Support Vector Regression for PM10 Concentration Modeling in Santa Marta Urban Area

Sanchez-Torres G., and Diaz Bolaño I.

Abstract—This paper presents a method for particulate material PM10 modeling based on support vector regression (SVR). Specifically, we applied ε-support vector regression (ε-SVR) and v-support vector regression (v-SVR) to a set of data recorded in the city of Santa Marta, Colombia, between 1999 and 2016. The set of data was initially pre-processed, filtered and normalized, and then was used to fit the SVR models. The parametrization and accuracy of each regression model are reported here. We used a month as the unit of time for the models and analyzed the accuracy for one-step predictions. The final results of this work show the best parameters and prediction properties of the SVR models for pollution data modeling in Santa Marta.

Index Terms—air quality, PM10, machine learning, support vector regression

I. INTRODUCTION

Air pollution in urban areas has become a relevant phenomenon for the scientific community. The study of this phenomenon is necessary to enact regulatory policies in order to reduce the impact of air pollution on human life and the environment [1].

A vast amount of documentation links the presence of air pollutants to several public health problems. Specifically, pollution is linked to a higher risk of respiratory, cardiovascular, and nervous-system diseases, general disabilities, and cognitive ability reduction [2]–[6]. Consequently, deeper understanding of the effect of air pollution on human health and the technological mechanisms for monitoring it are of great interest. Advancing in this field will allow governmental entities to propose optimal regulatory strategies to avoid health risks without affecting economic growth.

According to the literature, primary air contaminants can be classified as conventional or non-conventional. Typically, conventional pollutants include carbon monoxide (CO), nitrogen dioxide (NO2), ozone (O3), sulphur oxide (SO2) and particulate matter PM10 [7]–[9]. On the other hand, non-conventional pollutants include: Benzene (C6H6), lead (Pb) and its composites, cadmium (Cd), mercury (Hg), hydrogen Sulphur (H2S), etc. [10]. It is important to note that there are more studies about the impact of conventional contaminants on human health than that of non-conventional contaminants.

Particulate matter (PM) is often made up of small airborne particles with different diameters. These are classified as thick, fine, and ultra-fine particles with aerodynamic diameters of 2.5 to 10μm (PM10), less than 2.5μm (PM2.5), and smaller than 0.1μm (PM0.1) [11], respectively. These particles are mainly composed of dust and smoke particles emitted by wood burning, diesel vehicles, and industrial operations. Although, some studies indicate a correlation between health conditions and air pollution levels, the identification of toxic PM components is a complex task. The problem arises because PM is composed of a complex mixture of solid and liquid particles with significant variations in mass, size, shape, volume, chemical nature, acidity, solubility, and origin [12].

The World Health Organization (WHO) has published the guidelines for air quality in Europe from 1987 to the present [13]; and recently, also for the rest of the world [14]. In Colombia, air-quality norms have been established over the past decade [10], [15]. These regulations state the maximum permissible values of contaminants for annual exposure. Table 1 lists some of these contaminants, the maximum permissible levels in Colombia, and the maximum levels recommended by WHO. It also contains a general measure defined as Total Suspended Particles (PST, from the Spanish Partículas Suspendidas Totales).

In Colombia, the Ministerio de Ambiente (the Ministry of the Environment) elaborated a binding protocol for the monitoring and supervision of air quality. The protocol defines the target air quality for each year depending on a constant measurement process carried out by the Air Quality Vigilance System (SVCA, from the spanish Sistema de Vigilancia de Calidad del Aire). The design, location, and some other factors of the measurement process are also decided by the protocol [16]. Environmental authorities in Colombia have monitored 170 stations since 2010. Of these stations, 47%, are manually operated, 35% are automatically operated, and 18% are semi-automatically operated. PM10 pollutants have been monitored in 85% of these stations, while SO2 and NO2 contaminants have been monitored by 35%, O3 by 25%, PST by 24%, CO by 23%, and PM2.5 by 15% of these...
stations [17].

