Hindawi
Journal of Electrical and Computer Engineering
Volume 2017, Article ID 5106045, 14 pages
https://doi.org/10.1155/2017/5106045

Research Article

Modeling $PM_{2.5}$ Urban Pollution Using Machine Learning and Selected Meteorological Parameters

Jan Kleine Deters,[1] Rasa Zalakeviciute,[2] Mario Gonzalez,[2] and Yves Rybarczyk[2,3]

[1] *University of Twente, Enschede, Netherlands*
[2] *Intelligent & Interactive Systems Lab (SI[2] Lab), FICA, Universidad de Las Américas, Quito, Ecuador*
[3] *DEE, Nova University of Lisbon and CTS, UNINOVA, Monte de Caparica, Portugal*

Correspondence should be addressed to Yves Rybarczyk; y.rybarczyk@fct.unl.pt

Received 24 February 2017; Revised 23 April 2017; Accepted 11 May 2017; Published 18 June 2017

Academic Editor: Lei Zhang

Outdoor air pollution costs millions of premature deaths annually, mostly due to anthropogenic fine particulate matter (or $PM_{2.5}$). Quito, the capital city of Ecuador, is no exception in exceeding the healthy levels of pollution. In addition to the impact of urbanization, motorization, and rapid population growth, particulate pollution is modulated by meteorological factors and geophysical characteristics, which complicate the implementation of the most advanced models of weather forecast. Thus, this paper proposes a machine learning approach based on six years of meteorological and pollution data analyses to predict the concentrations of $PM_{2.5}$ from wind (speed and direction) and precipitation levels. The results of the classification model show a high reliability in the classification of low (<10 $\mu g/m^3$) versus high (>25 $\mu g/m^3$) and low (<10 $\mu g/m^3$) versus moderate (10–25 $\mu g/m^3$) concentrations of $PM_{2.5}$. A regression analysis suggests a better prediction of $PM_{2.5}$ when the climatic conditions are getting more extreme (strong winds or high levels of precipitation). The high correlation between estimated and real data for a time series analysis during the wet season confirms this finding. The study demonstrates that the use of statistical models based on machine learning is relevant to predict $PM_{2.5}$ concentrations from meteorological data.

1. Introduction

The effects of rapid growth of the world's population are reflected in the overuse and scarcity of natural resources, deforestation, climate change, and especially environmental pollution. Currently, more than half of the global population lives in urban areas, and this number is expected to grow to about 66% by 2050, mostly due to the urbanization trends in developing countries [1]. According to the latest urban air quality database, 98% of cities in low and middle income countries with more than 100,000 inhabitants do not meet the World Health Organization (WHO) air quality guidelines [2].

A recent study using a global atmospheric chemistry model estimated that 3.3 million annual premature deaths worldwide are linked to outdoor air pollution, which is expected to double by 2050, mostly due to anthropogenic fine particulate matter (aerodynamic diameter < 2.5 μm; $PM_{2.5}$) [3]. Over the last decade, evidence has been growing that exposure to fine particulate air pollution has adverse effects on cardiopulmonary health [4].

A recent air quality study in Quito, the capital of Ecuador, concurs that long-term levels of fine particulate pollution are not only exceeding the WHO's recommended levels of 10 $\mu g/m^3$ but also are higher than the national standards of 15 $\mu g/m^3$ [5]. And even though the overall levels of fine particulate pollution have been decreasing due to active efforts of the local and national governments in the last decade, in some locations of the city the air quality has continued to deteriorate. The latter reflects the global trends of urbanization and motorization.

In addition to the impact of urbanization and rapid population growth, the pollution levels in the cities are modulated by meteorological factors [6]. Most importantly, the depth of mixing layer (the lower layer of troposphere mixing surface emissions) often depends on solar radiation and thus temperature in the area. The shallower the mixing depth is,

the less diluted the daily emissions get. Therefore, temperature shows a reducing impact on fine particulate matter levels, through convection [7]. In addition, the formation and evolution of photochemical smog are dependent on solar radiation and temperature; meanwhile, wind speed tends to help ventilate air pollutants and/or transport them to other areas, even if the emission sources are not present in that region [8, 9]. This can result in increased levels of air pollution downwind from the original source, which directly depends on the wind direction [8]. Increased relative humidity has been shown to make even fine particles heavier, helping the dry deposition process of removal, while precipitation has a direct effect of scavenging by wet deposition [7, 8]. In addition, some studies differentiate between the seasons, as different parameters have different effects during the year, due to the combination of conditions [8, 9]. Thus, it is clearly impossible to rely on a single parameter to fully understand the urban pollution, especially if the study area is in a nonhomogeneous and complex terrain. This fact justifies the elaboration of models that take into account heterogeneous data to predict air quality.

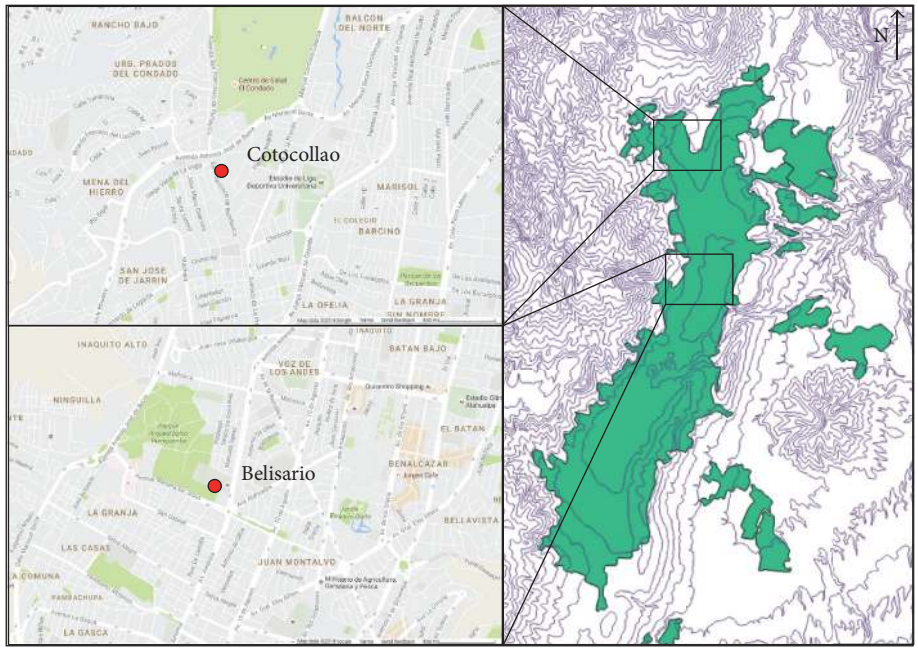
Currently, three major approaches are used to forecast $PM_{2.5}$ concentrations: statistical models, chemical transport, and machine learning. Statistical models, which are mainly based on single variable linear regression, have shown a negative correlation between different meteorological parameters (wind, precipitation, and temperature) and PM concentrations (PM_{10}, $PM_{2.5}$, and $PM_{1.0}$) [7]. Chemical transport and Atmospheric Dispersion Modeling are numerical methods, and the most advanced ones are WRF-Chem and CMAQ. These models can be used to predict atmospheric pollution, but their accuracy relies on an updated source list that is very difficult to produce [10]. In addition, complex geophysical characteristics of locations with complex terrain complicate the implementation of these models of weather and pollution forecast mostly due to the complexity of the air flows (wind speed and direction) around the topographic features [11, 12]. Unlike a pure statistical method, a machine learning approach can consider several parameters in a single model. The most popular classifiers to forecast pollution from meteorological data are artificial Neural Networks [13–15]. Other successful studies use hybrid or mixed models that combine several artificial intelligence algorithms, such as fuzzy logic and Neural Network [16], or Principal Component Analysis and Support Vector Machine [17], or numerical methods and machine learning [10].

Recent studies show that the machine learning approach seems to overcome the other two methods for forecasting pollution [9, 10]. This is the reason why it is increasingly used to predict air quality [13, 17–21]. However, the data mining does not only differ from one study to another, in terms of classification algorithms, but also regarding the used features. Some of them consider a quite exhaustive list of meteorological factors [15, 16], whereas others proceed with a careful selection [13, 14, 17, 22] or do not even use climatic parameters at all [18]. Since machine learning is a very promising method to forecast pollution, we propose applying this approach to predict $PM_{2.5}$ concentration in Quito. This prediction is based on a selection of meteorological features for two main reasons: first because a model using only meteorological data, which can be easily obtained in any urban area, is cheaper than an air quality monitoring system and second because a general model that may work for any city is not realistic [10], which implies that a selection of meteorological parameters must be performed in order to find the best model for the capital city of Ecuador. Quito is located in the Andes cordillera in the tropical climate zone, characterized by two seasons with different accumulation of precipitation. However, the temperature, the pressure, and even the amount of solar radiation do not vary much during the year. Moreover, the wind direction and speed highly depend on the topographic features of complex terrain in which a city is positioned and usually present one of the biggest challenges in forecasting weather and air quality. Therefore, this research aims to study the connectivity between three selected meteorological factors, wind speed, wind direction, and precipitation, and $PM_{2.5}$ pollution in two districts located in northwestern Quito.

