We are IntechOpen,
the world's leading publisher of
Open Access books
Built by scientists, for scientists

6,900
Open access books available

185,000
International authors and editors

200M
Downloads

154
Countries delivered to

Our authors are among the

TOP 1%
most cited scientists

12.2%
Contributors from top 500 universities

Selection of our books indexed in the Book Citation Index
in Web of Science™ Core Collection (BKCI)

Interested in publishing with us?
Contact book.department@intechopen.com

Numbers displayed above are based on latest data collected.
For more information visit www.intechopen.com

Regression Models to Predict Air Pollution from Affordable Data Collections

Yves Rybarczyk and Rasa Zalakeviciute

Abstract

Air quality monitoring is key in assuring public health. However, the necessary equipment to accurately measure the criteria pollutants is expensive. Since the countries with more serious problems of air pollution are the less wealthy, this study proposes an affordable method based on machine learning to estimate the concentration of $PM_{2.5}$. The capital city of Ecuador is used as case study. Several regression models are built from features of different levels of affordability. The first result shows that cheap data collection based on web traffic monitoring enables us to create a model that fairly correlates traffic density with air pollution. Building multiple models according to the hourly occurrence of the pollution peaks seems to increase the accuracy of the estimation, especially in the morning hours. The second result shows that adding meteorological factors allows for a significant improvement of the prediction of $PM_{2.5}$ concentrations. Nevertheless, the last finding demonstrates that the best predictive model should be based on a hybrid source of data that includes trace gases. Since the sensors to monitor such gases are costly, the last part of the chapter gives some recommendations to get an accurate prediction from models that consider no more than two trace gases.

Keywords: urban air pollution prediction, heterogeneous data sources, hybrid models, low-cost approach, real-time traffic monitoring, meteorological and chemical features

1. Introduction

Over the last century, the global human population has augmented more than four times. Most of the recent growth is accredited to the urban areas in the less developed parts of the world [1]. This has resulted in 80% of global cities and 98% of cities in low- and middle-income countries

to exceed the recommendations for air quality [2]. Apart from economic losses, reduced visibility, and climate change, ambient air pollution costs millions of premature deaths annually, mostly due to anthropogenic fine particulate matter ($PM_{2.5}$—particles with aerodynamic diameter less than 2.5 μm) [3]. In the case of business-as-usual, the global atmospheric chemistry models suggest that the contribution of outdoor air pollution to premature mortality could double by 2050 [4].

Even though the concentrations of $PM_{2.5}$ are 2–5 times higher in the developing countries, most of the air quality studies and measurements are concentrated in the developed countries [2, 5]. This is often due to the investments required to launch and support a reliable air quality monitoring station or network. High accuracy, standard air quality reference method equipment costs can range from $6000 to $36,000 per sensor [6], excluding the costs for maintenance, calibration and accessories, resulting in a price of a functional air quality monitoring station well over $100,000. Meteorological equipment is also essential for the evaluation of air quality, as high UV radiation, high winds, precipitation, or extreme temperatures can cause serious health concerns. Meteorological station, depending on accuracy requirements, can cost from $1000 to over $7000; although, the accuracy differences are not too great between the tiers (not including the lowest level equipment). Dynamic and nonhomogeneous urban systems contain different pollution sources, infrastructures, varying terrains, requiring more than one station for a comprehensive evaluation of air pollution conditions, consequently excluding poorer cities.

The question of economic limitations has recently been brought to attention resulting in the introduction of the lower cost sensors (<$500) or bundled platforms ($5000–10,000) to the market. Based on the comparative studies, evaluating sensor performance (fit for air quality monitoring), some air criteria pollutants compare quite well with the standard air quality reference methods, while some show lower correlation [6–8]. In addition, in some cases, adding a PM sensor to the platform increases costs significantly.

Recently, a different approach aims at using machine learning to estimate particulate pollution [9, 10]. This study proposes to evaluate the reliability for predicting air quality through a machine-learning approach and from data sources with a different scale of affordability. It focuses on the case study of Quito, the capital city of Ecuador, because it is a model example of complex terrain rapidly growing in mid-size cities in developing world with air pollution issues and economic limitations (e.g., poor quality fuel). In addition, Quito has many years of environmental data collection that can be used for data mining.

2. Machine-learning approach

2.1. Prediction by multiple regression

In regression, features derived from a dataset are used as input of the regression model to predict continuous valued output. This kind of prediction is obtained by learning the relationship between the input x and the output y. The simplest case of a regression model is a

simple regression, in which a single feature is used to estimate the value of the output. This relationship is acquired by fitting a linear or nonlinear curve to the data. In order to correctly fit the curve, it is necessary to define the goodness-of-fit metric, which allows us to identify the curve that fits better than the other ones. The optimization technique used in regression, and in several other machine-learning methods, is the gradient descent algorithm. In the case of a simple linear regression, the objective is to find the value of the slope and the intercept of the line that minimizes the goodness-of-fit metric. The residual sum of squares (*RSS*), also called sum of squared errors of prediction, is used to calculate this cost. The *RSS* adds up the squared difference between the estimated relationship between x and y (regression model) and the actual values of y (y_i), as described in Eq. (1)

$$RSS(w_0, w_1) = \sum_{i=1}^{N} \left(y_i - [w_0 + w_1 x_i] \right)^2 \qquad (1)$$

where N is the number of observations, x_i are the input values, and the coefficients w_0 and w_1 are the intercept and slope of the linear regression, respectively. For simplification, Eq. (1) is commonly rewritten as follows:

$$RSS(w_0, w_1) = \sum_{i=1}^{N} \left(y_i - \hat{y}_i(w_0, w_1) \right)^2 \qquad (2)$$

where $\hat{y}_i(w_0, w_1)$ is the predictive value of observation y_i, if a linear regression defined by w_0 and w_1 is used. In the case of a multiple regression model, more than one input (or feature) is considered to predict the output. The generic equation of such a model can be written as follows:

$$y_i = \sum_{j=0}^{D} w_j h_j(\vec{x}_i) + \varepsilon_i \qquad (3)$$

where D is the number of features, $h_j(\vec{x})$ are functions of the inputs (represented as a vector) that are weighted by different coefficients w_j, and ε_i is the error. Thus, the RSS is generically defined by Eq. 4 as

$$RSS(\vec{w}) = \sum_{i=1}^{N} \left(y_i - \hat{y}_i(\vec{w}) \right)^2 \qquad (4)$$

where $\vec{w}$ is a vector of the weights (or coefficients) of the whole parameters of the fit. The best regression model is the function that provides the smallest RSS. The model is obtained after a split of the dataset into two independent sets: a training set and a test set. The training set is used to build the model, and the calculation of the RSS is performed over the test set, only. The gradient descent is an iterative method that minimizes the RSS metric. It takes multiple steps to eventually provide the optimal solution as described in Algorithm 1. At first, all the parameters are initialized to be zero at the first iteration (t = 1). Then, the algorithm repeats while the magnitude of the RSS does not converge. The internal part of the loop calculates the partial derivative (partial[j]) for each feature of the multiple regression model, and then, the gradient step takes the jth coefficient at time t and subtracts the step size (η) times that partial

derivative. Once the algorithm cycled through all the features of the model, the t counter is incremented and the convergence condition is tested to decide whether the program must loop through or not. When the minimum is reached (RSS ≤ ε), the respective values of the regression coefficients are used as the model parameters to form the predictions.

Algorithm 1. Gradient descent algorithm for multiple regression.

1: init $\vec{w}^{(1)} = 0$, t = 1

2: while $\| \nabla \; RSS(\vec{w}^{(t)}) \| > \varepsilon$.

3:       for j = 0, …, D

4:           partial[j] = $-2\sum_{i=1}^{N} h_j(\vec{x}_i)(y_i - \hat{y}_i(\vec{w}^{(t)}))$

5:           $\vec{w}_j^{(t+1)} \leftarrow \vec{w}_j^{(t)} - \eta \text{partial[j]}$

6:       t ← t + 1

In addition, the final regression models of this study are obtained after an attribute selection using the M5 method, which steps through the attributes removing the one with the smallest standardized coefficient until no improvement is observed in the estimate of the error given by the Akaike information criterion (AIC) [11].

$$AIC \; = \; N \, ln\left(\frac{RSS}{N-D}\right) + 2D \tag{5}$$

where N is the number of observations (or instances), and D is the number of features (or attributes). The selected model is the model that gets the lowest AIC.

All the models presented in the manuscript are obtained after a normalization of the value of the variables, in order to avoid a dominance of the variables with the highest intrinsic values. The used method to evaluate the model accuracy is a 10-fold cross-validation. The regression modeling is performed with Pandas and scikit-learn machine-learning library for Python.

2.2. Cumulative modeling method

Air pollution data ($PM_{2.5}$) were collected in central Quito over a period of 2 months in June and July of 2017 by the city Secretariat of the Environment. Belisario (alt. 2835 m.a.s.l, coord.78°29′24″ W, 0°10′48″ S) measurement station was setup following the criteria of the Environmental Protection Agency of the United States (USEPA). For $PM_{2.5}$ concentration data Thermo Scientific FH62C14-DHS continuous ambient particulate monitor 5014i was used based on beta rays' attenuation method (EPA No. EQPM-0609-183). For all the data 1 hour averages were calculated, resulting in 1118 instances.

In this work, we present several regression models to provide a reliable estimation of the current level of $PM_{2.5}$ from data collection methods of different levels of affordability. In Section 3, we describe a prediction of $PM_{2.5}$ concentrations based on real-time traffic moni-toring, only. This type of data does not cost anything to the user as it is based on publicly available worldwide traffic data. Section 4 describes a prediction that adds meteorological

factors on top of the traffic data. Most of the meteorological equipment is not as costly as air quality sensors, thus still presenting a viable option for the prediction of $PM_{2.5}$ concentrations. Subsequently, Section 5 describes a prediction that includes traffic data, meteorological factors and trace gas concentrations. This way we build from the simplest to the most complex model, increasing the equipment costs with every step and improving the prediction performance. Finally, we finish our study by proposing the best simple model based on a feature selection method, letting us to reduce the costs significantly, but still producing a high performance.

