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

Sanchez-Torres G., and Díaz 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 ν-support vector regression (ν-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 (NO_2), ozone (O_3), sulphur oxide (SO_2) and particulate matter PM_{10} [7]–[9]. On the other hand, non-conventional pollutants include: Benzene (C_6H_6), lead (Pb) and its composites, cadmium (Cd), mercury (Hg), hydrogen Sulphur (H_2S), 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 aero dynamical diameters of 2.5 to $10\mu m$ (PM_{10}), less than $2.5\mu m$ ($PM_{2.5}$), and smaller than $0.1\mu m$ ($PM_{0.1}$) [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 Particulas 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. PM_{10} pollutants have been monitored in 85% of these stations, while SO_2 and NO_2 contaminants have been monitored by 35%, O_3 by 25%, PST by 24%, CO by 23%, and $PM_{2.5}$ by 15% of these

Manuscript received October 28, 2018. This work was supported by the Universidad del Magdalena.

Sanchez-Torres G., and Diaz I., are with the Faculty of Engineering, Universidad del Magdalena, Carrera 32 No. 22-08, Santa Marta, Colombia. (e-mail: gsanchez@unimagdalena.edu.co, idiaz@unimagdalena.edu.co).

(Advance online publication: 12 August 2019)

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.

TABLE 1.
MAXIMUM PERMISSIBLE VALUES FOR SEVERAL CONTAMINANTS IN COLOMBIA AND THE LEVELS RECOMMENDED BY THE WHO.

Contaminant	Maximum Permissible value $\mu g/m^3$		Average exposure time
	Colombia	WHO	
PST	100	150	Annual
PM_{10}	50	20	Annual
$PM_{2.5}$	25	10	Annual
$SO2$	250	20	24 hours
NO_2	100	40	Annual
O_3	80	100	8 hours
CO	10000	10000	8 hours

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 registerered 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.

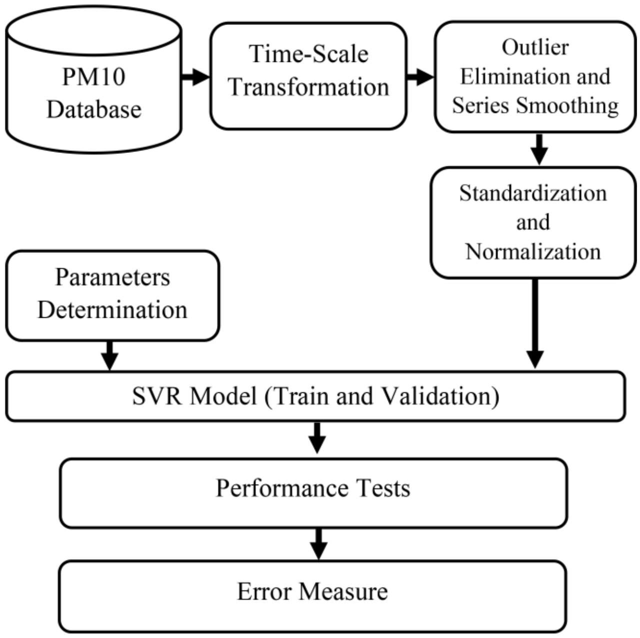

Fig. 1.Block diagram of the working methodology.

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.

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

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 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_{-n}, …, T_0, …, T_n) [26], and is defined as follows[24]:

$$\overline{T_j} = \frac{1}{n}\sum_{i=-m}^{m} C_i T_{j+1} \tag{2}$$

where $\overline{T_j}$ 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.
MONITORING STATIONS AND MEASURED VARIABLES.

Id	Name	Contaminant	State	Starting Date
1	Invemar	*PST*	Active	1999
2	C. Santa Marta	PM_{10}	Active	2007
3	C. Ejecutivo	*PST*	Active	1999
4	Cajamag	*PST*	Inactive	2003
5	Batallón	*PST*	Active	1999
6	Molinos	PM_{10}	Active	2011
7	Zuana	PM_{10}	Inactive	2007
8	Aeropuerto	*PST*	Inactive	1999
9	Don Jaca	PM_{10} *y PST*	Active	1999
10	Alcatraces	PM_{10} *y PST*	Inactive	1999
11	Papare	*PST*	Inactive	2005
12	Costa Verde	PM_{10} *y PST*	Active	2008

TABLE 3.
GEOGRAPHICAL LOCALIZATION OF ACTIVE STATIONS WITHIN THE LIMITS OF THE SANTA MARTA MUNICIPALITY.

Id	Geographical localization	
	Latitude	Longitude
1	11°15'02.8940"	11°15'02.8940"
2	11°14'25.6063"	11°14'25.6063"
3	11°14'23.3610"	11°14'23.3610"
5	11°13'57.2185"	11°13'57.2185"
6	11°11'40.5247"	11°11'40.5247"
9	11°05'54.5046"	11°05'54.5046"

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 $X \in R^N$. The fitting procedure transforms the observation set from a n-dimensional space to a m-dimensional space, such that m>n. The transformation is performed by a function or *kernel* Φ:n→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 regression ε (ε-SVR) [27]–[29]. Let $X=\{(x_1,y_1), \ldots, (x_i,y_i)\}$ be the observation set, where each element $x_i \in R^N$ represents an input, and $y_i \in R^1$ represents an output value. The optimization problem of a v-SVR regression can be expressed with the following restrictions:

Fig. 2. The monitoring stations located in Santa Marta, Colombia.