In this work, we first present a spatial visualization of the distribution of fine particulate matter trends according to wind (speed and direction) and precipitation parameters in two locations in Quito. This part includes a description of the preparation of the data for classification. Then, various machine learning models are exploited to classify different levels of $PM_{2.5}$, namely, Boosted Trees and Linear Support Vector Machines. Finally, a Neural Network regression and a time series analysis are applied to provide insight about the parametric boundaries, in which the classification models perform adequately. In the final section, we draw up the main conclusions and suggestions for future work.

2. Data Collection

2.1. Site Description. Unlike most of South America, the most urbanized continent on the planet (81%), Ecuador, is one of the few countries in the region with only 64% of total population living in urban areas [23]. However, the rate of urbanization has increased over the past decade. Quito sprawls north to south on a long plateau lying on the east side of the Pichincha volcano (alt. 4,784 m.a.s.l., meters above sea level) in the Andes cordillera at an altitude of 2,850 m.a.s.l. (see Figure 1). According to the 2010 census, Quito's metro area is currently 4,217.95 km^2 with a population over 2,239,191 and is expected to increase to almost 2.8 million by 2020, making the city the most populous city in the country, overgrowing Guayaquil [24]. The city is contained within a number of valleys at 2,300–2,450 m.a.s.l. and terraces varying from 2,700 to 3,000 m.a.s.l. altitude. Due to Quito's location on the Equator, the city receives direct sunlight almost all year round, and, due to its altitude, Quito's climate is mild, spring-like all year round. The region has two seasons, dry (June–August, average precipitation 14 mm/month) and wet (September–May, average precipitation 59 mm/month), with most of the rainfall in the afternoons. Quito's temperature is almost constant, around 14.5°C, with the prevailing winds from the east. However, due to a complex terrain, the winds in the city are highly variable most of the year (dry season is windier), challenging weather prediction in the region.



(a) (b)

Figure 1: Topographic map (b) of Quito's urban area (green areas) and Google maps images (a) of the air quality measurement sites (red dots) Cotocollao and Belisario.

For the purpose of this study, the two northwestern air quality monitoring points are presented: Cotocollao and Belisario (see red dots in Figure 1). These districts were chosen to show the variation and complexity of the prediction of fine particulate matter trends even within a relatively small area of Quito with similar topographical characteristics (approximately the same altitude and directly east of the Pichincha volcano).

2.2. Air Quality Measurements Monitoring Network and Instrumentation. The municipal office of environmental quality, *Secretaria de Ambiente*, has been collecting air quality and meteorological data since May 1, 2007, in several sites around the city. The measurement sites run by the *Secretaria de Ambiente* are located in representative areas throughout the city, varying by altitudes depending on municipal districts. We used the real meteorological and $PM_{2.5}$ concentration data from the two most northwestern automatic data collection stations: Belisario (alt. 2,835 m.a.s.l., coord. 78°29′24″W, 0°10′48″S) and Cotocollao (alt. 2,739 m.a.s.l., coord. 78°29′50″W, 0°6′28″S) (see Figure 1). These two sites are approximately 9 km apart from each other. The Belisario measurement site is less than 100 m west of a busy road (Avenida America), 200 m northwest of a busy roundabout, and less than 1,000 m to the east of a major outer highway (Ave. Antonio Jose de Sucre), which runs along the west side of the city, intended to reduce the traffic inside the city (Figure 1). The Cotocollao monitoring site is located in a residential area, with only a few busier streets, and the same outer highway (Ave. Antonio Jose de Sucre) 250 m to the north. Both monitoring sites are inside of the *"Pico y Placa"* zone, implemented in 2010, which, based on the last number of car license plates, limits rush hour traffic reducing the number of personal vehicles by approximately 20% during the weekdays.

The monitoring stations are positioned on the roofs of relatively tall buildings. Fine particulate matter ($PM_{2.5}$) measurements are conducted using instrumentation validated by the Environmental Protection Agency (EPA) of the United States. For $PM_{2.5}$ Thermo Scientific FH62C14-DHS Continuous, 5014i (EPA Number EQPM-0609-183), was used. The detection limit for this instrument is 5 $\mu g/m^3$ for one-hour averaging. The aerosol data is collected at 10 s intervals, and from this then 10 min, 1-hour, and 24-hour averages are calculated. The latter averaging data is presented in this work. Wind velocity is measured using MetOne/010C and wind direction using MetOne/020C instrumentation. The wind speed sensor and wind direction starting threshold is 0.22 m/s, and the accuracies are 0.07 m/s and 3°, respectively. The precipitation is measured using MetOne/382 and Thies Clima/5.4032.007 equipment. All meteorological parameters have been validated using Vaisala/MAWS100 weather station.

3. Data Preparation

In this section the method for the preparation of the data is presented, in order to proceed with the classification. It includes refining steps to discard useless data, transformations to visually examine and understand the data, and creation of an averaged intensity map of the $PM_{2.5}$ concentrations with respect to the selected meteorological parameters (wind and precipitation).

3.1. Data Refinement. For this study we analyzed the data of six years, starting June 2007 and ending July 2013. The two

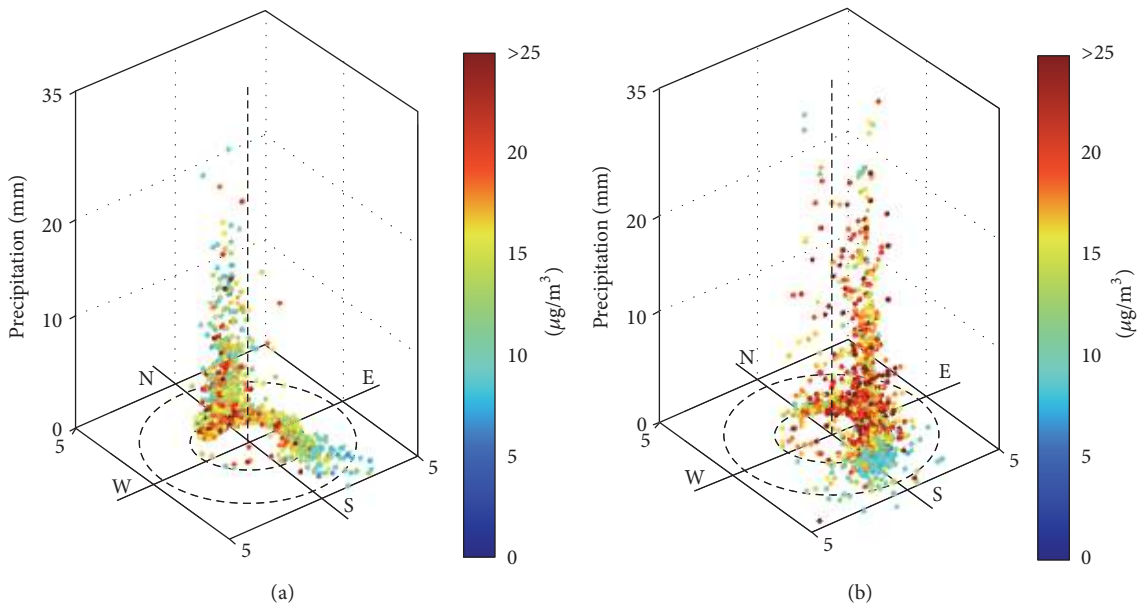


Figure 2: Data distribution for (a) Cotocollao and (b) Belisario, in terms of wind direction, wind speed, precipitation, and $PM_{2.5}$ concentrations (color scale). The inner circle represents wind speeds up to 2 m/s and the outer circle represents wind speeds up to 4 m/s.

datasets (one for each monitoring point) are composed out of 2,223 instances. Each data point consists of 4 parameters indicating daily values of precipitation accumulation (mm), wind direction (0–360°), wind speed (m/s), and observed fine particle concentrations (μg/m^3).

The datasets are cleaned by discarding data points that include any missing values. These data points represent 2.8% and 2.4% of the total data for Belisario and Cotocollao, respectively. It has been demonstrated that missing data of these magnitudes do not influence the classification performance [25]. In addition, considering the very low number of missing values, it is preferable to remove them instead of performing an interpolation, taking into account the following: (i) we proceed with an analysis on discrete variables (day-by-day) and not a time series forecasting and (ii) the $PM_{2.5}$ concentrations are very inconstant from one day to another. Weekend days are also removed from the dataset because the distribution of $PM_{2.5}$ concentrations during the weekdays and weekends is very different for Quito. This could introduce an additional level of complexity in data classification as during the weekdays there are clear rush hour peaks (morning and evening), while on Saturdays $PM_{2.5}$ levels increase between late morning and late afternoon hours. In addition, Sundays can be identified by a drop of $PM_{2.5}$ concentration. These patterns are dictated by human activity changes during the week, therefore, clearly showing $PM_{2.5}$ dependability on traffic. After cleaning, the final datasets are composed of 1,527 instances for Belisario and 1,536 instances for Cotocollao.