3. Prediction from real-time traffic monitoring

We propose a method to extract data from Google Maps Traffic, in which a simple request to the website enables us to build a database regarding the traffic in the city and, consequently, the level of urban air pollution.

3.1. Dataset

3.1.1. Data acquisition

3.1.1.1. Screenshot

A request to Google Maps Traffic is performed by the use of the library selenium for Python. A screenshot is carried out each 10 minute in a specific zone of Quito, which is centered on the neighborhood of Belisario. The exact coordinates of the geographic area of interest are −0.181661, −78.4987077, which is 1.2 km southwest from the center of the traffic map. Two kinds of images are stored: the one with traffic (**Figure 1a**) and the another without traffic (**Figure 1b**). It is necessary to save these two different types of pictures in order to proceed with the next step that consists of isolating the traffic information only (**Figure 1c**).

3.1.1.2. Background subtraction

A technique of background subtraction is used to eliminate picture information that is not related to traffic (**Figure 1**). The background removal is carried out through the process as follows [12]:

- Memorize the background image (picture without traffic).

- Check every pixel in the frame. If it is different from the corresponding pixel in the background image, it is a foreground pixel (traffic information). If not, it is a background pixel.

To get a clean image of the traffic, it is necessary to define a distance threshold of brightness when comparing the background image to the traffic + background images (see Algorithm 2). For every pixel, if the absolute difference of brightness between the image with traffic and the background image is lower than the threshold (empirically defined at 30), then the corresponding pixels are considered identical. In this case, the pixels are colored white
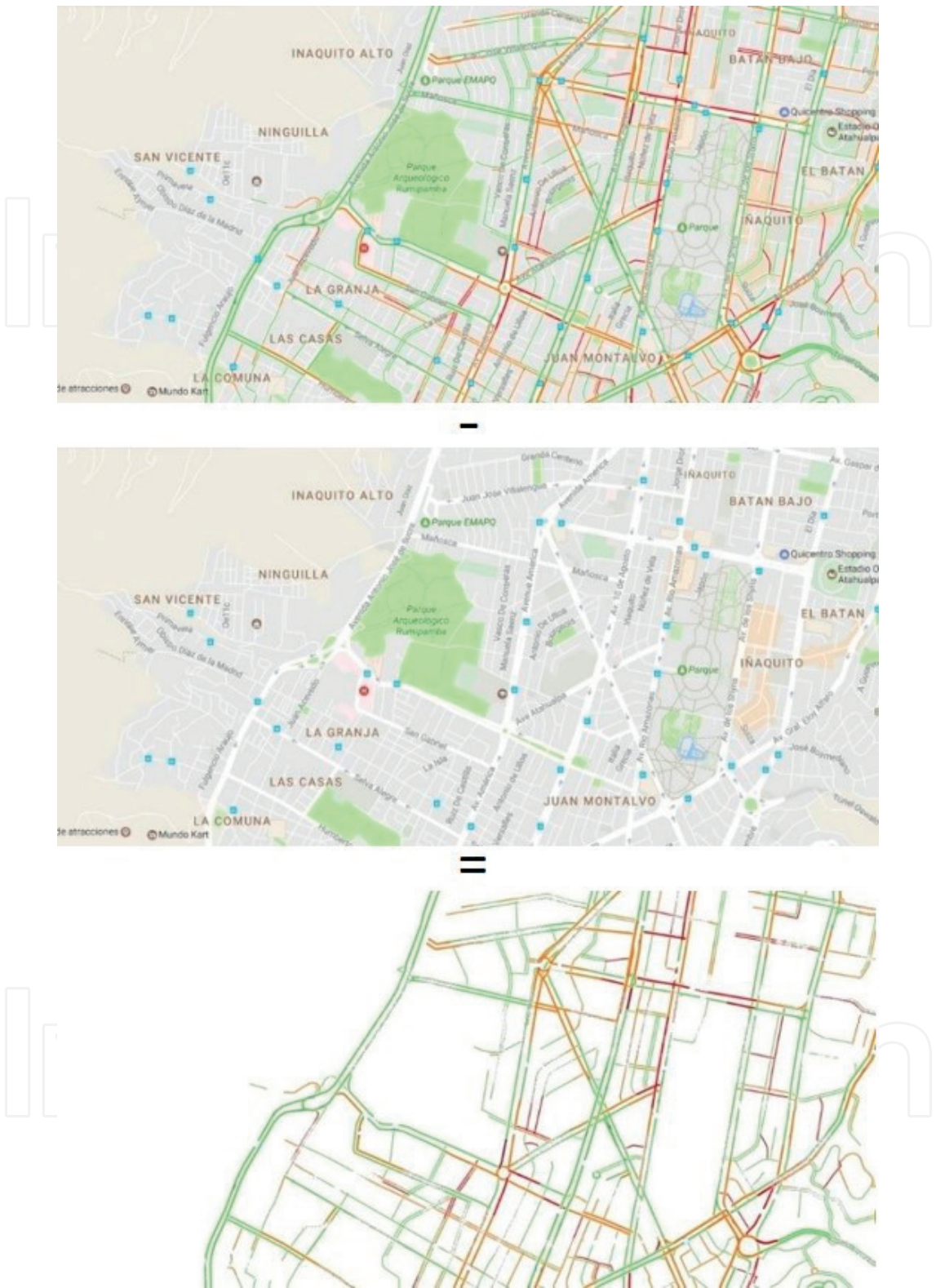
Figure 1. Description of the principle of background removal. The background image (b) is subtracted from the image that includes the traffic (a). The result is a picture with the traffic information only (c).

(lines 14 and 15 of Algorithm 2). On the contrary, if the brightness difference is higher than the threshold, the color of the pixel does not change (lines 16 and 17 of Algorithm 2). Thanks to this method, it is possible to extract only the color information of the traffic.

The calculation of the value of the difference is based on the computation of the distance between each color component of a pixel (RGB). In other words, colors are considered as points in a three-dimensional space (line 13 of Algorithm 2).

Algorithm 2. Generating the background subtraction.

1: for(int i = 0, i < allFrame, i++)

2:      for(int x = 0, x < width, x++)

3:          for(int y = 0, y < height, y++)

4:              int. pos = x + y * width

5:              color frameColor = frame[i].pixels[pos]

6:              color refColor = background.pixels[pos]

7:              float rFrame = red(frameColor)

8:              float gFrame = green(frameColor)

9:              float bFrame = blue(frameColor)

10:             float rRef = red(refColor)

11:             float gRef = green(refColor)

12:             float bRef = blue(refColor)

13:             float diff = dist(rFrame, gFrame, bFrame, rRef, gRef, bRef)

14:             if (diff <30)

15:                 image.pixels[pos] = color(255)

16:             else

17:                 image.pixels[pos] = frame[i].pixels[pos]

3.1.1.3. Pixel extraction

To identify traffic density, three categories of pixel colors are extracted: green, orange, and red (see Algorithm 3). The green, orange, and red pixels mean low, medium, and high amount of traffic, respectively. The pixel number of each category is obtained by getting the RGB component of the whole pixels in the image. After excluding the white pixels (line 6 of Algorithm 3), three rules are implemented to classify the remaining pixels in one or another category (lines 7 to 12 of Algorithm 3). Once the picture is entirely read, the percentage of each category is calculated by dividing the number of green, orange, and red pixels by the total number of colored pixels.

Algorithm 3. Generating the pixel color extraction.

1: for(int i = 0, i < allFrame, i++)

2:      float red = 0, orange = 0, green = 0

```
3:      image(frame[i], 0, 0)

4:      for(int x = 0, x < width, x++)

4:         for(int y = 0, y < height, y++)

5:            color c = get(x, y)

6:               if(red(c) < 200 || green(c) < 200 || blue(c) < 200)

7:                  if(red(c) > green(c) && abs(green(c)-blue(c)) < 20)

8:                     red++

9:                  else if(red(c) > green(c) && green(c) > blue(c))

10:                    orange++

11:                 else if(green(c) > red(c) && red(c) > blue(c))

12:                    green++
```

3.1.1.4. Hourly averaging

Since the machine-learning models are based on hourly data analysis, it is required to determinate for each hour the trend of the six 10 minute recording. To do so, the average of the six percentages per hour and for each color is calculated. Then, these values are added into the final dataset.

3.1.2. Data transformation

A last data preparation is necessary before running the machine-learning algorithms. The polar coordinates of time (think of time as an analog clock of 24 × 60 minutes, in which minute hand describes an angle) are transformed into Cartesian coordinates (Eqs. (6) and (7)). This mathematical transformation permits a more accurate feature representation of the data with respect to the traffic density at night. Otherwise, it would be impossible to find a correlation between time and traffic around midnight, since a similar traffic would correspond to a completely different number of minutes (before midnight ≈ 1440 minute, and after midnight ≈ 0 minute). This transformation is particularly relevant for machine-learning algorithms based on linear regression, because it relies on a continuous relationship between parameters [13].

$$Xminutes \ = \ cos\left(\frac{minutes \ . \ \pi}{720}\right) \tag{6}$$

$$Yminutes \ = \ sin\left(\frac{minutes \ . \ \pi}{720}\right) \tag{7}$$

Thus, the final dataset is composed of a number of five features, which are: Xminutes, Yminutes, %orange, %red, and $PM_{2.5}$ (= feature to predict). The %green can be discarded, because it provides a redundant data with the information brought by %orange and %red.

