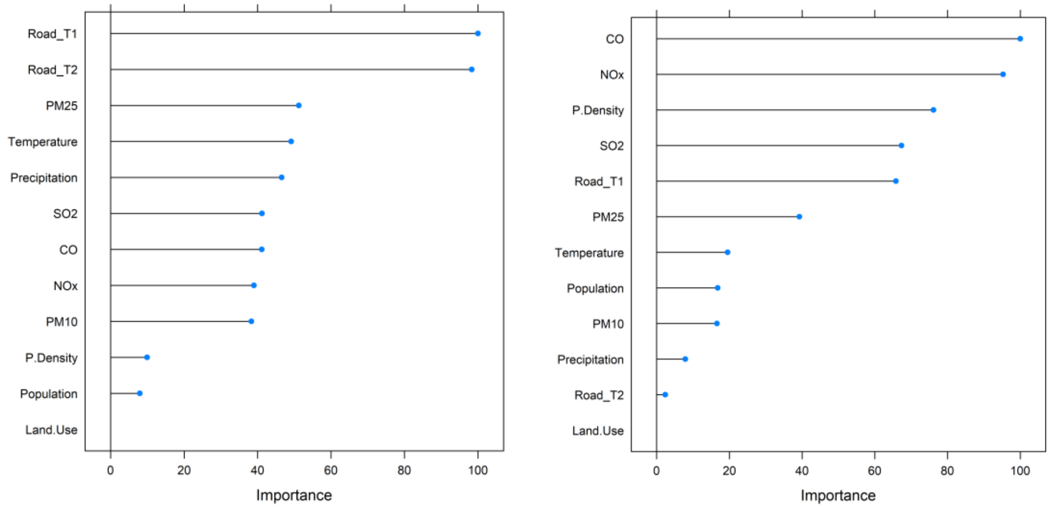


Fig. 6. Behavior of pollutants, meteorological variables, and field work results for 2016

Forecasting model with ML and GIS

Importance matrixes were created (see Fig. 7). The household proximity to roads (T1 and T2 in Fig. 7) variable had the strongest influence on the RF model, followed by temperature and $PM_{2,5}$. In the AdB model, household proximity to roads was the fifth most important variable. Variable behavior in the model is consistent with respect to the behavior recorded in

2016. The population-related variables are the least important in the RF model, while population density (P. Density in Fig. 7) plays an important role in the AdB model.



(a) (b)

Fig. 7. Hierarchization of predictor variables in the (a) RF and (b) AdB models

With respect to the models' performance, the RF model generated an AUC of 0.63 (see Fig. 8), in which the largest value was achieved through a model with 500 trees and 12 variables, which stabilized the error and prevented overfitting, resulting in an H measure of 0.10. The AdB model had an AUC of 0.52, for an H measure of 0.018.

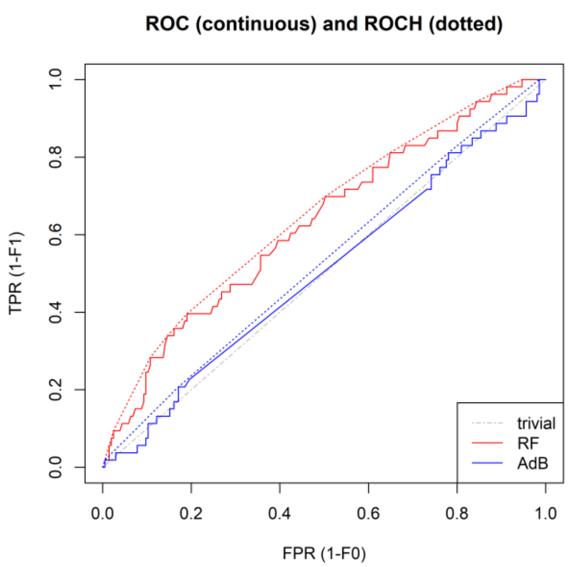


Fig. 8. ROC curve for RF and AdB

Forecasting zones with possible RD events

Based on the behavior of the variables introduced in the RF model, the most relevant zones in the locality related to exposed elements and external risk factors are Patio Bonito and Calandaima. Furthermore, considering the confluence of the most important variables in the RF model (proximity to T1 and T2 roads), the behavior of meteorological variables, and pollutants associated with both road quality and mobile source combustion ($PM_{2.5}$, CO, NO_x), there are hotspots present in each ZPU, which, depending on their intensity, enable the occurrence of possible RD cases (see Fig. 9).