3.2. Data Transformation. To represent the data according to a wind rose plot, the linear scale of wind direction (0–360°) is transformed from polar to Cartesian coordinates where angles increase clockwise and both 0° and 360° are north (N) (see Figure 2). This mathematical transformation (see (1)) permits a more accurate feature representation of the data for wind direction around the north axis. Otherwise, wind direction angles slightly higher than 0° and slightly lower than 360° would be considered as two opposing directions. This is useful for classification models that are implemented in the next stage. This relates to machine learning models that improve performance if there are continuous relationships between parameters (optimization: smoother clustering task) [26]. This transformation ensures both valid and more informative representation of the original data. In addition, this representation can be completed by the precipitation levels, which are plotted on the z-axis (Figure 2). The color range is mapped from concentrations 0 μg/m^3 to >25 μg/m^3. The threshold of 25 μg/m^3 indicates the values from which the 24-hour concentrations of $PM_{2.5}$ are harmful according to international health standards.

$$x = \sin\left(\frac{\text{Wind Direction}}{360^\circ} \cdot 2\pi\right) \cdot \text{Wind Speed},$$
$$y = \cos\left(\frac{\text{Wind Direction}}{360^\circ} \cdot 2\pi\right) \cdot \text{Wind Speed}. \quad (1)$$

A visual inspection of the transformed data shows that the wind directions corresponding to precipitation are north (N) for Cotocollao (Figure 2(a)) and east (E) for Belisario (Figure 2(b)). The stronger winds tend to take place between south (S) and southeast (SE) for Cotocollao and between southwest (SW) and SE in Belisario. As expected, in both cases these stronger winds seem to account for relatively low levels of $PM_{2.5}$.

3.3. Trend Analyses. In order to obtain general trends in the distribution of the $PM_{2.5}$ concentrations as a function of

wind speed and wind direction, the data are used to generate convolutional based spatial representations. Convolution-based models for spatial data have increased in popularity as a result of their flexibility in modeling spatial dependence and their ability to accommodate large datasets [27]. This generated Convolutional Generalization Model (CGM) [28] is an averaged value of the $PM_{2.5}$ pollution level (PL), in which the regional quantity of influence per data point is modeled as a 2D Gaussian matrix (see (2)). A Gaussian convolution is applied (i) to spatially interpolate data, in order to get a 2D representation from the points' coordinates calculated in (1) and (ii) to smooth the PL concentration values of this representation. A Gaussian kernel is used because it inhibits the quality of monotonic smoothing, and as there is no prior knowledge about the distribution, a kernel density function with high entropy minimizes the information transfer of the convolution step to the processed data [29]. This 2D Gaussian matrix is multiplied by the PL of the given data point and added to the CGM at the coordinates corresponding to the wind speed and direction of this point. Then, the quantity of influence is added to the point. The final step is to divide the total amount of each cell by the quantity of influence, which results in a generalized average value.

$$\text{CGM}(\text{rows}, \text{colums}) = \text{PL}\frac{1}{36}\begin{bmatrix}1\\4\\6\\4\\1\end{bmatrix}\begin{bmatrix}1 & 4 & 6 & 4 & 1\end{bmatrix}. \quad (2)$$

The general tendencies are as follows: (i) strong winds result in low $PM_{2.5}$ concentrations and (ii) the strongest winds generally come from the similar direction (SE for Cotocollao and S for Belisario). The results of CGMs for both sites are shown in Figure 3 as an overlay on top of the geographic location of their respective monitoring stations. Main highways are indicated in green. The highest concentrations of $PM_{2.5}$ (from yellow to red) tend to be brought by the winds coming from these main highways. It is to note that higher wind speeds for Cotocollao tend to be on the axis of Quito's former airport (grey-green area, center of the map, see Figure 3), currently transformed into a city park. This traffic and structure free corridor seems to accelerate wind speeds, which may explain the reduction of $PM_{2.5}$ concentrations due to better ventilation of this part of the city.

During the study, average $PM_{2.5}$ concentrations in Cotocollao and Belisario are 15.6 $\mu g/m^3$ and 17.9 $\mu g/m^3$, respectively, both exceeding the national standards. During the studied six years, the area of Belisario was more polluted with more variation in $PM_{2.5}$ concentrations (higher deviation, see Figure 4) and more turbulent (Figure 3) than Cotocollao. These factors could be the result of Belisario being more urbanized.

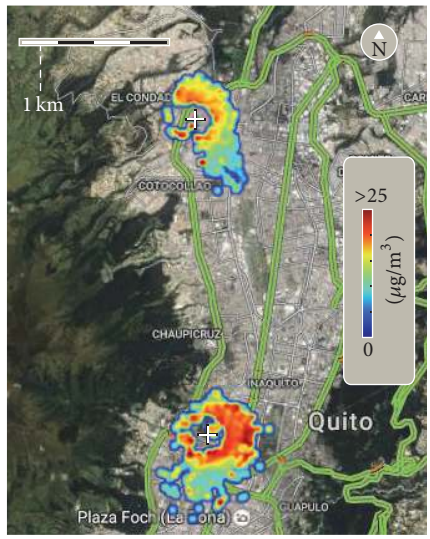


Figure 3: CGM visualization, positioned on top of the geographic location of the respective monitoring stations (northwestern part of Quito). The northern CGM visualization is Cotocollao and the southern one is Belisario. Main highways are represented in green.

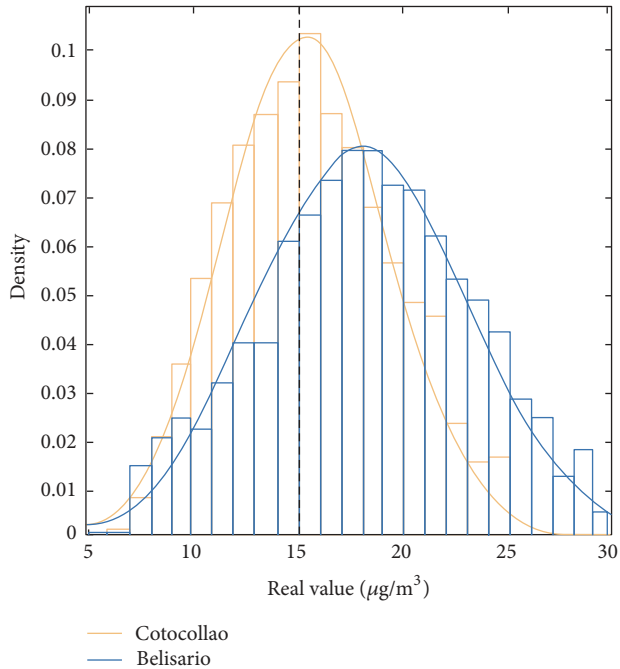


Figure 4: Distribution of $PM_{2.5}$ concentrations (June 2007 to July 2013) for Cotocollao and Belisario. Dashed black line represents the national standards and the class seperation boundary (15 $\mu g/m^3$).

4. Classification Models

Machine learning models are used to separate the data in different classes of $PM_{2.5}$ concentrations. Supervised learning techniques are applied to create models on this classification task. Here we introduce Boosted Trees (BTs) and Linear Support Vector Machines (L-SVM). A BT combines weak learners (simple rules) to create a classification algorithm, where each misclassified data point per learner gains weight. A following learner optimizes the classification of the highest weighted region. Boosted Trees are known for their

Table 1: Binary classification with class separation at 15 $\mu g/m^3$.

Model	Location	
	Belisario	Cotocollao
BT	83.2%	67.6%
L-SVM	79.8%	66.3%

insensibility to overfitting and for the fact that nonlinear relationships between the parameters do not influence the performance. A L-SVM separates classes with optimal distance. Convex optimization leads the algorithm to not focus on local minima. As these two models are well established and inhibit different qualities, they are used in this section. All computations and visualizations are executed in MathWorks Matlab 2015. Toolboxes for the classifications, the statistics, and machine learning processes are used in all the stages. Furthermore, Matlab's integrated tools for distribution fitting and curve fitting are applied for the different analyses. The initial parameters provided by the Matlab toolbox software are used in this work. ADA boost learning method with a total amount of 30 learners and a maximum number of splits being 20 at a learning rate of 0.1 are the default parameters for the BT. The SVM is initialized with a linear kernel of scale 1.0, a box constrained level of 1.0, and an equal learning rate of 0.1.