3.2. Single models

Two possible approaches can be considered to predict the level of $PM_{2.5}$ from other attributes. The first one is to build a single model for the whole day. Another approach is to consider several successive models, since the human activity and the atmospheric conditions change during the day. This section presents the former method.

A machine-learning algorithm based on a linear regression, as described in Section 2.1, is applied on the dataset. The models are trained and tested according to a 10-fold cross-validation technique. Then, the performance of the models is assessed by two metrics: the correlation coefficient and the root-mean-squared error (RMSE). The correlation coefficient (r) measures the strength of the linear relationship between two or more variables. The advantage of r over the other metrics is to be based on a scale with a maximum (±1) and a minimum (0) to quantify the strength of the relationship. The closer to 1 is the absolute value of r, the better is the correlation. The root-mean-squared error (RMSE) is the square root of the averaged squared error per prediction (MSE). RMSE is an intuitive evaluation metric that is frequently used, because it provides a performance in the same unit as the predicted attribute itself. The lower is the value of RMSE, the more accurate is the model prediction.

3.2.1. Time only

Since the transportation is the main source of pollution in Quito, and this human activity is relatively stereotypic all day long, the simplest approach is to build a predictive model of $PM_{2.5}$ based on time parameters, only. In this case, the number of features is limited to three, which are Xminutes, Yminutes, and $PM_{2.5}$.

The linear regression model obtained after running the algorithm is as follows:

$$PM_{2.5} =$$
$$-2.2242 \ * \ Xminutes \ +$$
$$-1.7366 \ * \ Yminutes \ +$$
$$13.8294$$

The prediction accuracy of the model is evaluated as

$$r = 0.21$$
$$RMSE = 8.76$$

In the present model, the coefficients attributed to both features are negative. It means that the higher are the two temporal attributes, the lower are the concentrations of fine particulate matter. However, the performance of this first model is quite low ($r \approx 0.2$). This is confirmed by the value of the RMSE, which is around nine out of an average level of $PM_{2.5} = 13.8$ µg/m^3 for the studied period.

3.2.2. Time and traffic

The result of the previous model suggests that it is necessary to consider additional information, such as traffic data, to improve the prediction accuracy of the regression model. To do so, the present analysis takes into account the traffic information provided by Google Maps and processed as described in Section 3.1.1. Thus, the used dataset is composed of five parameters, which are Xminutes, Yminutes, %red, %orange, and $PM_{2.5}$.

The linear regression model obtained after running the algorithm is as follows:

$$
\begin{aligned}
PM_{2.5} = \quad & \\
1.2093 \quad & * \quad Xminutes \quad + \\
2.0369 \quad & * \quad Yminutes \quad + \\
-23.3875 \quad & * \quad \%red \quad + \\
40.6166 \quad & * \quad \%orange \quad + \\
7.0578 &
\end{aligned}
$$

The prediction accuracy of the model is evaluated as

$$r = 0.32$$

$$RMSE = 8.48$$

The model shows that the parameters with the highest weight is %orange. It means that the quantification of the medium amount of traffic is an important feature to estimate the level of $PM_{2.5}$. It is to note that this model, which includes data regarding human activity (i.e., transportation), provides a higher prediction accuracy than a model based on temporal information, only.

3.2.3. Traffic only

One of the main objectives of a machine-learning approach is to produce the most accurate prediction with a model as simple as possible. Since the temporal features seem to have a lower weight than the traffic features, we propose to build a model based on traffic only and assessing its reliability. Here, the number of attributes is three: %orange, %red, and $PM_{2.5}$.

The linear regression model obtained after running the algorithm is as follows:

$$
\begin{aligned}
PM_{2.5} = \quad & \\
-18.8914 \quad & * \quad \%red \quad + \\
28.618 \quad & * \quad \%orange \quad + \\
9.2185 &
\end{aligned}
$$

The prediction accuracy of the model is evaluated as

$$R = 0.31$$

$$RMSE = 8.51$$

Again, the model shows that the weight of the %orange parameter is the largest. The higher is the medium amount of traffic, the higher is the level of $PM_{2.5}$. In terms of performance, this model based on two predictive features has an accuracy similar as the previous model with four features ($r \approx 0.3$ in both cases).

3.2.4. Simple regression model

Since the %orange parameter is the attribute with the highest weight, it would be possible to build a predictive model of $PM_{2.5}$ based on a simple regression. The advantage of such a model is its simplicity and the fact that it is visually interpretable from a bidimensional graph (see **Figure 2**). Thus, the used dataset for this analysis has two features, only: %orange and concentrations of $PM_{2.5}$.

The linear regression model obtained after running the algorithm is as follows:

$$PM_{2.5} = 20.1012 * \%orange + 9.6609$$

The prediction accuracy of the model is evaluated as

$$r = 0.31$$
$$RMSE = 8.53$$

The simple regression model and **Figure 2** show a growing trend of the level of $PM_{2.5}$ when the %orange parameter increases. This very elementary model (a single predictive feature) allows for a prediction performance quite comparable with the two preceding models ($r \approx 0.3$), which are more complex (four and two predictive features, respectively).

3.2.5. Interpretation of the results

The performance accuracy of the models evaluated by a metric in terms of correlation coefficient and RMSE between traffic and $PM_{2.5}$ is slightly above 0.3 and around 8.5, respectively. The models that consider traffic monitoring provide a higher accuracy than a model based on time only. This result means that traffic is more reliable than time to predict air quality. This difference could be reduced if the weekends (air pollution levels usually low) are excluded, since the traffic is quite stereotypic during the workdays. Also, the accuracy of a model based on traffic monitoring is not significantly improved by adding the time of day, because this information is mostly redundant with the traffic data.

Overall, it seems that Google Maps Traffic can provide a fair information to predict the level of $PM_{2.5}$. From this data source, the number of orange pixels (medium amount of traffic) would be the most relevant feature. It could be explained by the fact that the medium traffic has the largest amplitude of variation all day long, and thus, this is the category that best represents the traffic density in the city. Nevertheless, the accuracy of the model could be improved if we consider an air pollution modeling based on several daily models, defined by the variation of air pollution levels all day long (two peaks a day), instead of a single one.
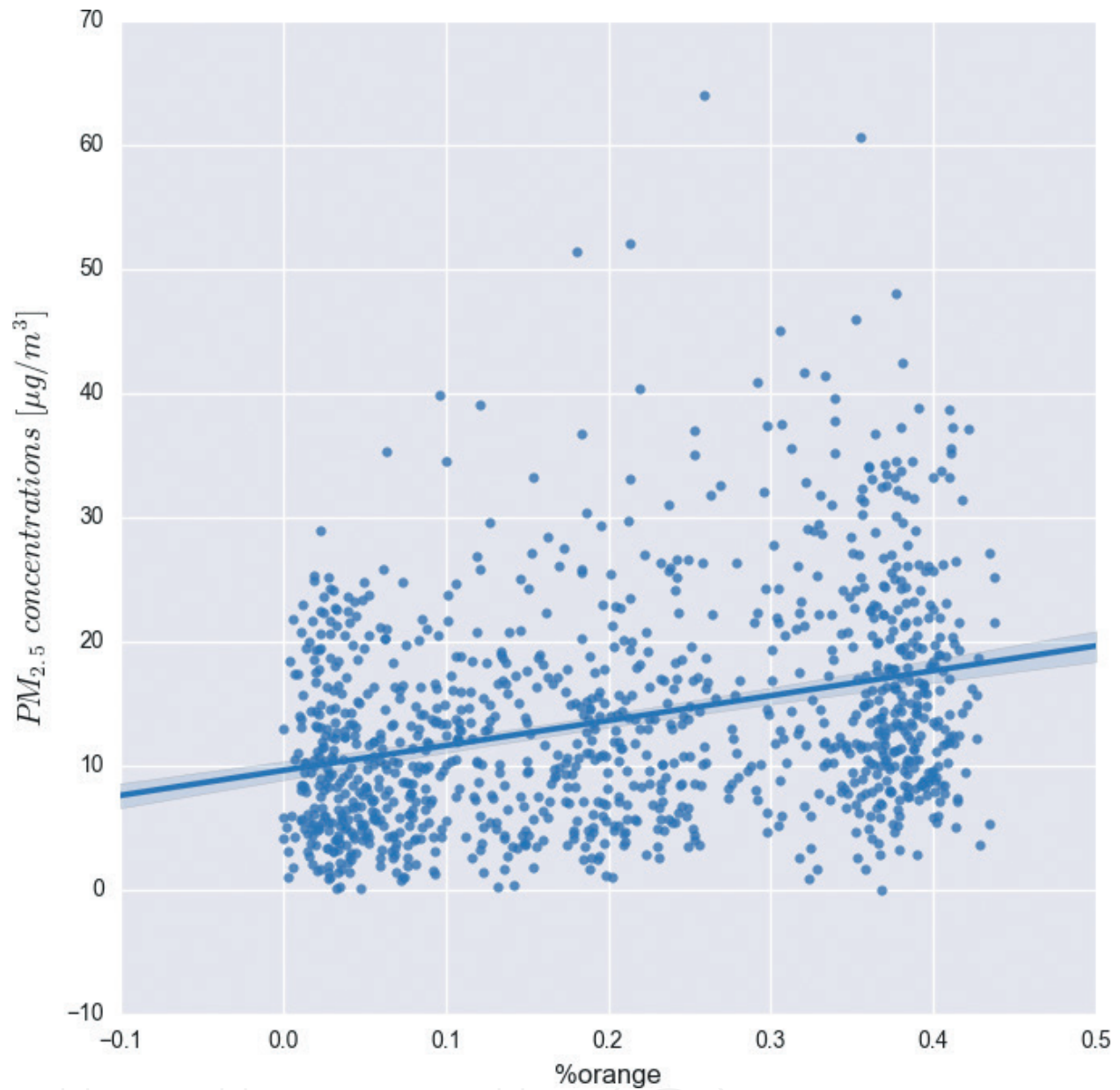
Figure 2. Representation of the value of PM$_{2.5}$ against the ratio of medium traffic (each dot is an observation) and the respective simple linear regression between these two features (line). The higher is the medium amount of traffic (%orange), the larger is the concentration of fine particulate matter (PM$_{2.5}$).