According to a Colombian case-study by the World Bank [18], people are mostly concerned with contaminants due to their effect on the population with a high risk of illness and mortality, especially for children under five and the elderly. Urban air contamination caused three times more deaths than inadequate water supply in 2002, and five times more deaths than indoor air contamination. The study also reveals that the analyses of the impact caused by $PM_{10}$ cost almost 0.8% of the Gross Domestic Product (GDP) in 2002, and 1.12% in 2010.

Specifically, in the Magdalena department, of which Santa Marta is the capital, PST measurement from 2007-2010 showed a progressive reduction from values of 184 $\mu g/m^3$ mainly in areas close to coal shipping harbors. Some stations registered values of 88 $\mu g/m^3$, which were under the threshold. However, this value was very close to the maximum permissible value (100 $\mu g/m^3$) [17].

Several studies have attempted to predict air pollution level through computational models [14]. Furthermore, it is of interest to include different atmospheric or social variables in these models to establish correlations.

The techniques used to forecast include multiple linear regression [15]-[18], data mining [19], wavelet analysis [20], hidden Markov models [21], [22], artificial neural networks [23]-[24], fuzzy logic, neuro-fuzzy logic, and stochastic simulation [25], among others [26].

In this study, we describe the fitting of prediction models for contaminant concentration applied to a dataset registered from 12 monitoring stations in Santa Marta, Colombia. The results exposed here were obtained by regression, using vector support machines, and were compared with previously reported work. The goal was to determine an accurate mathematical model to predict pollution concentration in our city. Our main contributions are the analysis of the data and the results of the regression models.

The document is organized as follows: Section II describes the employed methodology. Section III contains the results obtained from the computational models. Finally, Section IV presents some conclusions from the analysis of the results.

### II. MATERIALS AND METHODS

Constructing computational prediction models entails three main phases: data selection and preprocessing, model parametrization, and accuracy evaluation. In this study, we followed the general scheme shown in Fig. 1, which includes these three phases.

#### A. Data description

In Santa Marta, the public entity in charge of the SVCA, according to the law, is the Corporación Autónoma Regional (Corpamag). The purpose of this entity is to manage environmental issues and all issues related to renewable natural resources. Corpamag operates twelve air monitoring stations placed along the coastal area, within the municipal limits of the cities of Santa Marta and Ciénaga [19], as shown in Fig 2.

![Fig. 1. Block diagram of the working methodology.](image)

From these stations, only seven still contain fully functioning devices. The other stations are difficult to access, or have electrical supply issues. Eight measurement devices are available in the seven stations, the locations of which can be seen in white in Fig. 2. The other stations left can be seen in gray. Four out of the eight available stations measure PST, including $PM_{10}$.

The monitoring devices are mostly manually operated, and are able to manage high volumes of PST with volumetric flow controllers [20]. Table 2 shows the name of the twelve stations, the contaminants measured by each one, the operational state, and the year in which the station started to register data [17], [19]. Table 3 shows the geographical localization of the stations.

#### B. Scale transformation

In order to have a standard time unit for all the stations, we transformed the registered data into a monthly time series by using a geometric mean estimation. For every subset of $i$ samples taken during month $j$, the monthly $\overline{PM_{10}^j}$ value is estimated by Equation 1.

$$\overline{PM_{10}^j} = \sqrt[n]{\prod_{i=1}^{n} PM_{10}^{j,i}}$$

where $n$ is the number of valid samples taken in month $j$.

In this study we worked with the data obtained from the seven active stations in Table 2. This data is freely available

---

**TABLE 1.**

<table>
<thead>
<tr>
<th>Contaminant</th>
<th>Maximum value $\mu g/m^3$</th>
<th>Permissible level</th>
<th>Average exposure time</th>
</tr>
</thead>
<tbody>
<tr>
<td>PST</td>
<td>100</td>
<td>150</td>
<td>Annual</td>
</tr>
<tr>
<td>$PM_{10}$</td>
<td>50</td>
<td>20</td>
<td>Annual</td>
</tr>
<tr>
<td>$PM_{2.5}$</td>
<td>25</td>
<td>10</td>
<td>Annual</td>
</tr>
<tr>
<td>SO$_2$</td>
<td>250</td>
<td>20</td>
<td>24 hours</td>
</tr>
<tr>
<td>NO$_2$</td>
<td>100</td>
<td>40</td>
<td>Annual</td>
</tr>
<tr>
<td>$O_3$</td>
<td>80</td>
<td>100</td>
<td>8 hours</td>
</tr>
<tr>
<td>CO</td>
<td>10000</td>
<td>10000</td>
<td>8 hours</td>
</tr>
</tbody>
</table>