Fluctuations in yearly $PM_{2.5}$ concentrations are not taken into account in this classification process as a previous analysis showed a small variation in fine particulate matter pollution levels during the studied period [5]. A binary classification is performed to set a baseline comparison between the different sites. Then, a three-class classification is carried out to assess the separability between three ranges of concentrations of $PM_{2.5}$ (based on WHO guidelines) and provide insight into general classification rules.

4.1. Binary Classification. In this first classification two classes are used, which represent values above and below 15 $\mu g/m^3$. The latter value is selected as it is the National Air Quality Standard of Ecuador for annual $PM_{2.5}$ concentrations (equivalent to WHO's Interim Target-3) [30]. Due to the normal distribution of the datasets, as shown in Figure 4, a higher accuracy for Belisario than Cotocollao is expected, partially because of a priori imbalanced class distribution. A previous study using the same classification shows an accuracy of only 65% for Cotocollao by applying the trees.J48 algorithm, which is a decision tree implementation integrated in the WEKA machine learning workbench [5].

Classification with both BT and L-SVM shows similar results. Table 1 presents the results of this first classification. The implementation of the classification for Belisario outperforms that of Cotocollao. It also suggests that the extreme levels (low and high) of $PM_{2.5}$ could be more straightforward to classify with the current parameters, implying a higher class separability for the Belisario dataset (wider distribution). Tables 2 and 3 show that the concentrations above 15 $\mu g/m^3$ for both sites are better classified than those below the 15 $\mu g/m^3$ boundary. This is less surprising for Belisario due to the earlier mentioned class imbalance. For Cotocollao, however, the poor performance for this class can indicate that this class is less distinctive; thus the model optimizes the class above 15 $\mu g/m^3$. Note that it is crucial to be able to classify nonattainment ($PM_{2.5} > 15\,\mu g/m^3$) instances, as wrongly identified nonviolating national standards ($PM_{2.5} < 15\,\mu g/m^3$) levels would be a less costly error.

Table 2: Confusion matrix of binary classification for Cotocollao using a BT. Rows represent the true class and columns represent the predicted class.

Class	<15	>15	TPR/FNR
<15	51.1%	48.9%	**51.1**% **48.9**%
>15	20.3%	79.7%	**79.7**% **20.3**%

Table 3: Confusion matrix of Binary classification for Belisario using a BT. Rows represent the true class and columns represent the predicted class.

Class	<15	>15	TPR/FNR
<15	49.0%	51.0%	**49.0**% **51.0**%
>15	5.1%	94.9%	**94.9**% **5.1**%

In Figure 5(a) Receiver Operating Characteristic (ROC) curves comparison is shown for the binary classifiers presented in Table 1, namely, the BT and L-SVM classifiers. Figure 5(a) depicts the ROC curves for Cotocollao dataset and Figure 5(b) the ROC curves for Belisario dataset. Once the classifiers models are built for every dataset, a validation set is presented to the model, in order to predict the class label. It is also of interest to have the classification scores of the model which indicate the likelihood that the predicted label comes from a particular class. The ROC curves are constructed with this scored classification and the true labels in the validation dataset (Figure 5).

ROC curves are useful to evaluate binary classifiers and to compare their performances in a two-dimensional graph that plots the specificity versus sensitivity. The specificity measures the true negative rate, that is, the proportion of negatives that have been correctly classified: true negatives/negatives = true negatives/(true negatives + false positives). Likewise, the sensitivity measures the true positive rate, that is, the proportion of positives correctly identified: true positives/positives = true positives/(true positives + false negatives). The area under the ROC curve (AUC) can be used as a measure of the expected performance of the classifier, and the AUC of a classifier is equal to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance [31]. Figure 5(b) shows the performance of the BT and L-SVM classifiers for the Belisario dataset. The BT outperforms the L-SVM classifier in all regions of the ROC space, with [AUC(BT) = 0.72] > [AUC(L-SVM) = 0.66], which means a better performance

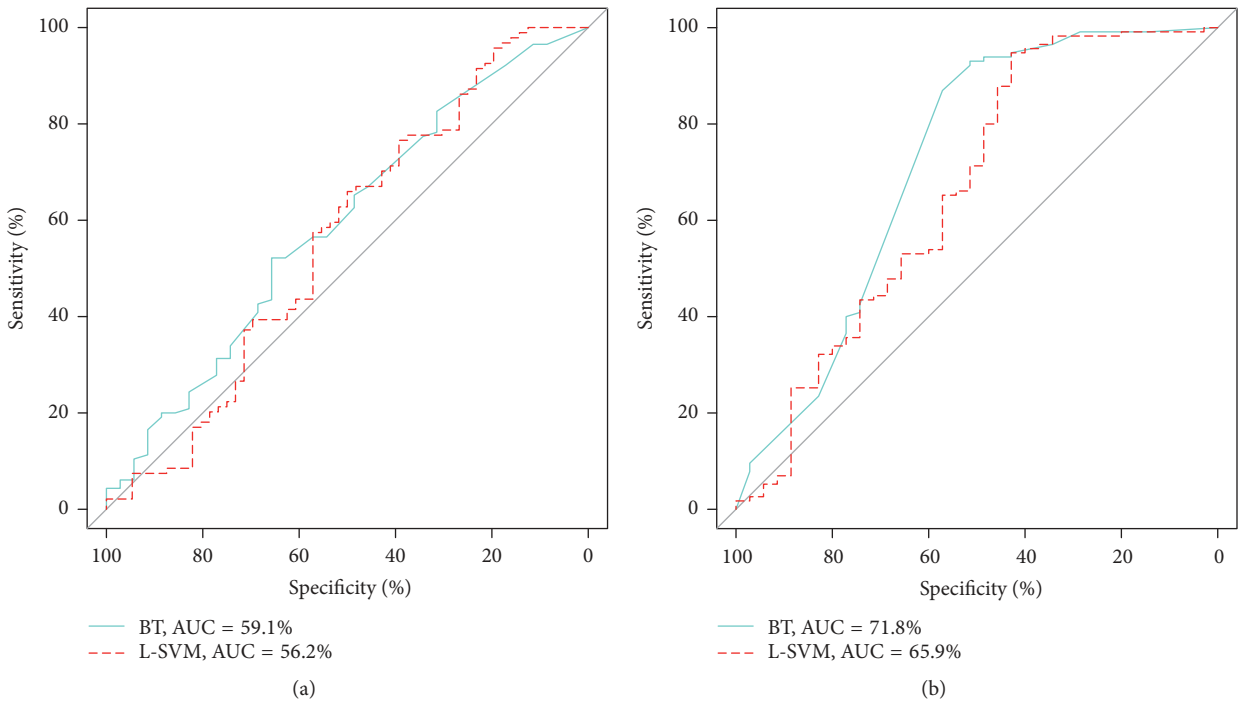


Figure 5: ROC curves for Cotocollao (a) and Belisario (b).

for the BT classifier. The BT classifier has a fair performance separating the two classes in the Belisario dataset.

In Figure 5(a) the ROC curves and AUC are presented for the Cotocollao dataset. Again, BT performs better than the L-SVM classifier with [AUC(BT) = 0.59] > [AUC(L-SVM) = 0.56]. This time the classifiers for the Cotocollao dataset have a poor performance separating the two classes, with a performance just slightly better when compared to a random classifier with AUC = 0.5. The classification result is clearly better for Belisario than for Cotocollao. Thus, a three-class classification should identify if for both sites; the extreme concentrations could be better classified than the moderate ones and clarify the low performance for Cotocollao.

4.2. Three-Class Classification. To further analyze the differences of multiple categories of concentration levels, a three-class classification is performed using WHO's guidelines for pollution concentrations as class boundaries. According to these guidelines, health risks are considered low if $PM_{2.5}$ < 10 $\mu g/m^3$ (long term, annual WHO's recommended level), moderate if 10 $\mu g/m^3$ > $PM_{2.5}$ < 25 $\mu g/m^3$, and high if $PM_{2.5}$ > 25 $\mu g/m^3$ (short term, 24-hour WHO's recommended level). The objective is to identify if these main pollution thresholds are indeed well separable and thus the weather parameters can account for $PM_{2.5}$ pollution in these three ranges of air quality.

In both studied districts the classes < 10 $\mu g/m^3$ and >25 $\mu g/m^3$ are relatively small with approximately 10% of the data compared to the class 10–25 $\mu g/m^3$. Due to this fact, an alternative BT algorithm is used to take into account these imbalanced classes. This RusBoosted Tree (RBT) approach endeavors to find an even distribution of performance for all classes instead of finding a global optimum [32]. This leads to a better representation of the separability. The true positive versus false negative rate (TPR/FNR) is shown for each class in the confusion matrices of Cotocollao (Table 4) and Belisario (Table 5).