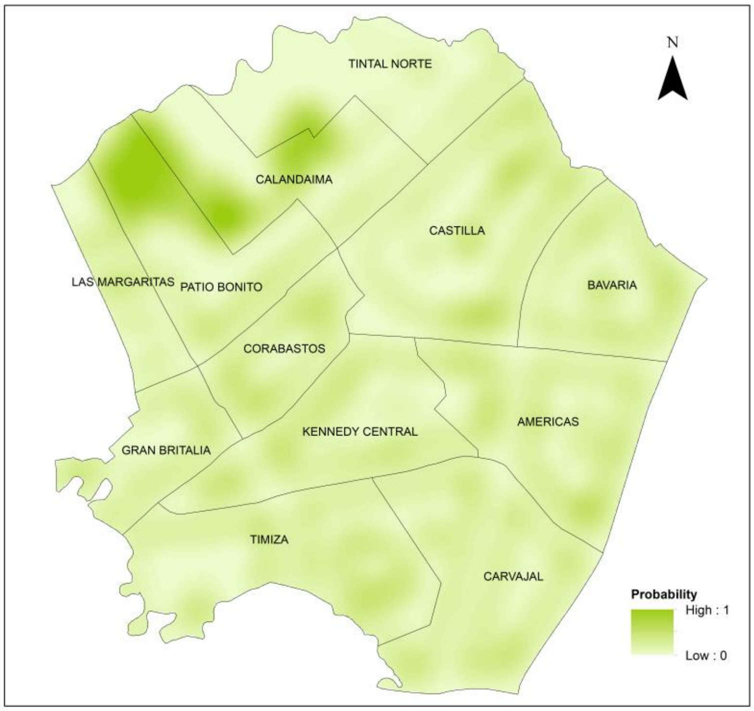


Fig. 9. Forecast of zones in which cases of RD could have occurred based on the RF model.

DISCUSSION

By applying ML tools, it was found that RF outperformed the AdB model for the H-measure, AUC, and accuracy (77.5% RF; 24.4% AdB). RF had a better odds ratio of 2.25, which reflects a sound diagnosis capacity and greater precision for the specific classification of possible cases of RD. In the same model, forecasting areas with high densities of possible cases of RD correlates with proximity to roads, $PM_{2.5}$, CO, SO_2 and meteorological variables. The high intensity classification of areas in Patio Bonito, Calandaima and specific points in Carvajal, Timiza, Kennedy Central and Castilla are supported primarily by the population's exposure to risk factors, including primary roadways on which different mobile sources of air pollution transit. However, there are other factors that influence RD events. For example, the low sensitivity of the model (11%) may be explained by the lack of variables that more accurately describe the behavior of the recorded events. This also responds to the influence exerted by important variables, as determined by their behavior.

It is worth noting that given their spatial behavior, in addition to identifying patterns of pollutant concentrations (Habibi and Alesheikh 2017), or forecasting individual contaminants (Ivanov et al. 2018), it is necessary to consider the spatial behavior of combined risk factors and relate this behavior with the exposed population. This fact is sustained in the multiple pollutant exposure phenomena, which has not been addressed in many studies (Billionnet et al. 2012), as well as the proximity of the population to pollutants (Mazenq et al. 2017; Yu et al. 2019). This study gathered these experiences and made progress towards an innovative method of identifying zones where possible cases of RD may occur through the combined use of geostatistical tools and ML.