(Advance online publication: 12 August 2019)
in Corpamag’s web system. The original samples were registered in a time interval of three days [21]. Usually, time series must be transformed and smoothed using different kinds of filters. Smoothing the series improves data interpretation and generates more realistic and accurate results. In this work we employed the Savitzky-Golay filter [24], which fits low-degree polynomials to a set of sequential data using linear least squares and convolution [25]. This filter is based on a $p$-grade polynomial regression with at least $2n+1$ equidistant points $(T_0, \ldots, T_n)$ [26], and is defined as follows[24]:

$$
T_j = \frac{1}{n} \sum_{i=m}^{j} C_i T_{j+1}
$$

where $T_j$ is the result after filtering. The index $j$ corresponds to the current datum being filtered, $T_{j+1}$ corresponds to the original values in the time series, $C_i$ is the coefficient for the $i$th value of the filter, $n$ is the number of convoluting integers equal to $2m+1$ (see Fig. 3).

### Table 2.
**MOONITORING STATIONS AND MEASURED VARIABLES.**

<table>
<thead>
<tr>
<th>Id</th>
<th>Name</th>
<th>Contaminant</th>
<th>State</th>
<th>Starting Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Invemar</td>
<td>PST</td>
<td>Active</td>
<td>1999</td>
</tr>
<tr>
<td>2</td>
<td>C. Santa Marta</td>
<td>PM10</td>
<td>Active</td>
<td>2007</td>
</tr>
<tr>
<td>3</td>
<td>C. Ejecutivo</td>
<td>PST</td>
<td>Active</td>
<td>1999</td>
</tr>
<tr>
<td>4</td>
<td>Cajamar</td>
<td>PST</td>
<td>Inactive</td>
<td>2003</td>
</tr>
<tr>
<td>5</td>
<td>Batalión</td>
<td>PST</td>
<td>Active</td>
<td>1999</td>
</tr>
<tr>
<td>6</td>
<td>Molinos</td>
<td>PM10</td>
<td>Active</td>
<td>2011</td>
</tr>
<tr>
<td>7</td>
<td>Zuana</td>
<td>PM10</td>
<td>Inactive</td>
<td>2007</td>
</tr>
<tr>
<td>8</td>
<td>Aeropuerto</td>
<td>PST</td>
<td>Inactive</td>
<td>1999</td>
</tr>
<tr>
<td>9</td>
<td>Don Jaca</td>
<td>PM10, y PST</td>
<td>Active</td>
<td>1999</td>
</tr>
<tr>
<td>10</td>
<td>Alcatraces</td>
<td>PM10, y PST</td>
<td>Active</td>
<td>1999</td>
</tr>
<tr>
<td>11</td>
<td>Papare</td>
<td>PST</td>
<td>Inactive</td>
<td>2005</td>
</tr>
<tr>
<td>12</td>
<td>Costa Verde</td>
<td>PM10, y PST</td>
<td>Active</td>
<td>2008</td>
</tr>
</tbody>
</table>

### Table 3.
**GEOGRAPHICAL LOCALIZATION OF ACTIVE STATIONS WITHIN THE LIMITS OF THE SANTA MARTA MUNICIPALITY.**

<table>
<thead>
<tr>
<th>Id</th>
<th>Latitude</th>
<th>Longitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11°15’02.8940”</td>
<td>11°15’02.8940”</td>
</tr>
<tr>
<td>2</td>
<td>11°14’25.6063”</td>
<td>11°14’25.6063”</td>
</tr>
<tr>
<td>3</td>
<td>11°14’23.3610”</td>
<td>11°14’23.3610”</td>
</tr>
<tr>
<td>4</td>
<td>11°13’57.2185”</td>
<td>11°13’57.2185”</td>
</tr>
<tr>
<td>5</td>
<td>11°11’40.5247”</td>
<td>11°11’40.5247”</td>
</tr>
<tr>
<td>6</td>
<td>11°05’54.5046”</td>
<td>11°05’54.5046”</td>
</tr>
</tbody>
</table>