3.3. Multiple models

In the city of Quito as in most of the cities worldwide, there are two peaks of PM$_{2.5}$ pollution during the day. The first peak is in the morning (around 10 am) and the second is in the evening (around 7 pm). **Figure 3** is a graphical representation of the two daily peaks of fine particulate contamination averaged over the last 10 years (2007–2016) for the district of Belisario (These peaks occur approximately at the same time in any district of Quito.) During the morning hours, the rush hour actually lasts longer than the visible PM$_{2.5}$ concentration peak, but a sudden decline can be observed due to the deepening of the planetary boundary

layer (PBL). PBL growth during the day is dependent on the solar heating of the surface and thus induced vertical mixing. The depth of maximum PBL can vary from 1 day to another due to the difference in solar radiation intensity, solar angle, and especially cloud cover [14]. PBL is shallow in the morning (up to a few hundred meters) and deepens during the day reaching up to few kilometers [15]. This has a consequence on the level of air contaminants, which are less diluted in the morning than in the afternoon. All of these variations would reduce the performance of a single regression model a day to predict $PM_{2.5}$ from the vehicle emissions in the city. Thus, the present section describes a prediction of fine particulate matters from three daily models determined by the two peaks of pollution, such as a morning model [6–10 h], a midday model [10–14 h], and an afternoon model [14–19 h]. It is not necessary to consider a night model, because the level of air pollution drops during this period.

3.3.1. Morning model

The morning model is defined between 6 am (360th minute) and 10 am (600th minute). **Figure 3** shows that there is a constant increase in the $PM_{2.5}$ concentration during this period. The two main factors that should explain this increase are the traffic intensification and the low morning PBL. If this assumption is correct, then the predictive accuracy of a regression model that considers traffic
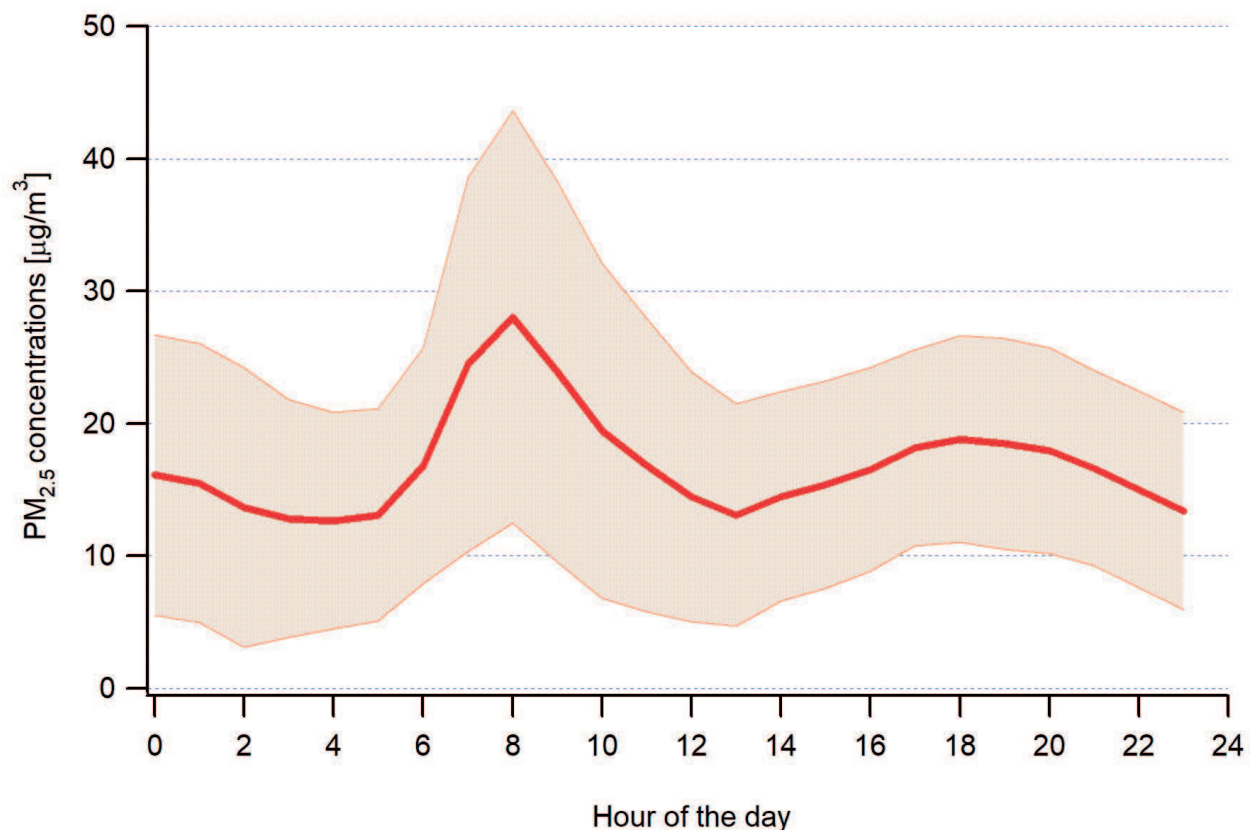

Figure 3. Typical profile of the $PM_{2.5}$ concentrations during the day in the Belisario district of Quito (2007–2016 data). Although, a slight reduction in the level of pollution was observed throughout the years, the air contamination peaks are always located at the same time of day (around 10 am and 7 pm).

data as features should be improved in comparison with the single models. The characteristic of the used dataset is as follows: 110 instances and 4 features (minutes, %red, %orange, and $PM_{2.5}$).

The linear regression model obtained after running the algorithm is as follows:

$$
\begin{aligned}
PM_{2.5} = \quad & \\
& 0.0444 * \text{minutes} + \\
& -123.0175 * \%\text{red} + \\
& 89.1856 * \%\text{orange} + \\
& -15.4187
\end{aligned}
$$

The prediction accuracy of the model is evaluated as

$$r = 0.49$$

$$RMSE = 10.13$$

As observed in the single model approach, the weights of the traffic attributes are significantly larger than the coefficient of time. The most representative feature, which is %orange, shows that the higher is the medium amount of traffic, the higher is the value of $PM_{2.5}$. In terms of performance, the prediction accuracy is around 0.5, for the correlation coefficient, and around 10 out of an average value of $PM_{2.5} = 17.4$ µg/m^3, for the RMSE. As hypothesized, this limited analysis on a morning window provides a regression model more accurate than the models based on the full day.

3.3.2. Midday model

The midday model is defined between 10 am (600th minute) and 2 pm (840th minute). **Figure 3** shows that there is a constant decrease in the $PM_{2.5}$ concentration during this period. The two main factors that should explain this drop are the traffic diminution and the elevation of the PBL that increases the dilution of air contaminants. In such a situation, the correlation between traffic and $PM_{2.5}$ should decrease. Here, the regression algorithm is applied on a dataset composed of 116 instances and 4 features (minutes, %red, %orange, and $PM_{2.5}$).

The linear regression model obtained after running the algorithm is as follows:

$$
\begin{aligned}
PM_{2.5} = \quad & \\
& -0.0354 * \text{minutes} + \\
& -68.1378 * \%\text{red} + \\
& 55.4262 * \%\text{orange} + \\
& 35.2107
\end{aligned}
$$

The prediction accuracy of the model is evaluated as

$$r = 0.29$$

$$RMSE = 10.36$$

The coefficients of the resulting model are lower than in the morning model, for all the features. It suggests that the weight of the traffic data to predict $PM_{2.5}$ is less important at midday than in the morning, as hypothesized. It is confirmed by the performance evaluation of the model, which is similar as the accuracy obtained from the single models ($r \approx 0.3$).

3.3.3. Afternoon model

The afternoon model is defined between 2 pm (840th minute) and 7 pm (1140th minute). **Figure 3** shows that there is a constant increase in the $PM_{2.5}$ concentration, although the evening peak is lower than the morning peak due to the fact that the PBL has reached its peak and is not changing at this time of day, until a nocturnal boundary layer starts forming due to the absence of surface heating. Besides the elevated PBL, the air pollution increases because of the traffic growth at the end of the day. Again, the important dilution of pollutants in the atmosphere should reduce the correlation between traffic and $PM_{2.5}$ concentrations. The used dataset to build the model is as follows: 145 instances and 4 features (minutes, %red, %orange, and $PM_{2.5}$).

The linear regression model obtained after running the algorithm is as follows:

$$PM_{2.5} =$$
$$0.0242 \quad * \quad minutes \quad +$$
$$20.7938 \quad * \quad \%orange \quad +$$
$$-14.6845$$

The prediction accuracy of the model is evaluated as

$$r \quad = 0.28$$
$$RMSE = 7.65$$

The feature with the maximum weight in the afternoon model is still %orange, although its value continues to decrease. The time coefficient is extremely low, and %red is filtered by the M5 attribute selection method. As expected, the model accuracy assessed by the correlation coefficient is relatively low ($r \approx 0.3$). It means that the traffic input is not a good predictor to estimate the level of $PM_{2.5}$ in the afternoon. The important dilution of the air contaminants in the atmosphere would explain this result. Surprisingly, the RMSE (<8) is lower than in the two previous models (>10). This reduced error of prediction can be explained by the lower standard deviation (SD) of the $PM_{2.5}$ values in the afternoon (SD = 8) than in the morning (SD = 11.6) and midday (SD = 10.8). In other words, the better power of prediction is not due to the reliability of the model per se (essentially based on the traffic), but due to the limited variation in the $PM_{2.5}$ concentrations in the afternoon.