Table 4: Confusion matrix of three-class classification for Cotocollao using a RBT. Rows represent the true class and columns represent the predicted class.

Class	<10	10–25	>25	TPR/FNR
<10	76.3%	16.3%	7.4%	**76.3%** **23.7%**
10–25	28.3%	28.8%	42.9%	**28.8%** **71.2%**
>25	6.3%	20.3%	73.4%	**73.4%** **26.6%**

Tables 4 and 5 show that the correctness in classifying concentrations < 10 $\mu g/m^3$ seems to perform adequately. Also, the correct classification for concentrations > 25 $\mu g/m^3$ in Cotocollao is fair. However, the false positive rate of this classification is extremely high, because 42.9% of the 10–25 $\mu g/m^3$ class gets classified as class > 25 $\mu g/m^3$. For Belisario, the separation of classes 10–25 $\mu g/m^3$ and >25 $\mu g/m^3$ is deficient. In both cases, only the extreme low values can be classified well. Thus, the hypothesis of the extreme concentrations in $PM_{2.5}$ being more straightforward to classify (see Section 4.1) is only partially verified.

Analyzing the wrongly classified samples of class 10–25 $\mu g/m^3$ shows that, for samples classified as <10 $\mu g/m^3$, the

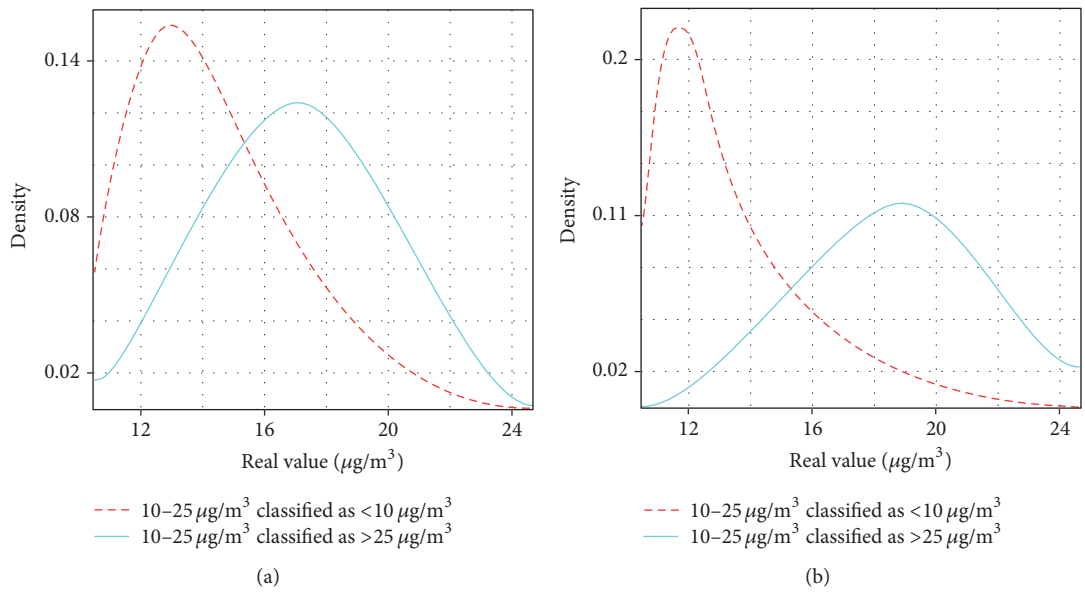


Figure 6: Wrongly classified samples of class 10–25 $\mu g/m^3$ with their real value distributions for Cotocollao (a) and Belisario (b).

Table 5: Confusion matrix of three-class classification for Belisario using a RBT. Rows represent the true class and columns represent the predicted class.

Class	<10	10–25	25	TPR/FNR
<10	84.8%	9.5%	5.7%	**84.8%** **15.2%**
10–25	12.3%	53.5%	34.2%	**53.5%** **46.5%**
>25	6.5%	45.1%	48.4%	**48.4%** **51.6%**

real values tend to be relatively close to 10 $\mu g/m^3$. This evidence is even stronger for Belisario (Figure 6(b)), than for Cotocollao (Figure 6(a)). This indicates a changeover in values around the decision boundary. The same does not apply to the wrongly classified samples that are grouped as >25 $\mu g/m^3$. As shown in Figure 6 these values are mostly normally distributed around the mean of class 10–25 $\mu g/m^3$. Even though for Belisario the mean is shifted, it is not evident that wrongly classified samples of class 10–25 $\mu g/m^3$ into class 25 $\mu g/m^3$ tend to be closer to values of 25 $\mu g/m^3$, as this shift is mainly caused by the fact that the mean value of the Belisario initial data is higher (see Figure 4). We can conclude that the low performance for Cotocollao in the previous section (Section 4.1) is mainly caused by the fact that the classifier tries to separate values in the range of 10–25 $\mu g/m^3$ and >25 $\mu g/m^3$, which are poorly separable according to the three-class classification.

These results show that values of 10–25 $\mu g/m^3$ and >25 $\mu g/m^3$ are not well separable and thus not largely influenced by the used meteorological parameters. On the contrary, lower values seem to be largely predictable by wind and precipitation conditions. This statement gains confidence by looking at the wrongly classified data points discussed previously (see Figure 6).

4.3. Classification Rules. Binary classification between all different classes with the use of RBTs provides general rules for classifying the different levels of $PM_{2.5}$ in terms of the parameter space. Here, the well performing rules in classifying $PM_{2.5}$ concentrations < 10 $\mu g/m^3$ are discussed. The rules and their performance can be seen in Table 6. This table shows that rules separating classes < 10 $\mu g/m^3$ versus 10–25 $\mu g/m^3$ and <10 $\mu g/m^3$ versus >25 $\mu g/m^3$ have a high percentage of accuracy. On the contrary, the separation between 10–25 $\mu g/m^3$ and >25 $\mu g/m^3$ is less accurate.

Figure 7 provides a visualization of the data according to the class separation in Table 6 for the example of Cotocollao. The RBT classification of the data as seen in Figures 7(a) and 7(b) creates two clusters for class < 10 $\mu g/m^3$. In the case of Belisario, the RBT classifications result in identifying only one cluster for class < 10 $\mu g/m^3$.

It is to note that, for Cotocollao, the performance increases drastically comparing the binary classifications of <10 $\mu g/m^3$ versus 10–25 $\mu g/m^3$ and <10 $\mu g/m^3$ versus >25 $\mu g/m^3$ (from 73.2% up to 88.9%, see Table 6). In contrast, the performance for Belisario for these two classifications does not differ (from 86.7% to 88.8%). This indicates that the data for Cotocollao are less separable at the 10–25 $\mu g/m^3$ class than for Belisario.

To sum up the outcomes of the classification models, the binary classification utilizing the National and International Air Quality Standards as class labels ($PM_{2.5}$ < 15 $\mu g/m^3$, $PM_{2.5}$ > 15 $\mu g/m^3$) showed a high difference in performance

Table 6: Classification rules and pairwise comparisons between the different classes and their respective performance.

Classification	Location	
	Cotocollao	Belisario
	Classification rules	
$<10\ \mu g/m^3$ versus $10–25\ \mu g/m^3$	Wind speed > 2.5 m/s Wind direction = S-SE Wind direction = NW-NE Precipitation > 15 mm	Wind speed > 2.2 m/s Wind direction = SE-SW
	Classification performance	
	73.2% (Figure 7(a))	86.7%
	Classification rules	
$<10\ \mu g/m^3$ versus $>25\ \mu g/m^3$	Wind speed > 2 m/s Wind direction = S-SE Wind direction = NW-NE Precipitation > 1 mm	Wind speed > 2 m/s Wind direction = SE-SW
	Classification performance	
	88.9% (Figure 7(b))	88.8%
$10–25\ \mu g/m^3$ versus $>25\ \mu g/m^3$	60.0%	64.1%