C. Regression with support vector machines – SVR

The purpose of SVR is to fit a multivariate regression function $f(x)$ over a set of $N$ observations $X \in \mathbb{R}^d$. The fitting procedure transforms the observation set from a $d$-dimensional space to a $m$-dimensional space, such that $m>n$. The transformation is performed by a function or kernel $\Phi: \mathbb{R}^d \rightarrow \mathbb{R}^m$. Afterward, the procedure continues applying multiple linear regression methods in the new feature space. In this study, we use support vectors $v$ (v-SVR) and epsilon regulated function $\varepsilon$ (c-SVR) [27]–[29]. Let $X=[(x_1,y_1), \ldots, (x_n,y_n)]$ be the observation set, where each element $x_i \in \mathbb{R}^d$ represents an input, and $y_i \in \mathbb{R}^l$ represents an output value. The optimization problem of a v-SVR regression can be expressed with the following restrictions:

$$
\min \frac{1}{2} w^T W + C(v) = \frac{1}{2} \sum_{i=1}^{n} \xi_i + \xi^*_i
$$

$$
\begin{cases}
(w^T \Phi(x_i) + b) - y_i \leq \varepsilon + \xi_i \\
(y_i - (w^T \Phi(x_i) + b)) \leq \varepsilon + \xi^*_i
\end{cases}
$$

where $0 \leq \theta \leq 1$, $C$ is the regularization parameter, and $\Phi(x_i)$ is the space transformation function, or kernel. The $\varepsilon$ value is the cost function, or loss, and represents an error tolerance. Thus, if $w^T \Phi(x_i)$ is in the $y_i \pm \varepsilon$ range, it is not considered to be an error. As described by [27] and [29], a proper estimation of $\varepsilon$ is difficult. The reason is that the
procedure introduces a new parameter $v$ to control the number of support vectors and training errors. Equations 3 and 4 can be solved by introducing Lagrange multipliers $\alpha^*, \eta^*$, $\beta \geq 0$. In this way, we obtain a dual Lagrange formulation for $v \geq 0$, $C > 0$:

$$\max \sum_{i=1}^{n} (\alpha_i^* - \alpha_i) y_i - \frac{1}{2} \sum_{i,j=1}^{n} (\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)k(x_i, x_j)$$

(5)

Subject to:

$$\begin{cases} \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) = 0 \\ 0 \leq \alpha_i^* \leq C \\ \sum_{i=1}^{n} (\alpha_i - \alpha_i^*) \leq C, v \end{cases}$$

(6)

Therefore, the regression function is approximated as follows:

$$y = f(x) = \sum_{i=1}^{n} (\alpha_i^* - \alpha_i)k(x_i, x) + b$$

(7)

where $b$ represents a systematic error or noise, and $k$ is the dot product in the feature space yielded by the transformation function $\Phi$ or $\text{kernel}$:

$$k(x, y) = (\Phi(x), \Phi(y))$$

(8)

D. Parameter selection for the v-SVR model

Three parameters must be established to build the SVR regression model: $C, \varepsilon$, and $\delta$. $C$ is the penalty parameter, $\varepsilon$ is the tolerance parameter, which measures the degree of fitting of the regression model to the training data. $\delta$ is a parameter related to the selected $\text{kernel}$ function. Model performance and accuracy depend on parameter selection. There is no widely accepted procedure to determine hyper-parameters for SVR models [30]–[32]. However, multiple approaches have been proposed to address the parameter selection issue [30], including exhaustive search, analytical techniques, and metaheuristic techniques, such as genetic algorithms and particle swarm optimization.