3.3.4. Interpretation of the results

There is a significant improvement in the prediction of $PM_{2.5}$ in the morning ($r \approx 0.5$). The performance can be explained by the fact that the PBL is relatively low in the morning. Thus, the pollution dilution is reduced and consequently the level of $PM_{2.5}$ becomes strongly

correlated with the pollution produced by the vehicles. The higher is the traffic activity, the higher is the concentration of fine particulate matter (see the high weight of the %orange parameter).

For the two other models, the accuracy is around the same value as a global model (r ≈ 0.3). Their predictive performance seems reduced, because the depth of the PBL increases with the augmentation of the solar radiation (maximal around noon). The poor power of prediction of these two models would be caused by the reduction of the influence of the traffic on the level of $PM_{2.5}$, since the weight of the %orange parameter drops at midday and afternoon.

Nevertheless, the average performance of an approach based on three models per day provides an accuracy slightly better than the single model (see Eq. (8)). It suggests that the best prediction of $PM_{2.5}$ from the traffic monitoring is obtained by analyzing the typical daily fluctuation of $PM_{2.5}$ concentration and applying a specific model according to the occurrence of the pollution peaks, especially in the morning.

$$\bar{r} = \frac{0.49 + 0.29 + 0.28}{3} = 0.35 \tag{8}$$

This performance could be further improved by analyzing a reduced image of the traffic map that closely matches the footprint of $PM_{2.5}$ concentrations measured by the monitoring station. In this study, the used picture represents an area of 22.4 km^2 and the footprint area for Belisario station (monitoring station height 10 m) would be around 3 km^2, only [16]. However, we chose a bigger traffic map area to have a more representative traffic situation of the city.

4. Adding meteorological factors

The ambient air pollution levels are mainly modulated by meteorological conditions [9, 17]. Consequently, considering these parameters in a model should improve the prediction of the concentration of fine particulate matter. Since the required equipment to proceed with the recording of these data is significantly cheaper than the air quality sensors, we present models that can predict the level of $PM_{2.5}$ from the selected meteorological features as follows: solar radiation (SR), temperature (T), pressure (P), precipitation (rain), relative humidity (RH), wind speed (WS), and wind direction (WD).

4.1. Dataset

4.1.1. Data acquisition

Seven meteorological parameters (wind speed and direction, temperature, relative humidity, atmospheric pressure, precipitation, and solar radiation) were measured using Vaisala WXT536 instrumentation, with an exception of Kipp&Zonnen netradiometer to measure solar radiation. To get the hourly value of SR, T, P, rain, RH, and WS, we simply have to calculate the average value from the six records per hour of the used dataset (one record each 10 minutes).

However, the calculation of the WD is a bit more complex. It is not possible to compute the mean direction per hour, because it can provide a completely wrong result. For instance, if the wind angle is four times around the east (90°) and the two other times is around the west (270°), the mean WD will be the south-southeast (150°), even if the wind never originated in that direction. To tackle this issue, the calculation of the most representative WD for each hour is carried out through the process as follows:

- Sampling of the WD to transform continuous values into discrete values.

- Fit a normal distribution to the data.

- Take the mean of the Gaussian as the hourly WD.

Figure 4 represents an example regarding the approach the WD is obtained.

4.1.2. Data transformation

Another data preparation is required before running the machine-learning algorithms. The polar coordinates of the WD (0–360°) are transformed into Cartesian coordinates, by consider-
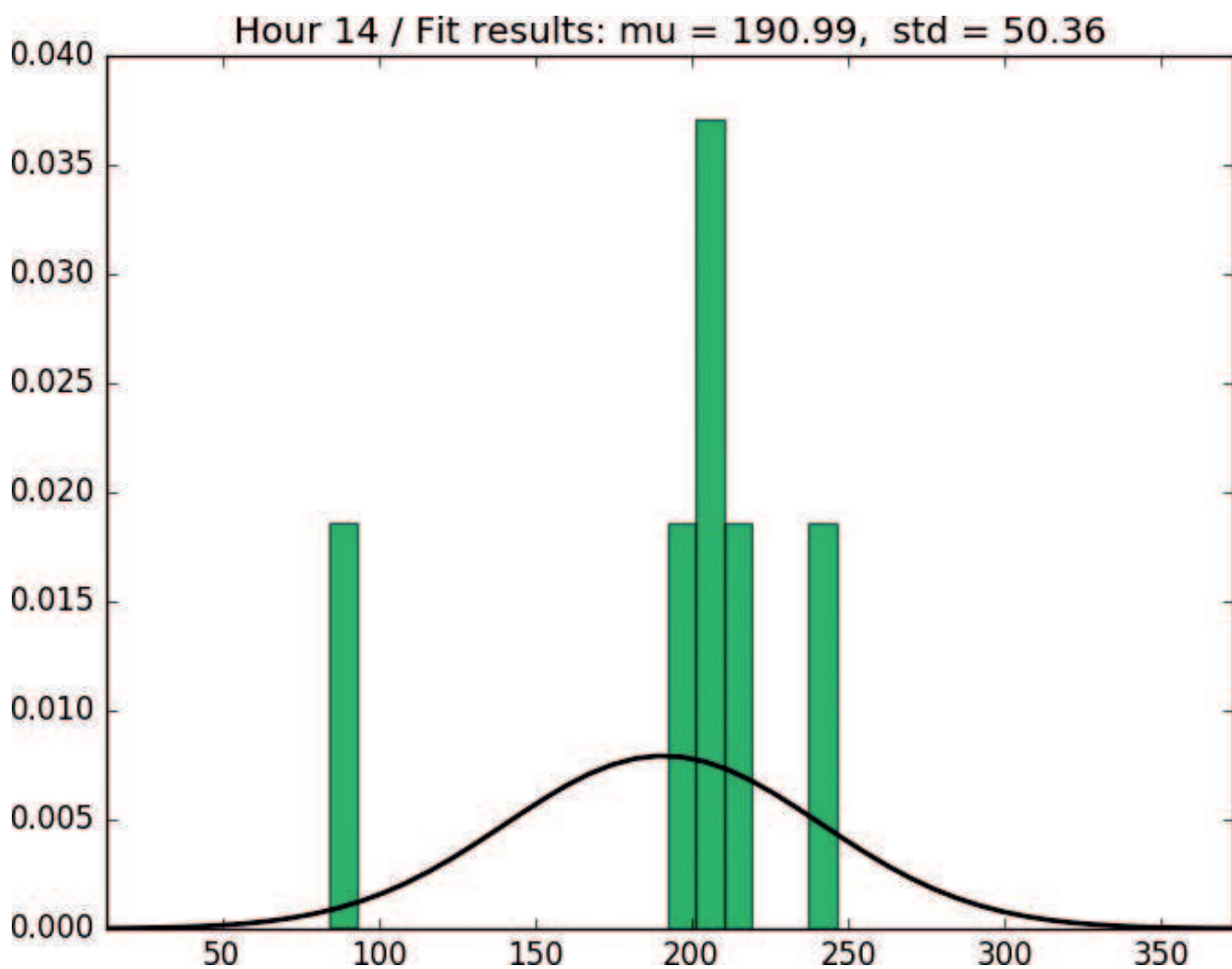

Figure 4. Representation of the calculation of the WD for a specific hour. The graphic indicates the WD angles, in degrees (x-axis), and their respective ratio of occurrences (y-axis). The black curve represents the normal distribution that fits the data. Here, the value of the hourly WD is mu ≈ 191°.

ing both WD and WS in a same formula (see Eqs. (9) and (10)). This mathematical transformation permits a more accurate feature representation of the data with respect to the WD around the north axis. Otherwise, it would be impossible to find a correlation between WD and $PM_{2.5}$, since some similar WD pointing north could have completely different values (slightly higher than 0° or slightly lower than 360°) according to the polar coordinates. This transformation is particularly relevant for machine-learning algorithms based on linear regression, because this modeling relies on a continuous relationship between parameters [9].

$$Xwind = cos\left(\frac{WD \cdot \pi}{180°}\right) \cdot WS \tag{9}$$

$$Ywind = sin\left(\frac{WD \cdot \pi}{180°}\right) \cdot WS \tag{10}$$

Thus, the final dataset is composed of 13 features, which are Xminutes, Yminutes, %orange, %red, SR, T, P, rain, RH, WS, Xwind, Ywind, and $PM_{2.5}$ (= feature to predict).

4.2. Single models

Two models are proposed. The first one is based on a multiple regression algorithm as described in Section 2.1. The second one implements a model tree that allows for a larger flexibility (but also complexity) than a linear regression for modeling the data.