In this order of ideas, interpolation via IDW established the behavior of atmospheric pollutants and the meteorological variables, which were consolidated in the information bases for the correlation matrix and ML models. This aspect is consistent with prior experiences supported in studies developed by Gorai et al. (2018), Habibi and Alesheikh (2017), and Sajjadia et al. (2017), in which the transformation of information from observed points to continuous information was carried out to compare spatial behavior patterns. In different cases, atmospheric pollutant behavior is the primary variable for the analysis of its effects on a population's health, which is consistent with this study that used ML and found that $PM_{2.5}$ was one of the most relevant variables.

Moreover, when information is collected in the field, conducting surveys has a related bias effect, such as selection bias. To reduce this effect, the survey was carried out in Kennedy's twelve ZPUs based on the sample, distributed by ZPU and land use. However, due to security concerns, entering some zones was difficult, which hindered the completion of the total number of surveys. This was the case in the Las Margaritas and Calandaima ZPUs. Furthermore, in the field data review process, it was found that due to spatial effects, some regions were not covered. As such, the number of surveys in these zones was increased to 912 for the final sample size value. Conducting a survey made it possible to identify possible respiratory system-related morbidity events in a spatial manner. Therefore, related biases may be limited in future uses with spatial location information that is recorded at health entities, and for security reasons, is not shared with the public.

Not having complete information of atmospheric pollutants and meteorological variables was one of the study's constraints. The air quality monitoring and tracking protocol (MAVDT 2010) establishes a minimum data validity standard of 75%. That is, data whose information is at least 75% complete for the period analyzed is considered valid. Of the data used, an average of 78% met the temporal validity parameter. The NOx and SO_2 data represent 40% of the data that did not meet this requirement. $PM_{2.5}$ and CO had values of 38% and 14%, respectively. However, given the need to have information to feed the ML model and generate a weighted average of pollutant behavior and the meteorological variables for each year, the recorded information was evaluated in terms of its data trend in an analyzed time series based on the standard deviation of the variable for the pollutants that did not meet the required validity percentage in a given year and monitoring station. Thus, the data used exhibited a behavior recorded in its trend, which is largely in line with the required quality standard.

Detecting hotspots through spatial analysis with geostatistical and ML tools is useful to establish measures to reduce the vulnerability of people who are exposed to different health risk factors. Moreover, this approach facilitates the identification of important variables for the model, which is a prioritization tool. Nonetheless, due to different factors that influence people's health, the model could be strengthened through more available information to refine the characterization of the study area. This study's approach is useful as a support mechanism for urban planning projects, including the evaluation of territories' sustainable development performance. This approach could be applied in other fields to identify potential areas of interest, such as the agriculture sector to identify suitable soils, earth that is ready for the sowing of future crops, or to detect possible polluted soils due to different activities.

CONCLUSIONS

- This research developed a tool based on ML that presents the necessary stages to forecast hotspots in which possible RD cases may occur, based on the behavior of a territory's characterizing variables. Its application in a densely urban area is useful and replicable as it is a common characteristic in certain territories in developing countries. The micro-territorial nature of the study is relevant and innovative, as it differs from capital city and country approaches. This approach also enables researchers to generate useful technical support data for early warnings and contingency plans to mitigate impacts on air quality and population health, which also influences territories' economies.

- Using open-source software such as R and spatialization by means of open-source ML codes makes this study an easily replicable tool. These tools become stronger as more specific and spatialized information becomes available, and their advantages strengthen environmental health governance by public entities and the academic sector.