Analytical parameter selection technique

This technique uses standard SVM parametrization which employs the training data statistics and noise variance to set up the model’s parameters. The following equations compute parameters $C$ and $\varepsilon$:

$$C = \max \left( \bar{y} + 3\sigma, \left| \bar{y} - 3\sigma \right| \right)$$

(9)

$$\varepsilon = \tau \sigma_{noise} \sqrt{\frac{\ln(n)}{n}}$$

Where $\bar{y}$ is the mean value of the training data corresponding to the outputs $y_i \in R^1$, and $\sigma_i$ is the standard deviation of values $y$ in the training data. The $\sigma_{noise}$ parameter is the standard deviation of the noise estimation for the training set, and $\tau$ is a constant, experimentally set to 3 in [32]. In addition, kernels such as the Radial Basis Function (RBF) and the sigmoid function require an additional parameter $\delta$ that can be approximated by:

$$\delta \sim k \cdot \text{Range}(x)$$

(10)

where $k$ is a constant in the range $(0.2, 0.5)$, and $x$ is the input component from the training set. Thus, the width parameter $\delta$ of the RBF kernel reflects the distribution or range of the training set’s $x$ values. Although the estimated parameters do not generate high precision models, they constitute a more convenient approximation than using default values.

Exhaustive search

Exhaustive search is the most widely used parameter selection technique. It is a straightforward approach for parameter estimation. This technique is based on evaluating model effectiveness on a grid formed by parameter tuples, commonly comprised of values for $C$ and $\varepsilon$. This approach is computationally expensive as it is a brute force method. However, the exhaustiveness also makes this technique the most precise when a big enough grid is searched.

In order to perform exhaustive search, we select $v$ training sets using the cross-validation procedure. Then, we select all the tuples $(C, \varepsilon)$ in a grid that contains all the possible values in a given range, generated with a given step size. Due to the high computational cost it is recommended to divide the search into two phases. In the first phase (coarse search), the purpose is to localize regions where the optimal values can be found. This first search is carried out with a large step size. In the second phase (fine search), the search is carried out within the candidate regions, therefore, a smaller step size is used. During the search, the model is parametrized and all tuples $(C, \varepsilon)$ are evaluated on the training set. After carrying out grid search, the tuple which resulted in the highest accuracy is selected to compute the final SVR model [33]. Algorithm 1 describes the complete search procedure.

Meta-heuristic techniques

Meta-heuristic techniques are employed to reduce computational costs in complex and wide search spaces. The heuristics aim to minimize the cost of finding global optima. Although these techniques are more complicated to implement when compared to analytical or exhaustive search, the availability of repositories and libraries in modern programming languages makes it easier to employ meta-heuristic methods. Some of the methods that are employed the most in the literature include genetic algorithms [34]–[36], differential evolution [37], and particle swarm optimization [31], [38].

![Algorithm 1](image)

This procedure can be easily extended to three parameters.
E. Kernel function selection

The selection of the kernel function relies on the application knowledge domain and should be based on the training data distribution. Some typical kernel functions are linear, polynomial, gaussian, and RBF-based functions:

- Linear \( k(x_i, x_j) = x_i \cdot x_j \)
- Polynomial \( k(x_i, x_j) = (x_i \cdot x_j + 1)^d \)
- Gaussian \( k(x_i, x_j) = \exp \left( -\frac{(x_i - x_j)^2}{2\sigma^2} \right) \)
- RBF \( k(x_i, x_j) = \exp \left( -\frac{p\|x_i - x_j\|^2}{2} \right) \)