4.2.1. Multiple regression model

The linear regression model obtained after running the algorithm is as follows:

$PM_{2.5}$ =

2.199	*	**Yminutes**	+
−18.0966	*	**%red**	+
39.7399	*	**%orange**	+
0.2636	*	**RH**	+
1.0088	*	**pressure**	+
0.8186	*	**temperature**	+
1.3403	*	**Xwind**	+
−753.8078			

The prediction accuracy of the model is evaluated as

r = 0.58

RMSE = 7.32

The result shows that the regression model considers all the three classes of parameters (time, traffic, and weather) to predict the value of $PM_{2.5}$. Nevertheless, in terms of meteorological
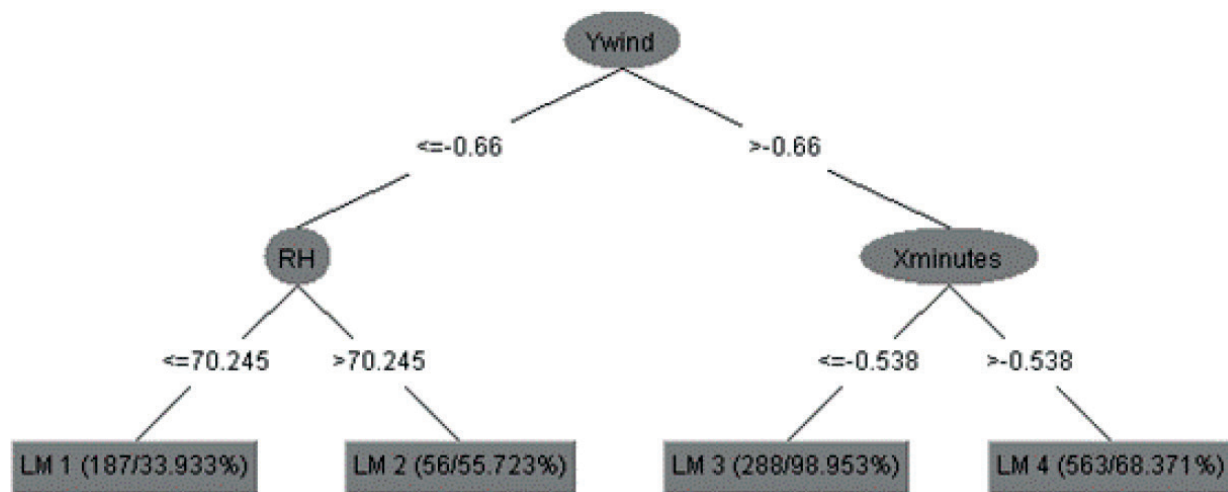
Figure 5. Graphical representation of the model tree and its respective decision rules to invoke the best regression models (LM 1–4) to predict the value of $PM_{2.5}$.

factors, the solar radiation and the precipitation are filtered by the M5 method (see Section 2.1 for more details). The rain attribute is certainly removed, since it occurs only 71 times, which represents 6.4% of the total instances. The SR is also excluded from the model, because it is an attribute mostly redundant with some of the other meteorological factors, and the filtering method is essentially based on an elimination of the redundant information. As hypothesized, including the weather conditions in the model allows for a significant improvement of the prediction accuracy. The value of correlation coefficient is almost twice higher than a model that does not consider meteorological data.

4.2.2. Regression model tree

A model tree is a more complex and flexible modeling of the data, since it is composed of several rules and each of these rules are associated with a regression model [18]. In other words, in such a tree representation, there is a different linear regression model at the leaves to predict the response of the instances that reach the leaf. In the present modeling, we use a pruned tree, in which the minimum number of instances allowed at a leaf node is nine.

Figure 5 represents the resulting model tree. It is composed of four rules as follows:

1: if Ywind ≤ −0.66 and RH ≤ 70.245

        model = LM 1

2: else if Ywind ≤ −0.66 and RH > 70.245

        model = LM 2

3: else if Ywind > −0.66 and Xminutes ≤ −0.538

        model = LM 3

4: else if Ywind > −0.66 and Xminutes > −0.538

        model = LM 4

The linear regression models associated to each rule are:

- LM 1

$PM_{2.5}$ =

3.3209	*	Xminutes	+
0.1278	*	Yminutes	+
−1.0521	*	%red	+
22.0077	*	%orange	+
0.1359	*	RH	+
0.0587	*	pressure	+
0.0101	*	SR	+
−0.3479	*	temperature	+
0.8434	*	Xwind	+
−41.7637			

- LM 2

$PM_{2.5}$ =

0.5269	*	Xminutes	+
0.1278	*	Yminutes	+
−1.0521	*	%red	+
6.595	*	%orange	+
0.3362	*	RH	+
0.0587	*	pressure	+
−0.0346	*	SR	+
1.5505	*	temperature	+
0.2383	*	Xwind	+
−72.0163			

- LM 3

$PM_{2.5}$ =

−0.0904	*	Xminutes	+
10.1893	*	Yminutes	+

−41.9183	*	%red	+
51.2883	*	%orange	+
0.3139	*	RH	+
1.6439	*	pressure	+
−0.0056	*	SR	+
1.7683	*	temperature	+
2.2056	*	WS	+
3.1792	*	Xwind	+
−0.0401	*	Ywind	+
−1233.2713			

- LM 4

$PM_{2.5}$ =

−0.0474	*	Xminutes	+
−2.2031	*	Yminutes	+
8.5034	*	%red	+
14.6847	*	%orange	+
0.2603	*	RH	+
−0.9338	*	pressure	+
−0.0001	*	SR	+
0.048	*	temperature	+
0.6414	*	WS	+
0.3914	*	Xwind	+
−1.3052	*	Ywind	+
669.5642			

The prediction accuracy of the model is evaluated as

r = 0.63

RMSE = 6.95

The root node of the tree is Ywind. It means that wind direction and wind speed are the fundamental factors to proceed with the selection of one or another regression model. Then, the second level of discrimination is based on two other important parameters, which are

relative humidity and Xminutes. The regression models that depend on the RH threshold (nine features) are slightly simpler than the models that depend on the Xminutes threshold (11 features). To note that when the tree algorithm is applied, the SR is included in the model, even though its weight is quite low. As expected, the model tree (four rules and an average of 10 features per rule) is more complex than the linear regression model (seven features). Nevertheless, the model tree is still easy to interpret and provides a prediction performance slightly better than the linear regression (+0.05 for the correlation coefficient of the tree).

4.2.3. Interpretation of the results

This analysis shows that including meteorological factors as model inputs improves the prediction accuracy of $PM_{2.5}$ concentrations (r = 0.58). The performance is slightly improved by applying a model tree, which is composed of four linear regressions (r = 0.63).

Thus, the results suggest that the use of a quite affordable meteorological station enables us to significantly improve the prediction of the concentration of fine particulate matter (The correlation coefficient is twice higher than with the traffic monitoring only.) All the meteorological factors are relevant for the prediction, except the precipitation accumulation. Rain seems to be excluded from the model, because it is a very rare event.

Next, it is studied if a multiple model approach, based on three models a day, could improve the prediction accuracy.

4.3. Multiple models

The same division of the dataset into three periods as in Section 3.3 is carried out. Since the day is analyzed into three independent parts, the dataset can be reduced to 12 features: minutes, %orange, %red, SR, T, P, rain, RH, WS, Xwind, Ywind, and $PM_{2.5}$ (= feature to predict). The three datasets are composed of 110, 116, and 145 instances for the morning, midday, and afternoon models, respectively.

4.3.1. Morning model

The linear regression model obtained after running the algorithm is as follows:

$$PM_{2.5} =$$

0.0513	*	**minutes** +	
41.7958	*	**%orange**	+
−0.23	*	**RH**	+
−2.8397	*	**temperature**	+
2.5325	*	**Xwind**	+

$$8.5432 \quad * \quad \text{Ywind} \quad +$$

$$38.6386$$

The prediction accuracy of the model is evaluated as

$$r \quad = 0.58$$

$$\text{RMSE} = 9.56$$

The model presents six features, only. It means that many attributes are filtered, especially in terms of meteorological factors (SR, pressure, rain, and WS are removed). It can be explained by the fact that the prediction of the level of $PM_{2.5}$ in the morning would be mainly correlated with the density of the traffic (see Section 3.3). However, the morning model does not seem to be significantly different than the single multiple regression neither in terms of features (five identical attributes) nor in terms of performance (r = 0.58 in both cases).

4.3.2. Midday model

The linear regression model obtained after running the algorithm is as follows:

$$PM_{2.5} \quad =$$

−0.0636	*	minutes +	
28.7942	*	%orange	+
0.4791	*	RH	+
−10.0519	*	rain	+
−0.0141	*	SR	+
2.5065	*	temperature	+
3.8358	*	Xwind	+
−2.4909			

The prediction accuracy of the model is evaluated as

$$r \quad = 0.56$$

$$\text{RMSE} = 9.13$$

The model is still composed of the same nucleus of features: minutes, %orange, RH, temperature, and wind. The only new parameter that appears as predictive feature is the precipitations. It can be explained by the fact that the rain events usually occur in Quito at midday. This factor has a negative coefficient, because the precipitation has a cleaning effect on the concentration of fine particulate matter [19]. The performance of the model is maintained at a constant accuracy (r = 0.56).

4.3.3. Afternoon model

The linear regression model obtained after running the algorithm is as follows:

$$PM_{2.5} =$$

$$-0.02 \quad * \quad minutes \quad +$$

$$28.0895 \quad * \quad \%red \quad +$$

$$0.4498 \quad * \quad RH \quad +$$

$$-2.7491 \quad * \quad pressure \quad +$$

$$-2002.1108$$

The prediction accuracy of the model is evaluated as

$$r = 0.56$$

$$RMSE = 6.61$$

This model is simpler (only four features) and does not consider exactly the same attributes than the two previous models (Pressure is used, and %red is preferred to %orange.) In Section 3.3.3, differences were already noted in the afternoon model with respect to the morning and midday. The explanation seemed to be related to the difficulty to get a reliable predictive model of $PM_{2.5}$ when the particulates are strongly diluted in the atmosphere. In such a situation, the fair performance of the model (r = 0.56; RMSE = 6.61) would be more caused by the reduced fluctuation of the $PM_{2.5}$ values (**Figure 3** shows a maximum peak at around 20 µg/m³, against 30 µg/m³ in the morning) than the reliability of the prediction per se.