- The level of importance of pollutants such as $PM_{2.5}$, CO, SO_2 and meteorological variables influences the ML model's behavior. Relevant variables regarding the characteristics of the study area include: high vehicle flow of fossil fuel-powered automobiles (which explains the level of importance of the $PM_{2.5}$, CO and SO_2 variables); non-standard operating conditions; deterioration of local roads with the consequent generation of resuspended material; and residential areas with high population densities that are grouped together, where mixed land uses are integrated with commercial, industrial and service provision activities. It is necessary to continue carrying out detailed studies on the exposed population that observe factors such as dose, duration, form of contact, age, sex, diet, personal characteristics, lifestyle and health condition, in order to determine the relative risk and establish these factors' behavior in the study area. In this study, 21.4% of the individuals surveyed reported having been diagnosed with respiratory diseases, of which 14.3% were individuals over 60 years of age, 51.8% are working-aged individuals, and 49.2% of those surveyed stated that they had lived in the study area for more than 10 years, demonstrating that exposure time is another variable of interest. These indicative figures include the broad spectrum of respiratory diseases, from the common cold to chronic and acute respiratory system diseases.

- Different areas reflect the confluence of risk factors and exposed elements. As such, the RF model established that an area of great interest could be in the Patio Bonito and Calandaima ZPU's. However, the residential characteristics of the Timiza, Kennedy Central and Carvajal ZPUs draw attention to the exposed population. RF perform better in terms of a model driver (AUC: 0.63; H measure: 0.1; accuracy: 77.5%), meaning that the results generated by the RF model are more accurate than those generated by an AdB model. Similarly, it can be concluded that it is possible to replicate this model in other areas or municipalities, and its accuracy can be improved by introducing specific data on the location with the highest exposure of patients attended to in consultations, emergency room visits and hospitalizations related to RD, as well as information on the explicative variables for the analysis period. The combination of the tools applied in this study together with a pollutant dispersion model could increase the AUC, as well as the model's classification metrics.

- Lastly, it must be stressed that sustainable development refers to an increase in quality of life, through the interaction of social, environmental, and economic dimensions for equitable, livable, and viable development. A model of these characteristics becomes a preventive tool, which can contribute to reducing costs by addressing events associated with air pollution. As a territorial planning component, determining the influence of air pollution on a territory's sustainability can contribute to implementing policies instituted in the international framework. As air pollution increases, so does the number of workdays lost, reducing productivity. A better understanding of this phenomenon could contribute to zonal planning and determining the territorial organization of each zone.

Acknowledgments

Many thanks to the members of the Intelligence and Territorial Analysis Group of the Universidad Santo Tomás for their collaboration in conducting the fieldwork.

References

Altman Douglas G, Bland J. Martin (1994) Diagnostic Tests 3: Receiver Operating Characteristic Plots. BMJ 309 (6948):188. https://doi.org/10.1136/bmj.309.6948.188

Billionnet C, Sherrill D, Annesi-Maesano I (2012) Estimating the health effects of exposure to multi-pollutant mixture. Ann Epidemiol 22:126–141. https://doi.org/10.1016/J.ANNEPIDEM.2011.11.004

Bobb JF, Valeri L, Claus Henn B, et al (2015) Bayesian kernel machine regression for estimating the health effects of multi-pollutant mixtures. Biostatistics 16:493–508. https://doi.org/10.1093/biostatistics/kxu058

Borja-Aburto VH (2000) Ecological studies. Salud Pública de México 42:533–538.

Breiman L (2001) Random Forests. Machine Learning 45:5–32. https://doi.org/10.1023/A:1010933404324

CRAN Comprensive R Archive Network (2018) R-3.5.2 for Windows (32/64 bit). https://cran.r-project.org/bin/windows/base/old/3.5.2/. Accessed 10 June 2019

DANE National Administrative Department of Statistics (2018) Multi-Purpose Survey -MS 2017. Bogotá, Colombia.

Del Valle Benavides AR (2017) ROC curves (Receiver-Operating-Characteristic) and their applications. Universidad de Sevilla.

DHS (2019) SALUDATA- Health Observatory of Bogota http://saludata.saludcapital.gov.co/osb/index.php/datos-de-salud/salud-ambiental/consultaurgencias14anios/. Accessed 11 April 2019

Franceschi F, Cobo M, Figueredo M (2018) Discovering relationships and forecasting PM_{10} 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

Galindo WG (2013) Construction dynamics by use, the locality of Kennedy 2002/2012. Bogotá.