F. Error estimation

Similar to parameter selection, a standard method for error estimation does not exist. We applied the Mean Absolute Error (MAE), the Root of the Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). However, since MAPE is vulnerable to divisions by zero or values close to zero, we also used the error measure called Mean Arc-Tangent Absolute Error (MAAPE). According to [39], MAAPE keeps the ideas behind MAPE and overcomes the zero or close-to-zero values problem. The method uses bounded influences for outliers, considering the ratio as an angles instead of a slope. Additionally, we used the Index of Agreement (IA) metric, which varies between 0 and 1, as a standardized measure of prediction error. The error metrics are defined as follows:

\[
MAE = \frac{1}{N} \sum_{i=1}^{N} |y_i - \hat{y}_i| \quad (11)
\]

\[
RMSE = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \quad (12)
\]

\[
MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \times 100\% \quad (13)
\]

\[
MAAPE = \frac{1}{N} \sum_{i=1}^{N} \arctan \left( \frac{y_i - \hat{y}_i}{y_i} \right) \quad (14)
\]

\[
IA = 1 - \frac{\sum_{i=1}^{N} (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N} (\hat{y}_i - \bar{y} + y_i - \bar{y})^2} \quad (15)
\]

where \( y_i \) is the response value or ground truth, \( \hat{y}_i \) is the value predicted by the SVR model, \( \bar{y}_i \) is the mean of the ground truth data, and \( N \) is the number of datums for which predictions were carried out.

III. RESULTS AND DISCUSSION

We selected four stations for modeling particulate matter concentration behavior in Santa Marta. The criteria we employed were to select stations that were active and that had devices for measuring PM\(_{10}\) (see Table 2). Data for each station was published by the SVCA between 1999 and 2015.

Each dataset was transformed into a monthly time series as explained in Section 2.2. Fig. 4 shows the histogram of PM\(_{10}\) concentration for each station, along with the estimated normal density function in blue, and the kernel density estimation function in red.

Each set of data was also smoothed by means of a Savitsky-Golay filter (Eq. 2). In order to adjust the filter’s parameters, we tried out different values for \( n \). Fig. 5 shows the effect of the filter on a segment of a data series as the number of samples varies from 5 to 17. This way, we set \( n=13 \) and \( p=7 \).

Fig. 6 displays the results of the smoothing algorithm for each data series. We can observe that the smoothed series maintain the behavior of the original data. The filter reduces pronounced high and lows values, as well as small fluctuations in short time lapses.

We normalized the smoothed series before further processing, since normalization is a standard preprocessing step for facilitating convergence in SVR-based models.

The next step was to parametrize the model for the training phase. We employed Algorithm 1 to perform an exhaustive search of the SVR model parameters. In the first search, we evaluated values in the (0,1) range for \( \varepsilon \), with step size 0.1, and values in the \((2^{-1},2^0)\) range for parameter \( C \), with step size 1. In the second phase we evaluated values in in the ranges \((B-0.25,B+0.25)\) and \((2^{C-1},2^{C})\), for \( \varepsilon \) and \( C \) respectively, where \( B \) represents the best value found in the coarse search for both parameters. For this phase, the step sizes were 0.01 for \( \varepsilon \) and 1/6 for \( C \). We used Linear, Radial, and Sigmoid kernel functions for each parameter tuple we evaluated. Radial and Sigmoid kernels required parameters \( \delta \) and \( \gamma \), which were also optimized with exhaustive search.

Fig. 7 shows several examples of errors produced by different parameter tuples using \( \varepsilon \)-SVR modeling. The left side of the figure shows the behavior of error for the first phase, while the right side shows the behavior of error for the second phase. Figures 7a, 7b, 7c, and 7d show the results for stations 2, 6, 9 and 12 respectively. Blue areas represent regions with the best precision, and red areas represent regions with the worst precision. The scale on the right side of each figure is used for reference, as error levels colors in the two optimization phases can not be directly compared.

Once we obtained the best set of parameters for each time series, we evaluated two regression models: \( \varepsilon \)-SVR and \( \nu \)-SVR. The \( \nu \) parameter for the latter is usually related to the proportion of desired support vectors with respect to the number of samples in the dataset.
Contrary to $\varepsilon$-SVR, $\nu$-SVR controls the amount of data employed for each support vector. Nevertheless, $\varepsilon$-SVR controls error by penalizing values bigger than $\varepsilon$ based on the value assigned to $C$.