4.3.4. Interpretation of the results

Eq. (11) presents the average prediction accuracy by modeling the air pollution through the three daily models.

$$\bar{r} = \frac{0.58 + 0.56 + 0.56}{3} = 0.57 \tag{11}$$

Although the morning model is slightly more accurate than the two other ones, the mean value of the regression coefficient is not better than the regression coefficient of the single model, especially if this model is obtained by a model tree algorithm.

Thus, when meteorological factors are taken into account, it does not seem to be advantageous to consider three regression models per day. It can be explained by the fact that the weather conditions have a very strong effect on the levels of $PM_{2.5}$ (e.g., rain and wind tend to clean the atmosphere). Thus, including these factors as model features reduces the mere influence of the traffic on the value of $PM_{2.5}$. And since the impact of this human activity is more

significant in the morning than in the rest of the day, because of the low dilution of the vehicle emissions in the atmosphere, adding meteorological parameters in the model decreases the performance differences between the three daily models.

5. Adding trace gas concentrations

This part intends to verify the prediction accuracy of the methods as described in the previous sections. To do so, the precision of the prediction based on low-cost data collection is compared with a pollution monitoring that makes use of costlier technologies (i.e., EPA-approved chemical sensors). Then, a hybrid model is proposed from a selection of the most relevant features to minimize the prediction error.

5.1. Prediction from chemical monitoring

The concentrations of $PM_{2.5}$ are commonly correlated with other air pollutants, such as SO_2, NO_2, CO, etc. [20]. However, the monitoring of these substances involves a more specialized equipment than traffic or weather monitoring. The performance of the models built in this section is used as referential to assess the quality of the previous models and investigates if a selection of the most affordable chemical records can significantly improve the overall prediction accuracy. Four additional criteria pollutants were measured (CO, NO_2, SO_2, and O_3). For SO_2 concentrations, ThermoFisher Scientific 43i high-level SO_2 analyzer was used based on ultraviolet florescence (EPA No. EQSA-0486-060). For O_3 concentration data collection, ThermoFisher Scientific 49i ozone analyzer was used based on ultraviolet absorption (EPA No. EQOA-0880-047). For NOx concentration data collection, ThermoFisher Scientific 42i NOx analyzer was used based on chemiluminescence method (EPA No. RFNA-1289-074). Finally, for CO concentration data collection, ThermoFisher Scientific 48i was used based on infrared absorption (EPA No. RFCA-0981-054). The used dataset is composed of 1118 observations and 5 features: CO, NO_2, O_3, SO_2, and $PM_{2.5}$ (= feature to predict).

The prediction accuracy of the model is evaluated as

$$r = 0.75$$

$$RMSE = 5.89$$

The evaluation of this model demonstrates that only the chemical factors are very high predictors of the level of fine particulate matter. A model built with these parameters provides a significantly lower RMSE and higher r than the traffic and meteorology based models. This outcome was expected as the levels of anthropogenic $PM_{2.5}$ that are directly related to the emission of other air pollutants, such as a number of different contaminants that come from the same sources. It can be concluded from this analysis that selecting some low-cost chemical recordings should improve the prediction accuracy of the affordable models.

5.2. Prediction from full data sources

This section explores the possibility to get a better prediction of air pollution if we build a hybrid model that uses a combination of the whole data sources mentioned previously. The objective is to define the best predictive model to estimate the concentration of $PM_{2.5}$ from all the available types of data.

5.2.1. Single model

The full dataset is used for this analysis. There is a total number of 17 features, which are Xminutes, Yminutes, %red, %orange, relative humidity, precipitation, pressure, solar radiation, temperature, wind Speed, Xwind, Ywind, CO, NO_2, O_3, SO_2, $PM_{2.5}$ (= feature to predict).

The linear regression model obtained after running the algorithm is as follows:

$$
\begin{aligned}
PM_{2.5} = \quad & \\
1.4412 \quad * \quad & Yminutes \quad + \\
0.2212 \quad * \quad & RH \quad + \\
-0.0035 \quad * \quad & SR \quad + \\
0.9367 \quad * \quad & temperature \quad + \\
1.2377 \quad * \quad & WS \quad + \\
0.7501 \quad * \quad & Xwind \quad + \\
0.3971 \quad * \quad & Ywind \quad + \\
0.2691 \quad * \quad & NO_2 \quad + \\
0.1878 \quad * \quad & O_3 \quad + \\
1.0463 \quad * \quad & SO_2 \quad + \\
8.3473 \quad * \quad & CO \quad + \\
-30.8553 \quad &
\end{aligned}
$$

The prediction accuracy of the model is evaluated as

$$r = 0.81$$

$$RMSE = 5.31$$

The results show that the regressive model considers three classes of parameters (time, meteorology, and criteria pollutants) out of four to predict the value of $PM_{2.5}$. Traffic information is filtered, certainly because of its redundancy with time. After attribute selection (M5 method), the final model is composed of 11 features out of 16. As hypothesized, a model based on a hybrid data source allows for a significant improvement of the prediction

accuracy. The values of the correlation coefficient and the RMSE are better for the hybrid than the chemical model.

5.2.2. Multiple models

5.2.2.1. Morning model

The linear regression model obtained after running the algorithm is as follows:

$PM_{2.5}$	=		
0.0379	*	minutes	+
0.3438	*	RH	+
−1.7248	*	pressure	+
−0.6846	*	temperature	+
4.5902	*	CO	+
0.4294	*	NO_2	+
2.0133	*	SO_2	+
0.6343	*	O_3	+
1209.4494			

The prediction accuracy of the model is evaluated as

r   = 0.85

RMSE = 6.04

5.2.2.2. Midday model

The linear regression model obtained after running the algorithm is as follows:

$PM_{2.5}$	=		
−0.0362	*	minutes	+
−1.1911	*	pressure	+
−0.0122	*	SR	+
2.3857	*	temperature	+
1.4346	*	Ywind	+
0.2274	*	RH	+
14.8788	*	CO	+

0.3632	*	NO_2	+
0.796	*	SO_2	+
0.2348	*	O_3	+
835.1936			

The prediction accuracy of the model is evaluated as

r	= 0.87
RMSE	= 5.33

5.2.2.3. Afternoon model

The linear regression model obtained after running the algorithm is as follows:

$PM_{2.5}$ =

21.026	*	%red	+
−14.9417	*	%orange	+
0.3291	*	RH	+
0.8285	*	temperature	+
1.2914	*	WS	+
−1.1325	*	pressure	+
−0.0109	*	SR	+
0.3909	*	NO_2	+
0.6993	*	SO_2	+
0.2503	*	O_3	+
790.3383			

The prediction accuracy of the model is evaluated as

r	= 0.66
RMSE	= 6.29

5.2.2.4. Interpretation of the results

The results of the Eq. (12) shows that the average prediction accuracy (evaluated by the regression coefficient metrics) by modeling the air pollution through three models is

$$\bar{r} \; = \; \frac{0.85 + 0.87 + 0.66}{3} \; = \; 0.79 \tag{12}$$

Thus, it seems that using several models with all the available features for the prediction of fine particulate matter is only justified to predict the level of $PM_{2.5}$ from 6 am to 2 pm ($r \approx 0.86$). After this period, the model gets more complex and less reliable. This result confirms the previous analyses that tend to demonstrate that the model accuracy to estimate $PM_{2.5}$ concentrations from traffic, meteorology, and air pollutants is stronger when the gases and particulates are less diluted in the atmosphere.

6. Simplification and recommendations

6.1. The simplest best model

Since the full feature model (Section 5.2) is quite complex, the present stage consists of removing insignificant and/or redundant features in order to optimize the modeling. The goal is to find a simple model that is still able to provide a reliable estimation of $PM_{2.5}$ concentrations. The simplest best model is defined as a model that maintains a high accuracy ($r \geq 0.8$) with a maximum number of features equal to eight. The method used to get this model is the ranker search method. This technique sorts the attributes according to their evaluation and allows for a specification of the number of attributes to retain.

The linear regression model obtained after running the algorithm is as follows:

$$
\begin{aligned}
PM_{2.5} =\ & 0.2032 * RH + \\
& 0.6507 * temperature + \\
& -0.0021 * SR + \\
& 0.4549 * Xwind + \\
& 0.225 * NO_2 + \\
& 0.2159 * O_3 + \\
& 1.0707 * SO_2 + \\
& 8.8163 * CO + \\
& -23.9476
\end{aligned}
$$

The prediction accuracy of the model is evaluated as

$$r = 0.8$$

$$RMSE = 5.34$$

Table 1 represents the ranked attributes, in which the features are sorted in the descending order of their individual performance to predict the output value.

Ranking	Performance	Feature
1	0.0311	SO_2
2	0.0256	CO
3	0.0193	Relative humidity
4	0.0172	NO_2
5	0.0133	O_3
6	0.0125	Solar radiation
7	0.0109	Xwind
8	0.0065	Temperature

Table 1. Ranked attributes.

The simplest best model is composed of the whole chemical parameters and a selection of meteorological factors (RH, SR, Xwind, and T). As suggested by the previous analyses, the individual performance to accurately estimate the values of $PM_{2.5}$ is globally higher for the chemical (first, second, fourth, and fifth positions) than the meteorological features (third, sixth, seventh, and eighth positions). In other words, $PM_{2.5}$ are firstly correlated with the emission of chemical substances (especially SO_2 and CO) and secondly with the weather conditions (especially relative humidity and solar radiation). It is to note the negative correlation between the value of SR and the concentration of $PM_{2.5}$. This result can be explained by the fact that the larger is the SR, the deeper is PBL, and consequently, the bigger is the dilution of fine particulate matter in the boundary layer. The other factors are positively correlated with $PM_{2.5}$. Besides its simplicity (eight features only), the model is able to predict the level of fine particulate matter with the same accuracy than a model using all the features (r = 0.8 and RMSE = 5.3, in both cases).