García-Ubaque JC, García-Ubaque CA, Vaca-Bohórquez ML (2011) Medical consultation in productive age population related with air pollution levels in Bogota city. Procedia Environ Sci 4: 165–169. https://doi.org/10.1016/j.proenv.2011.03.020

Gorai AK, Tchounwou PB, Biswal S, et al (2018) Spatio-Temporal Variation of Particulate Matter ($PM_{2.5}$) Concentrations and its health impacts in a mega city, Delhi in India. Environ Health Insights 12:1-9. https://doi.org/10.1177/1178630218792861

Habibi R, Alesheikh AA, Mohammadinia A, et al (2017) An assessment of spatial pattern characterization of air pollution: A case study of CO and $PM_{2.5}$ in Tehran, Iran. ISPRS Int J Geo-Inf 6:270. https://doi.org/10.3390/ijgi6090270

Hand DJ (2009) Measuring classifier performance: A coherent alternative to the area under the ROC curve. Mach Learn 77:103–123. https://doi.org/10.1007/s10994-009-5119-5

Hernández B, Velasco-Mondragón HE (2000) Cross-sectional surveys. Salud Pública de México 42: 447–455

Huang K, Xiao Q, Meng X, et al (2018) Predicting monthly high-resolution $PM_{2.5}$ concentrations with random forest model in the North China plain. Environ Pollut 242:675–683. https://doi.org/10.1016/j.envpol.2018.07.016

IDEAM Institute of Hydrology,Meteorology and Environmental Studies (2016) State of air quality in Colombia, 2011 – 2015 Report. Bogotá D.C.

Ivanov A, Voynikova D, Stoimenova M, et al (2018) Random forests models of particulate matter PM_{10}: A case study, in: AIP Conference Proceedings 2025, 030001. https://doi.org/10.1063/1.5064879

Jenks, George F (1967) The Data Model Concept in Statistical Mapping. International Yearbook of Cartography 7: 186-190

Kami JA (2019) A random forest partition model for predicting NO_2 concentrations from traffic flow and meteorological conditions. Sci Total Environ 651:475–483. https://doi.org/10.1016/j.scitotenv.2018.09.196

Kassomenos P, Petrakis M, Sarigiannis D, et al (2011) Identifying the contribution of physical and chemical stressors to the daily number of hospital admissions implementing an artificial neural network model. Air Qual Atmos Health 4:263–272. https://doi.org/10.1007/s11869-011-0139-2

Kestenbaum B (2019) Epidemiology and Biostatistics. Seattle, USA. https://doi.org/10.1007/978-3-319-96644-1

Kuhn, Max, Kjell Johnson (2016) Applied Predictive Modeling. New York, USA. https://doi.org/10.1007/978-1-4614-6849-3

Lazcano-Ponce E, Fernández E, Salazar-Martínez E, et al (2000) Cohort studies. Methodology, biases and application. Salud Pública de México 42:230–241

Li S, Batterman S, Wasilevich E, et al (2011) Asthma exacerbation and proximity of residence to major roads: a population-based matched case-control study among the pediatric Medicaid population in Detroit, Michigan. Environ Health 10:34. https://doi.org/10.1186/1476-069X-10-34

Li Jin, Heap Andrew D (2014) Spatial interpolation methods applied in the environmental sciences: A review. Environ Modell Software 53:173-189. http://dx.doi.org/10.1016/j.envsoft.2013.12.008

Ly S, Charles C, Degr A (2011) Geostatistical interpolation of daily rainfall at catchment scale: the use of several variogram models in the Ourthe and Ambleve catchments, Belgium. Hydrol Earth Syst Sci 15:2259-2274. https://doi.org/10.5194/hess-15-2259-2011

MAVDT Ministry of Environment, Housing and Territorial Development (2010) Protocol for air quality monitoring. Bogota, Colombia.