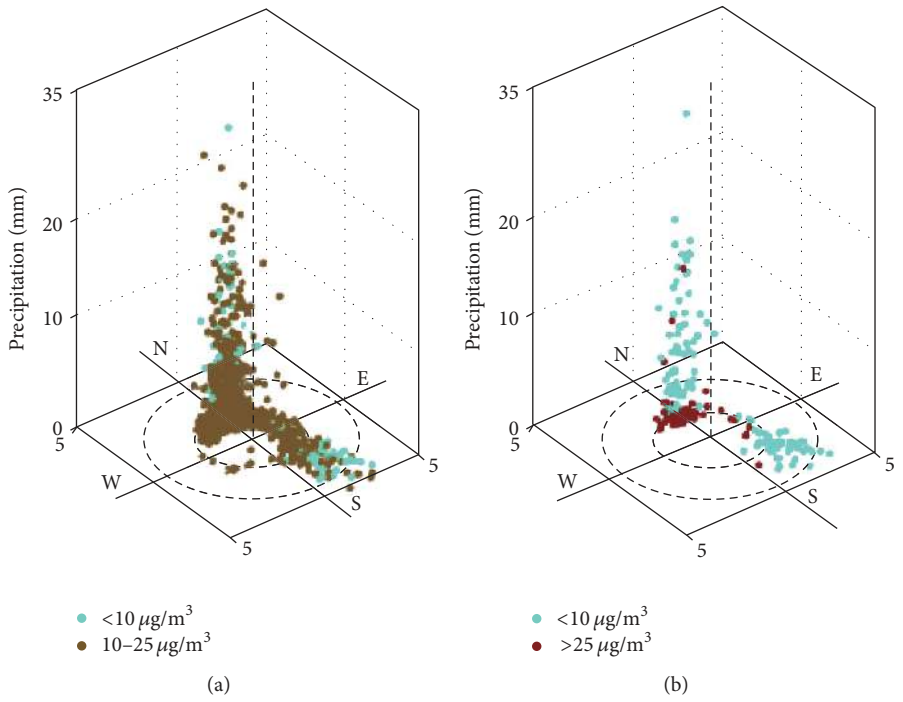


Figure 7: Data split for three different classes (see Table 6): (a) $<10\ \mu g/m^3$ versus $10–25\ \mu g/m^3$ and (b) $<10\ \mu g/m^3$ versus $>25\ \mu g/m^3$. Both (a) and (b) are results for Cotocollao mapped in terms of wind direction, wind speed, and precipitation. The inner circle represents wind speeds up to 2 m/s and the outer circle represents wind speeds up to 4 m/s.

between the two sites. In order to explain this difference and the misclassifications, the analysis was refined to a three-class classification based on WHO's guidelines regarding the consequences of $PM_{2.5}$ concentrations on health risks as low ($PM_{2.5}$ $< 10\ \mu g/m^3$), moderate ($PM_{2.5}$ = $10–25\ \mu g/m^3$), and high ($PM_{2.5}$ $> 25\ \mu g/m^3$). This classification showed high performance in categorizing low concentrations in contrast to high concentrations. Next, we propose a regression analysis to pinpoint the upper boundary of $PM_{2.5}$ values, for which the weather parameters are still able to explain variation in pollution levels that are not described by the classification analysis.

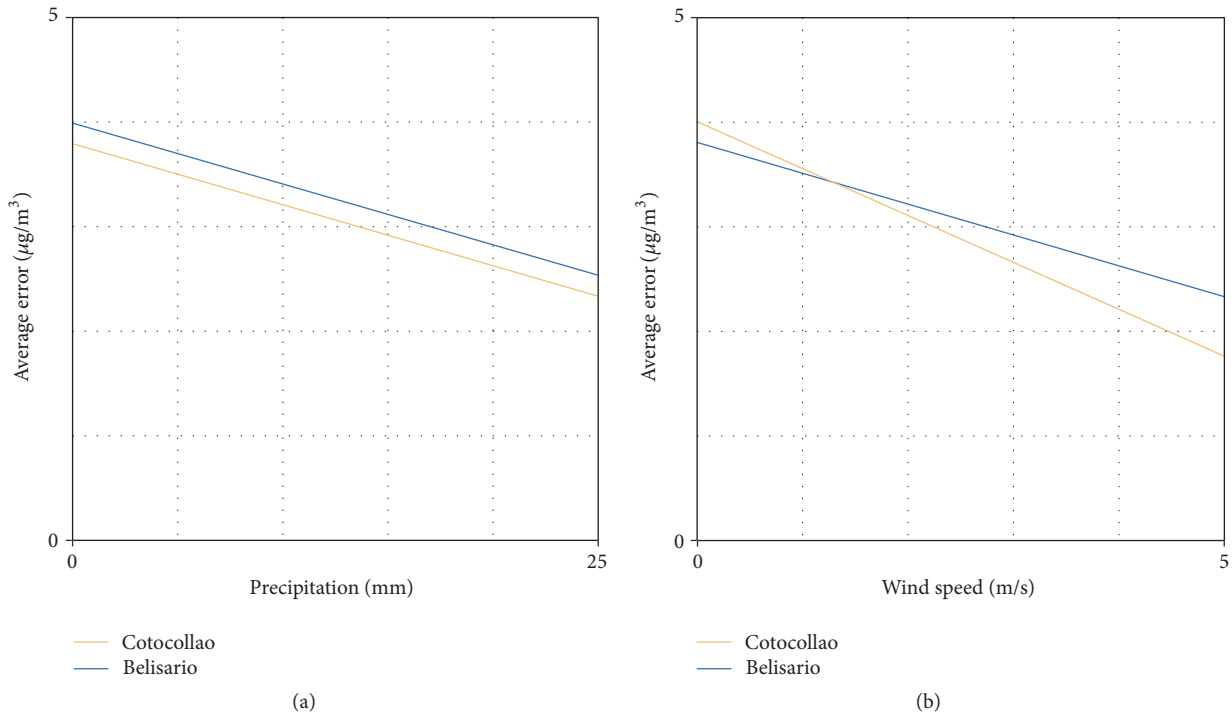


Figure 8: Decrease in average prediction error with increasing parameter values (precipitation and wind speed) for Cotocollao (orange) and Belisario (blue).

5. Regression Analyses

In this section an additional machine learning analysis, based on BT, L-SVM, and Neural Networks (NN), is used to perform a regression for both sites. Default parameters provided by the Matlab toolbox software are used to set up the models. NN are appropriate models for highly nonlinear modeling and when no prior knowledge about the relationship between the parameters is assumed. The NN consist of 10 nodes in 1 hidden layer, trained with a Levenberg-Marquardt procedure, in combination with a random data division. Identifying the correlation between the real and predicted values gives us the topological coherence between the input and output parameter values. In addition, the error related to the parameter values provides insight regarding the prediction confidence for determined weather conditions. Also, the analysis of the data trend over time will inform on the applicability of a time series forecasting. Finally, the CGM is used to remark on the possibility of optimizing the regression.

5.1. Regression Models. A regression is performed with three different classifiers. Bin sizes of 0.5 μg/m^3 (0–35 μg/m^3 range) are used for the models that output discrete class values (BT and SVM). This relatively small bin size permits these models to perform regression as their output values closely approach continuous values. The additional parameters of the models are set up as explained in the binary and three-class classification (Sections 4.1 and 4.2). The models are trained with 10-fold cross-validation. The test set is 20% of the original data. Unlike the NN continuous output values, the discrete output values of the other models can have an effect on the classification error. However, as the bin size is relatively small, we expect the errors related to these types of output to be marginal.

$$\mathrm{MSE} = \frac{1}{n} \cdot \sum_{i=1}^{n} \left(y_i - \widehat{y}_i\right)^2. \tag{3}$$

The mean squared error (MSE) is used to measure the classification performance (see (3)). The MSE is the averaged squared error per prediction. The mean absolute percentage error (MAPE) is used to express the average prediction error in terms of percentage of a data point's real value (see (4)). The MAPE function provides a more intuitive understanding of the performance.

$$\mathrm{MAPE} = \frac{\sum_{i=1}^{n} \left|\left(y_i - \widehat{y}_i\right)/y_i\right|}{n}. \tag{4}$$

An analysis of the confidence levels in relation to the precipitation and wind speed parameters is shown in Figure 8. The prediction confidence rises when the parameter values increase. A level of confidence is explained as the average prediction error (absolute difference between the real and the predicted values, root of MSE) at a certain interval with respect to an input parameter. In Figure 8, fitted lines represent the predicted data in terms of their absolute error with respect to precipitation and wind speed for both sites. The decrease in errors can be seen with respect to increasing

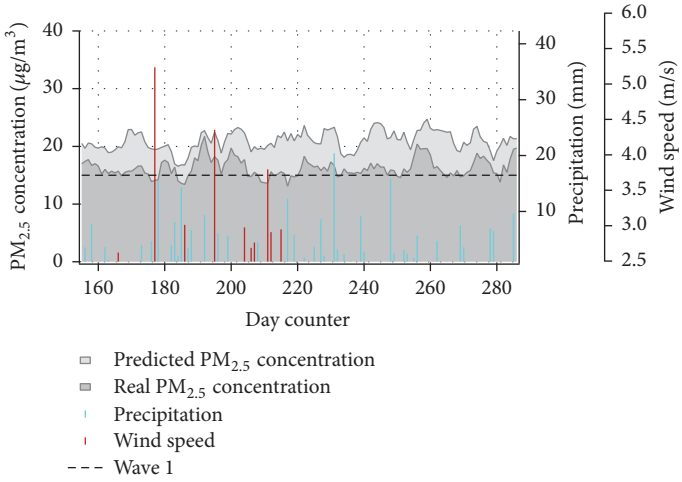


Figure 9: Neural Network's regressive prediction of Cotocollao $PM_{2.5}$ concentration (light grey) compared to the real data (dark grey) during the wet season plotted against daily rain accumulation and wind speed thresholds, >1 mm and >2.5 m/s, respectively (see Table 6, thresholds obtained from 3-class classification). The dashed black line represents the national standards for $PM_{2.5}$ annual concentrations.

values of these specified input parameters. It suggests that the prediction of $PM_{2.5}$ concentration is more reliable for extreme than moderate climatic conditions.