6.2. Recommendations based on model performances

The final objective of this study is to find the best predictive model that uses the less costly data recording of relevant features. As previously mentioned, the accurate measurement of trace gases requires expensive equipment. Thus, the best affordable model can be defined as the model that gets the best performance with no more than two trace gases. The model performances with the whole affordable attributes and only one or two trace gases are presented in **Table 2**. The model accuracy is assessed according to the value of r. The main diagonal represents the performance by considering a single trace gas, whereas the other cells take into account two gases.

The results show that it is still possible to build a model with high prediction accuracy with two trace gases, only. The best performance is obtained by considering SO_2 and NO_2 (r = 0.78). It can be explained by the fact that these two trace gases are strongly correlated with the values of $PM_{2.5}$ (see **Table 1**). In the case that only one trace gas sensor is affordable, it has to be a device that measures the levels of CO or NO_2 (r = 0.73). It is to note that O_3 is a gas that can

	SO_2	CO	NO_2	O_3
SO_2	0.7			
CO	0.77	0.73		
NO_2	0.78	0.76	0.73	
O_3	0.7	0.75	0.73	0.58

Table 2. Model performance (r value) with all the affordable attributes (e.g., time, traffic, and meteorology) and only one (main diagonal) or two (other cells) trace gases.

be automatically discarded, since its power of prediction is the lowest (Section 4 shows that models without O_3 get a better r). This finding could be expected as there is no direct relationship between the level of O_3 (a secondary pollutant) and the concentrations of $PM_{2.5}$.

7. Conclusions and perspectives

This study demonstrates that the $PM_{2.5}$ prediction performance depends on the available input information. The first finding shows that it is possible to get a reasonable prediction of $PM_{2.5}$ concentrations only using public access traffic data. Ambient $PM_{2.5}$ pollution prediction based on traffic can be significantly improved by using three models a day instead of a single one, especially for the morning hours. During the morning rush hour, planetary boundary layer is shallow, resulting in a continuous traffic emission buildup showing a cumulative growth of $PM_{2.5}$ concentrations. The latter start decreasing with the dilution effect of the PBL deepening, due to surface heating, increase in temperatures and ventilating wind effect. Thus, using an affordable meteorological station data further improves the prediction accuracy. In this case, a regression model tree gives a better prediction than a linear regression model. As expected, the best model is obtained by including a hybrid data sources as features (time, traffic, meteorological, and the concentrations of atmospheric criteria pollutants). The complexity of the resulting model can be reduced from seventeen to eight most relevant features without reducing the performance (r ≈ 0.8, and RMSE ≈ 5.3). These eight selected attributes are composed of criteria pollutants (CO, NO_2, O_3, SO_2) and meteorological factors (humidity, solar radiation, temperature, wind speed, and direction). Thus, our results suggest to proceed with a selection of chemical sensors based on the best ratio prediction/cost. For example, if only one trace gas sensor is affordable, the best performance can be reached with CO or NO_2 concentrations, while the use of two trace gases (SO_2 and NO_2) are sufficient to get very close to the best possible accuracy. In contrast, O_3 is a secondary pollutant that can be excluded from the models with no significant consequences on the prediction of $PM_{2.5}$, suggesting a low impact of photochemical component in $PM_{2.5}$ formation.

The proposed approach is easily generalizable to other cities worldwide. A storage and regression analysis of 2-month data were sufficient to build models that are able to predict fine particulate matter with high accuracy. The main limitation of the present method is to

predict PM$_{2.5}$ when the PBL is deep. Nevertheless, it is often less of an issue in terms of air quality since an elevated PBL enhances dilution and, consequently, reduces the concentration of atmospheric contaminants. Further work will focus on improving the model performance at evening rush hours. More refined models are expected to be obtained by including additional observations and features into the dataset. For example, some additional studies are anticipated to investigate the impact of PBL depth on the dilution of the PM$_{2.5}$ pollution.

Furthermore, it is motivating to investigate the current model performance with the data acquired by the lower tier equipment. In this study, the air pollution and meteorology were measured with USEPA-approved equipment, not affordable to a large fraction of cities in the developing countries, thus limiting air pollution studies and awareness to the main cities. It has been shown, however, that small cities are often more polluted than the big agglomerations, presenting the necessity for a wide set of options to promote the consciousness of the air quality [21].

Author details

Yves Rybarczyk[1,2]* and Rasa Zalakeviciute[1]

*Address all correspondence to: y.rybarczyk@fct.unl.pt

1 Intelligent & Interactive Systems Lab (SI² Lab), Universidad de Las Américas, Quito, Ecuador

2 Department of Electrical Engineering – CTS/UNINOVA, Nova University of Lisbon, Monte de Caparica, Portugal

References

[1] United Nations, Department of Economic and Social Affairs, Population Division. World Population Prospects: The 2015 Revision, Key Findings and Advance Tables. 2015. Working Paper No. ESA/P/WP.241. Retrieved from: https://esa.un.org/unpd/wpp/publications/files/key_findings_wpp_2015.pdf

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

[3] UNEP. Status of Fuel Quality and Vehicle Emission Standards Latin America [Internet]. 2016. Available from: http://www.unep.org/urban_environment/Issues/urban_air. [Accessed: 22 October 2016]

[4] Lelieveld J, Evans JS, Fnais M, Giannadaki D, Pozzer A. The contribution of outdoor air pollution sources to premature mortality on a global scale. Nature. 2015;**525**(7569): 367-371

[5]   Karagulian F, Belis CA, Dora CFC, Prüss-Ustün AM, Bonjour S, Adair-Rohani H, Amann M. Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level. Atmospheric Environment. 2015;**120**:475-483

[6]   Castell N, Dauge FR, Schneider P, Vogt M, Lerner U, Fishbain B, Broday D, Bartonova A. Can commercial low-cost sensor platforms contribute to air quality monitoring and exposure estimates? Environment International. 2017;**99**:293-302

[7]   Borrego C, Costa AM, Ginja J, Amorim M, Coutinho M, Karatzas K, Sioumis T, Katsifarakis N, Konstantinidis K, De Vito S, Esposito E, Smith P, André N, Gérard P, Francis LA, Castell N, Schneider P, Viana M, Minguillón MC, Reimringer W, Otjes RP, von Sicard O, Pohle R, Elen B, Suriano D, Pfister V, Prato M, Dipinto S, Penza M Assessment of air quality microsensors versus reference methods: The EuNetAir joint exercise. Atmospheric Environment. 2016;**147**:246-263.

[8]   USEPA. Evaluation of emerging air pollution sensor performance. 2017. Available from: https://www.epa.gov/air-sensor-toolbox/evaluation-emerging-air-pollution-sensor-per-formance. [Accessed: 28 August 2017]

[9]   Kleine Deters J, Zalakeviciute R, Gonzalez M, Rybarczyk Y. Modeling PM2.5 urban-pollution using machine learning and selected meteorological parameters. Journal of Electrical and Computer Engineering. 2017. 14 pages. Article ID 5106045. DOI: 10.1155/2017/5106045

[10] Brokamp C, Jandarov R, Rao MB, LeMasters G, Ryan P. Exposure assessment models for elemental components of particulate matter in an urban environment: A comparison of regression and random forest approaches. Atmospheric Environment. 2017;**151**:1-11

[11] Quinlan RJ. Learning with continuous classes. In: Proceedings of the 5th Australian Joint Conference on Artificial Intelligence. Singapore; 1992. pp.343-348. Retrieved from: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.34.885&rep=rep1&type=pdf

[12] Rybarczyk Y. 3D markerless motion capture: A low-cost approach. In: Rocha A, Correia AM, Adeli H, Reis LP, Teixeira MM, editors. New Advanced in Information Systems and Technologies; Recife, Brazil. Switzerland: Springer; 2016. pp. 731-738. DOI: 10.1007/978-3-319-31232-3

[13] Mierswa I, Wurst M, Klinkenberg R, Scholz M, T. Yale E. Rapid prototyping for complex data mining tasks. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. Philadelphia: USA; 2006; pp. 935-940. Retrieved from: https://pdfs.semanticscholar.org/5722/e63d03edba571262ba258fe5aaf-fed4147c9.pdf

[14] Stull RB. An Introduction to Boundary Layer Meteorology. Boston, Massachusetts: 13Kluwer Academic Publishers; 1988.

[15] Cazorla M. Air quality over a populated Andean region: Insights from measurements of ozone, NO, and boundary layer depths. Atmospheric Pollution Research. 2016;**7**:66-74

[16]   Hsieh CI, Katul G, Chi TW. An approximate analytical model for footprint estimation of scalar fluxes in thermally stratified atmospheric flows. Advances in Water Resources. 2000;**23**:765-772

[17]   Rybarczyk Y, Zalakeviciute R. Machine learning approach to forecasting urban pollution: A case study of Quito, Ecuador. In: Ecuador Technical Chapters Meeting (ETCM). Guayaquil, Ecuador: IEEE; 12-14 Oct. 2016, 2016. DOI: 10.1109/ETCM.2016.7750810

[18]   Wang Y, Witten I H. Induction of model trees for predicting continuous classes. In: van Someren M, Widmer G, editors. Proceedings of the 9th European Conference on Machine Learning; April 1997. Prague, Czech Republic. Springer; 1997

[19]   Li Y, Chen Q, Zhao H, Wang L, Tao R. Variations in PM10, PM2.5 and PM1.0 in an urban area of the Sichuan basin and their relation to meteorological factors. Atmosphere. 2015;**6**:150-163

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

[21]   Zalakeviciute R, Rybarczyk Y, López-Villada J, Diaz Suarez M. Quantifying decade-long effects of fuel and traffic regulations on urban ambient PM 2.5 pollution in a mid-size south American city. Atmospheric Pollution Research. 2017. DOI: 10.1016/j.apr.2017.07.001