Mazenq J, Dubus J-C, Gaudart J, et al (2017) City housing atmospheric pollutant impact on emergency visit for asthma: A classification and regression tree approach. Respir Med 132:1–8. https://doi.org/10.1016/j.rmed.2017.09.004

Pandey G, Zhang B, Jian L (2013) Predicting submicron air pollution indicators: a machine learning approach. Environ Sci Processes Impacts. 15:996–1005. https://doi.org/10.1039/c3em30890a

Polezer G, Tadano YS, Siqueira HV, et al (2018) Assessing the impact of $PM_{2.5}$ on respiratory disease using artificial neural networks. Environ Pollut 235:394-403. https://doi.org/10.1016/j.envpol.2017.12.111

Ramírez O, Sánchez de la Campa AM, Amato F, et al (2018) Chemical composition and source apportionment of PM_{10} at an urban background site in a high–altitude Latin American megacity (Bogota, Colombia). Environ Pollut 233:142–155. https://doi.org/10.1016/j.envpol.2017.10.045

Reid CE, Jerrett M, Tager IB, et al (2016) Differential respiratory health effects from the 2008 northern California wildfires: A spatiotemporal approach. Environ Res 150:227–235. https://doi.org/10.1016/J.ENVRES.2016.06.012

Rodríguez-Villamizar LA, Rojas-Roa NY, Blanco-Becerra LC, et al (2018) Short-Term effects of air pollution on respiratory and circulatory morbidity in Colombia 2011−2014: A multi-city, time-series analysis. Int J Environ Res Public Health 15:2-12. https://doi.org/10.3390/ijerph15081610

Rokach, Lior, and Oded Maimon (2015) Data Mining with Decision Trees: Theory and Applications. 2nd ed. Singapore: World Scientific Publishing Co. Pte. Ltd. 5.

Salam MT, Islam T, Gilliland FD (2008) Recent evidence for adverse effects of residential proximity to traffic sources on asthma. Curr Opin Pulm Med 14:3–8. https://doi.org/10.1097/MCP.0b013e3282f1987a

Sajjadia SA, Zolfagharib G, Adabc H, et al (2017) Measurement and modeling of particulate matter concentrations: Applying spatial analysis and regression techniques to assess air quality. MethodsX 4:372–390. https://doi.org/10.1016/j.mex.2017.09.006

Schapire RE, Freund Y (2012) Boosting: foundations and algorithms, Adaptive computation and machine learning. MIT Press, London.

SDA District Secretariat for the Environment (2017) Air quality annual report of Bogota, 2016. Bogotá, Colombia.

SDP District Planning Secretariat (2018) Monograph 2017 Assessment of the main territorial, infrastructure, demographic and socio-economic aspects of the locality of Kennedy 08. Bogotá, Colombia.

Weizhen H, Zhengqiang L, Yuhuan Z, et al (2014) Using support vector regression to predict PM_{10} and $PM_{2.5}$, in: IOP Conference Series: Earth and Environmental Science. IOP. https://doi.org/10.1088/1755-1315/17/1/012268

Westerlund J, Urbain JP, Bonilla J (2014) Application of air quality combination forecasting to Bogota. Atmos Environ 89:22-28. https://doi.org/10.1016/j.atmosenv.2014.02.015

WHO (2006) WHO Air quality guidelines for particulate matter, ozone, nitrogen dioxide and sulfur dioxide. Global update 2005. Geneva, Switzerland.

Yu Y, Yao S, Dong H, et al (2019) Association between short-term exposure to particulate matter air pollution and cause-specific mortality in Changzhou, China. Environ Res 170:7–15. https://doi.org/10.1016/j.envres.2018.11.041

Zhan Y, Luo Y, Deng X, et al (2017) Spatiotemporal prediction of continuous daily $PM_{2.5}$ concentrations across China using a spatially explicit machine learning algorithm. Environ Pollut 233:464-473. https://doi.org/10.1016/j.atmosenv.2017.02.023