Figure 9 shows an example of the comparison of the predictive models of $PM_{2.5}$ concentration and the real $PM_{2.5}$ concentration for Cotocollao during six months of a wet season (first half of 2008). The graph shows the 5-point box-smoothed data to demonstrate the good prediction of the tendency of the $PM_{2.5}$ concentrations. Besides a certain gap, the estimated values seem to fairly correlate with the real data. The correlation analysis shows a significant positive correlation between the real concentrations and the predicted concentrations, $r(130) = 0.5$, $p < 0.000$. Also, the model performance is relatively good throughout the study period. The correlation analysis for all of the data shows a significant positive correlation between the real and predicted $PM_{2.5}$ concentrations, $r(1534) = 0.34$, $p < 0.000$.

This visualization shows that the error of predicted concentration seems to increase when $PM_{2.5}$ concentration increases. The reduction in both real and estimated $PM_{2.5}$ concentrations coincides with rain events and wind speeds above the thresholds defined in Table 6 (>1 mm and >2.5 m/s, resp.).

The results of the MSE for the regression show that in both city sites a NN performs the best (see Table 7). The correlation analysis shows that there is a logarithmic relationship between the real particle concentration values and the prediction (Figure 10). It means that there is an overprediction for low values and an underprediction for high values and an overall decrease in correlation as values get higher. The correlation seems the best for values around 17 $\mu g/m^3$ for Cotocollao and 19 $\mu g/m^3$ for Belisario.

To sum up, the present input parameters do not well describe an increase in $PM_{2.5}$ concentrations if these levels are transcending values over 20 $\mu g/m^3$, as errors increase at this point and prediction values stagnate. Thus, additional parameters must be considered for the prediction of $PM_{2.5}$ levels beyond this concentration threshold, since meteorological factors alone are not able to account for the whole particulate matter concentrations. For instance, considering human activity (e.g., car traffic), which is the main source of pollution, should contribute to the reduction of the overprediction and underprediction observed in our model.

Table 7: MSE and MAPE of the NN, L-SVM, and BT on regression.

Model	Location	
	Belisario	Cotocollao
NN	22.1 (26%)	40.7 (40%)
L-SVM	26.8 (28%)	41.8 (41%)
BT	28.5 (30%)	44.4 (42%)

Table 8: MSE and MAPE of CGM and NN regression.

Model	Location	
	Belisario	Cotocollao
CGM	15.6 (22%)	15.0 (25%)
NN	22.1 (26%)	40.7 (40%)

5.2. Optimization. The CGM, as applied in Section 3.3, could be used in classification tasks. In this section a 10-fold cross-validation on regression with this model is applied to compare it with the best performing model (NN).

The results show a substantial reduction in MSE with the CGM regression compared to the NN regression for the two city sites (see Table 8). It is to note that this diminution is particularly high in the case of Cotocollao. It seems that the model is able to better handle the dense (see Figure 4) and noisy (as stated in Section 4.3) data of Cotocollao than the NN. The similar performance in both sites means that this model has the potential to be applied in various situations with similar expected error rates. Further development

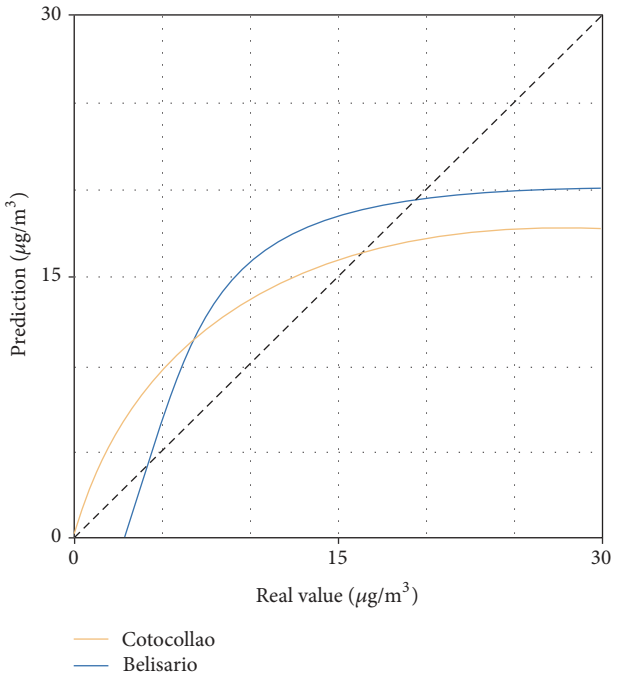


FIGURE 10: Fitted lines representing the correlation between predicted values and real values through a NN algorithm for Cotocollao (orange) and Belisario (blue).

should aid in qualifying the true robustness of this approach by exploiting the possibility of modeling with other spatial dependencies, such as density of measurements and day-by-day shifts, which represent the degree of freedom of parameters related to readings of the previous day(s). The latter dependency could be combined with linear quadratic estimation (LQE) techniques such as Kalman filters to improve the precision.

6. Conclusions and Perspectives

This study proposes a machine learning approach to predict $PM_{2.5}$ concentrations from meteorological data in a high-elevation mid-sized city (Quito, Ecuador). Standard levels of fine particulate matter are classified by using different machine learning models. This classification is performed on six years' records of daily meteorological values of wind speed (m/s), wind direction (0–360°), and precipitation accumulation (mm) for two air quality monitoring sites located in Quito (Cotocollao and Belisario). Although these sites are both in Quito's urbanized area, they exhibit differences in spread and dominance regarding wind features (speed and direction) that account for high $PM_{2.5}$ concentrations and distribution of pollution levels over the years. This could be caused by the fact that Belisario is more urbanized than Cotocollao and more importantly due to the extremely complex terrain of the city.

For these two different districts the results show a high reliability in the classification of low (<10 μg/m^3) versus high (>25 μg/m^3) and low (<10 μg/m^3) versus moderate (10–25 μg/m^3) $PM_{2.5}$ concentrations. We found well defined clusters, within the parameter space, for $PM_{2.5}$ concentrations < 10 μg/m^3. The regression analysis shows that the used parameters can predict $PM_{2.5}$ concentrations up to 20 μg/m^3 and the accuracy of the predictions is improved in conditions of strong winds and high precipitation for both Cotocollao and Belisario. There is a significant positive correlation between the real concentrations and the predicted concentrations for all the study period. The slightly higher correlation during the rainy season confirms that the model can predict $PM_{2.5}$ concentrations better for more extreme weather conditions.

Using a convolutional based spatial representation (CGM) to perform regression shows improving performance compared to various used machine learning algorithms (NN, L-SVM, and BT). In addition to this model, finding trends over periods of time with the use of time series algorithms could further improve the prediction and would make a long-term forecasting of $PM_{2.5}$ concentrations possible [13].

The main contribution of this study is to propose an alternative approach to chemical transport numerical modeling, such as WRF-Chem or CMAQ, the performance of which depends on several input parameters (emission inventory, orography, etc.) and the accuracy of built-in meteorological models (WRF, MM5). The application of numerical models for complex terrain regions is challenging, since important topographic features are not well represented [11, 33]. This produces imprecisions in not only forecasting air quality, but also relevant meteorology [10, 12, 34, 35]. Here, the proposed model provides a more reliable and more economical alternative to predict $PM_{2.5}$ levels, as it only requires meteorological data acquisition. In addition, accurate meteorological technology is far more affordable compared to air quality sensors that can exceed the price over 100 times. Finally, this model is based on the three basic meteorological parameters (wind speed, wind direction, and precipitation), which have a straightforward effect on pollution. Thus, by considering that our model has a good prediction efficiency for a city of such a complex topography, we argue that it could be successfully applied in other tropical locations (regions of reduced changes in solar angle, temperature, and relative humidity).

Also, this work provides an insight into the main limitations regarding $PM_{2.5}$ prediction from meteorological data and machine learning. The classification and regression show that concentrations > 20 μg/m^3 seem to be influenced more by additional parameters than the meteorological factors used in this study. For example, although daily temperature, solar radiation, and pressure do not vary much during the year, they might make a difference if analyzed during different times of the day, causing different pollution levels in the city. An interesting approach to tackle this limitation would be to consider a hybrid model that would mix a numerical method (WRF-Chem or CMAQ) with machine learning algorithms [10].