Table 4 shows the results of each regression model for each time series using different kernels. Each value in the table represents the error values obtained in the coarse and fine exhaustive search phases. Additionally, the number of support vectors for each model is shown. The lowest error
for all time series was obtained by employing a radial kernel and ε-SVR regression. It is important to note that ε-SVR used significantly more support vectors than γ-SVR in the majority of the cases.

After parameter selection and training, we tested the best model on ~70% of the data for each time series. Table 5 shows accuracy measurements according to the metrics described in section II.F. Due to the varying properties of each metric, interpretation can be difficult. Some metrics depend on scale, while others can be symmetrical or asymmetrical.

A widely employed metric is the Index of Agreement [40], which takes values in the [0-1] range, where 1 is a perfect match, and 0 indicates no agreement at all. The IA metric resulted in values close to 1 for all tests, showing high agreement between the model predictions and the true values for each test sample. One important feature of the IA index is the sensitivity to extreme. However, the Savitsky-Golay filter smoothed the series, and reduced the effect of such values.

Although forecast was not our primary goal, we also estimated model forecast accuracies. For this purpose, we analyzed one-step forecast by training the models and computing error levels over all the available data. Three scenarios were chosen for training: using raw data, using smoothed data, and using smoothed and normalized data. In the first two scenarios, the error was estimated based on the original series values, while in the third scenario, the error was estimated based on a normalized version of the original data. Table 6 shows the errors estimated for each scenario. The third scenario yielded the smallest error because SVM assumes data to be in normalized.

Scaling or normalizing prevents large numerical values from dominating the model and small values from being treated as irrelevant. According to [41], the advantage of preprocessing data for SVR model is due to the fact that kernel values usually depend on the inner product of the feature vectors. In this case, normalization avoids numerical issues such as floating point overflow and underflow.

In [26], we show the accuracy of a neural network for predicting PM10 levels. Table 7 compares the results obtained by SVR and the Neural Network model using one-step prediction. SVR-based models obtained the best results when predicting a single step. However, the models lose accuracy when predicting farther away future values. In these cases, Neural Network-based models tend to be more accurate; in other words, the SVR model does not have good forecasting capabilities when more than one-step prediction is needed.

<table>
<thead>
<tr>
<th>Table 4. Accuracy results for each kernel and regression type. The parameters for each experiment were set up based on Algorithm 1.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Type</strong></td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>E. id</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 5. Error estimation for the best model for each time-series.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>E. id</strong></td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 6. Error measure for one-step (month) forecasting.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>E. id</strong></td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Table 7. Error comparison.</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>E. id</strong></td>
</tr>
<tr>
<td>---------</td>
</tr>
<tr>
<td>2</td>
</tr>
<tr>
<td>6</td>
</tr>
<tr>
<td>9</td>
</tr>
<tr>
<td>12</td>
</tr>
</tbody>
</table>
Fig. 7. Color map representation of coarse (left) and fine (right) parameter optimizations. The explored parameters are $\epsilon$ and $C$ for SVR, on four time-series. From top to bottom a), b), c), and d) correspond to stations 2, 6, 9, and 12 in Table 2, respectively.

IV. CONCLUSIONS

Methods for modeling pollution are essential for enacting regulatory policies. Traditional approaches for analyzing pollution data use the highest possible granularity, that is, at least one hour or a day measurement periods. The data used in this work, available for Santa Marta city, does not allow for this granularity level due to inadequate measurement equipment and human error in data collection procedures. For this reason, we scaled the time series to monthly data. Previous works focused to bio-inspired techniques for pollution modelling, while the method developed in this paper focuses on pre-processing the time series and using different SVR models. Our method generated more accurate models in comparison to previous work in one-step prediction scenarios. The best model on the Santa Marta pollution data was obtained using a $\nu$-SVR regressor with a radial kernel function.

(Advance online publication: 12 August 2019)
One of the main limitations of the models proposed in this work is the lack of additional variables, such as environmental or correlating factors like wind direction, industry locations, traffic, and temperature, among others. The construction of a more robust and general model should include additional information, such as the aforementioned factors, as well as data acquired more regularly and with better measuring devices and methodology.

REFERENCES