Other climatic conditions and unusual impactful events causing higher pollution levels (festivities, wild fires, accidents, seasonal variability, or natural calamities) could also explain changes in $PM_{2.5}$ concentrations exceeding 20 μg/m^3.

Future work will consist of identifying the parameters or events causing values above this threshold. Furthermore, we intend to improve our CGM and use it to classify outliers and find their cause. Considering the diverse machine learning models used in air quality prediction, such as Neural Network [13–15], regression [18], decision trees, and Support Vector Machine [17], we applied and tested most of these classifiers in this study. Alternative approaches to improve the accuracy of our model would consist of performing a prediction based on an ensemble of different algorithms of data processing and modeling [16, 17, 22].

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

The authors would like to thank David R. Sannino for editing the text.

References

[1] United Nations, Department of Economic and Social Affairs (2015). World Population Prospects, the 2015 Revision, in Population Division edited, UN.

[2] World Health Organization, Media Centre (2016). Air pollution levels rising in many of the world's poorest cities. http://www.who.int/mediacentre/news/releases/2016/air-pollution-rising/.

[3] J. Lelieveld, J. S. Evans, M. Fnais, D. Giannadaki, and A. Pozzer, "The contribution of outdoor air pollution sources to premature mortality on a global scale," *Nature*, vol. 525, no. 7569, pp. 367–371, 2015.

[4] C. A. Pope and D. W. Dockery, "Health effects of fine particulate air pollution: lines that connect," *Journal of the Air and Waste Management Association*, vol. 56, no. 6, pp. 709–742, 2006.

[5] Y. Rybarczyk and R. Zalakeviciute, "Machine learning approach to forecasting urban pollution: a case study of Quito," in *Proceedings of the IEEE Ecuador Technical Chapters Meeting, (ETCM '16)*, Guayaquil, Ecuador, 2016.

[6] M. A. Pohjola, A. Kousa, J. Kukkonen et al., "The spatial and temporal variation of measured urban PM_{10} and $PM_{2.5}$ in the Helsinki metropolitan area," *Water, Air and Soil Pollution: Focus*, vol. 2, no. 5, pp. 189–201, 2002.

[7] Y. Li, Q. Chen, H. Zhao, L. Wang, and R. Tao, "Variations in pm10, pm2.5 and pm1.0 in an urban area of the sichuan basin and their relation to meteorological factors," *Atmosphere*, vol. 6, no. 1, pp. 150–163, 2015.

[8] J. Wang and S. Ogawa, "Effects of meteorological conditions on PM2.5 concentrations in Nagasaki, Japan," *International Journal of Environmental Research and Public Health*, vol. 12, no. 8, pp. 9089–9101, 2015.

[9] F. Zhang, H. Cheng, Z. Wang et al., "Fine particles (PM2.5) at a CAWNET background site in central China: chemical compositions, seasonal variations and regional pollution events," *Atmospheric Environment*, vol. 86, pp. 193–202, 2014.

[10] X. Xi, Z. Wei, R. Xiaoguang et al., "A comprehensive evaluation of air pollution prediction improvement by a machine learning method," in *Proceedings of the 10th IEEE International Conference on Service Operations and Logistics, and Informatics, SOLI 2015 - In conjunction with ICT4ALL '15*, pp. 176–181, Hammamet, Tunisia, November 2015.

[11] P. A. Jimenez and J. Dudhia, "Improving the representation of resolved and unresolved topographic effects on surface wind in the WRF model," *Journal of Applied Meteorology and Climatology*, vol. 51, no. 2, pp. 300–316, 2012.

[12] R. Parra and V. Díaz, "Preliminary comparison of ozone concentrations provided by the emission inventory/WRF-Chem model and the air quality monitoring network from the Distrito Metropolitano de Quito (Ecuador)," in *Proceedings of the 8th annual WRF User's Workshop*, NCAR, Boulder, Colo, USA.

[13] X. Ni, H. Huang, and W. Du, "Relevance analysis and short-term prediction of PM2.5 concentrations in Beijing based on multi-source data," *Atmospheric Environment*, vol. 150, pp. 146–161, 2017.

[14] J. Chen, H. Chen, Z. Wu, D. Hu, and J. Z. Pan, "Forecasting smog-related health hazard based on social media and physical sensor," *Information Systems*, vol. 64, pp. 281–291, 2017.

[15] J. Zhang and W. Ding, "Prediction of air pollutants concentration based on an extreme learning machine: the case of Hong Kong," *International Journal of Environmental Research and Public Health*, vol. 14, no. 2, p. 114, 2017.

[16] P. Jiang, Q. Dong, and P. Li, "A novel hybrid strategy for PM2.5 concentration analysis and prediction," *Journal of Environmental Management*, vol. 196, pp. 443–457, 2017.

[17] K. P. Singh, S. Gupta, and P. Rai, "Identifying pollution sources and predicting urban air quality using ensemble learning methods," *Atmospheric Environment*, vol. 80, pp. 426–437, 2013.

[18] C. Brokamp, R. Jandarov, M. B. Rao, G. LeMasters, and P. Ryan, "Exposure assessment models for elemental components of particulate matter in an urban environment: a comparison of regression and random forest approaches," *Atmospheric Environment*, vol. 151, pp. 1–11, 2017.

[19] M. Arhami, N. Kamali, and M. M. Rajabi, "Predicting hourly air pollutant levels using artificial neural networks coupled with uncertainty analysis by Monte Carlo simulations," *Environmental Science and Pollution Research*, vol. 20, no. 7, pp. 4777–4789, 2013.

[20] A. Russo, F. Raischel, and P. G. Lind, "Air quality prediction using optimal neural networks with stochastic variables," *Atmospheric Environment*, vol. 79, pp. 822–830, 2013.

[21] M. Fu, W. Wang, Z. Le, and M. S. Khorram, "Prediction of particular matter concentrations by developed feed-forward neural network with rolling mechanism and gray model," *Neural Computing and Applications*, vol. 26, no. 8, pp. 1789–1797, 2015.

[22] W. Sun and J. Sun, "Daily $PM_{2.5}$ concentration prediction based on principal component analysis and LSSVM optimized by cuckoo search algorithm," *Journal of Environmental Management*, vol. 188, pp. 144–152, 2017.

[23] United Nations Development Programme (UNDP), Human development report 2014, Sustaining Human Progress: Reducing Vulnerabilities and Building Resilience.

[24] Instituto Nacional de Estadistica y Censos (INEC), Quito, el cantón más poblado del Ecuador en el 2020, 2013.

[25] E. Acuña and C. Rodriguez, "The treatment of missing values and its effect on classifier accuracy," in *Classification, Clustering, and Data Mining Applications*, D. Banks, F. R. McMorris, P. Arabie, and W. Gaul, Eds., pp. 639–647, Springer, Berlin, Heidelberg, 2004.

[26] I. Mierswa, M. Wurst, R. Klinkenberg, M. Scholz, and T. Euler, "Yale: rapid prototyping for complex data mining tasks," in *Proceedings of 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 935–940, Philadelphia, PA, USA, 2006.

[27] C. A. Calder and N. Cressie, "Some topics in convolution-based spatial modeling," in *Proceedings of the 56th Session of the International Statistics Institute*, International Statistics Institute, Netherlands, 2007.

[28] F. Fouedjio, N. Desassis, and J. Rivoirard, "A generalized convolution model and estimation for non-stationary random functions," *Spatial Statistics*, vol. 16, pp. 35–52, 2016.

[29] J. Babaud, A. P. Witkin, M. Baudin, and R. O. Duda, "Uniqueness of the Gaussian kernel for scale-space filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 8, no. 1, pp. 26–33, 1986.

[30] MA, "Ministerio Del Ambiente: Norma de Calidad del Aire Ambiente o Nivel de Inmision Libro VI Anexo 4, 2015".

[31] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.

[32] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A hybrid approach to alleviating class imbalance," *IEEE Transactions on Systems, Man, and Cybernetics Part A:Systems and Humans*, vol. 40, no. 1, pp. 185–197, 2010.

[33] P. A. Jimenez and J. Dudhia, "On the ability of the WRF model to reproduce the surface wind direction over complex terrain," *Journal of Applied Meteorology and Climatology*, vol. 52, no. 7, pp. 1610–1617, 2013.

[34] A. Meij, A. De Gzella, C. Cuvelier et al., "The impact of MM5 and WRF meteorology over complex terrain on CHIMERE model calculations," *Atmospheric Chemistry and Physics*, vol. 9, no. 17, pp. 6611–6632, 2009.

[35] P. Saide, G. Carmichael, S. Spak et al., "Forecasting urban PM10 and PM2.5 pollution episodes in very stable nocturnal conditions and complex terrain using WRF-Chem CO tracer model," *Atmospheric Environment*, vol. 45, no. 16, pp. 2769–2780, 2011.