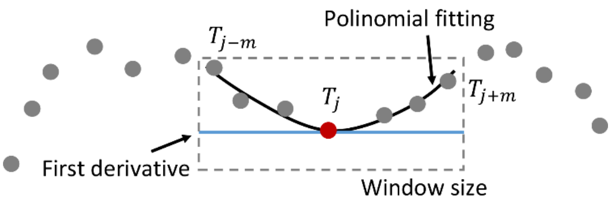

Fig. 3. Savitzky-Golay filter.

$$\min \frac{1}{2} w^T W + C(v\varepsilon + \frac{1}{i}\sum_{i=1}^{l}\xi_i + \xi^*) \tag{3}$$

$$\begin{cases} \left(w^T\Phi(x_i)+b\right) - y_i \le \varepsilon + \xi_i \\ y_i - (wT\Phi(x_i)+b) \le \varepsilon + \xi^* \\ \xi_i, \xi^* \ge 0, i = 1,\ldots,l; \varepsilon \ge 0 \end{cases} \tag{4}$$

where $0 \le v \le 1$, C is the regularization parameter, and $\Phi(x_i)$ is the space transformation function, or *kernel*. The ε value is the cost function, or loss, and represents an error tolerance. Thus, if $w^T\Phi(x_i)$ is in the $yi \pm \varepsilon$ range, it is not considered to be an error. As described by [27] and [29], a proper estimation of ε 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 α*, η*, β ≥ 0. In this way, we obtain a dual Lagrange formulation for *v≥0, C>0:*

$$\max \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)y_i - \frac{1}{2}\sum_{i,j=1}^{l}(\alpha_i^* - \alpha_i)(\alpha_j^* - \alpha_j)k(x_i, x_j) \tag{5}$$

Subject to:

$$\begin{cases} \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \\ 0 \le \alpha_i^{(*)} \le \dfrac{C}{l} \\ \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) \le C, v \end{cases} \tag{6}$$

Therefore, the regression function is approximated as follows:

$$y = f(x) = \sum_{i=1}^{l}(\alpha_i^* - \alpha_i)k(x_i, x) + b \tag{7}$$

where *b* represents a systematic error or noise, and *k* is the dot product in the feature space yielded by the transformation function *Φ* or *kernel*:

$$k(x, y) = (\Phi(x).\Phi(y)) \tag{8}$$

D. Parameter selection for the v-SVR model

Three parameters must be established to build the SVR regression model: *C*, *ε*, and *δ*. *C* is the penalty parameter. *ε* is the tolerance parameter, which measures the degree of fitting of the regression model to the training data. *δ* is a parameter related to the selected *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 *ε*:

$$C = \max(|\bar{y} + 3\sigma_y|, |\bar{y} - 3\sigma_y|)$$
$$\varepsilon = \tau\sigma_{noise}\sqrt{\frac{\ln(n)}{n}} \tag{9}$$

Where $\bar{y}$ is the mean value of the training data corresponding to the outputs $y_i \in R^1$, and σ_y is the standard deviation of values *y* in the training data. The σ_{noise} parameter is the standard deviation of the noise estimation for the training set, and τ 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 δ that can be approximated by:

$$\delta \sim k \cdot Range(x) \tag{10}$$

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 δ 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 *ε*. 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, ε)* 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, ε)* 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. AN EXHAUSTIVE SEARCH FOR PARAMETERS *C* AND *ε*

```
Select v subsets using cross-validation
Establish Δc and Δε as the step-sizes, define
(C1, ε1) and (Cf, ε1)
For C = C1 through Cf step = Δc
   For ε= ε1 through εf step = Δε
     Accu(C, ε) = Evaluate (C, ε)
    Fstop
Fstop
Select the candidate region and define (C'1, ε'1)
y (C'f, ε'f)
For C=C'1 through C'f step =  Δ'f
   For ε = ε'1 through ε'f step = Δ'ε
     Accu(C, ε) = Evaluate (C, ε)
   Fstop
Fstop
(Coptimal, εoptimal)=minAccuracy(C, ε)
```

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

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:

$$MAE = \frac{1}{N}\sum_{i=1}^{N}\left|y_i - \overline{y_i}\right| \tag{11}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}\left(y_i - \overline{y_i}\right)^2}{N}} \tag{12}$$

$$MAPE = \frac{1}{N}\sum_{i=1}^{N}\frac{\left|y_i - \overline{y_i}\right|}{y_i} x100\% \tag{13}$$

$$MAAPE = \frac{1}{N}\sum_{i=1}^{N}\arctan\left(\left|\frac{y_i - \overline{y_i}}{y_i}\right|\right) \tag{14}$$

$$IA = 1 - \frac{\sum_{i=1}^{N}(y_i, \hat{y}_i)^2}{\sum_{i=1}^{N}(\left|\hat{y}_i - \overline{y}\right| + \left|y_i - \overline{y}\right|)^2} \tag{15}$$

where y_i is the response value or ground truth, $\hat{y}_i$ is the value predicted by the SVR model, $\overline{y}_i$ 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 ε, with step size 0.1, and values in the (2^{-2}, 2^{10}) 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 ($2^{B-1,}$ 2^{B+1}), for ε 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 ε 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 δ and γ, which were also optimized with exhaustive search.

Fig. 7 shows several examples of errors produced by different parameter tuples using *ε-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: *ε-SVR* and *v-SVR*. The *v* parameter for the latter is usually related to the proportion of desired support vectors with respect to the number of samples in the dataset.

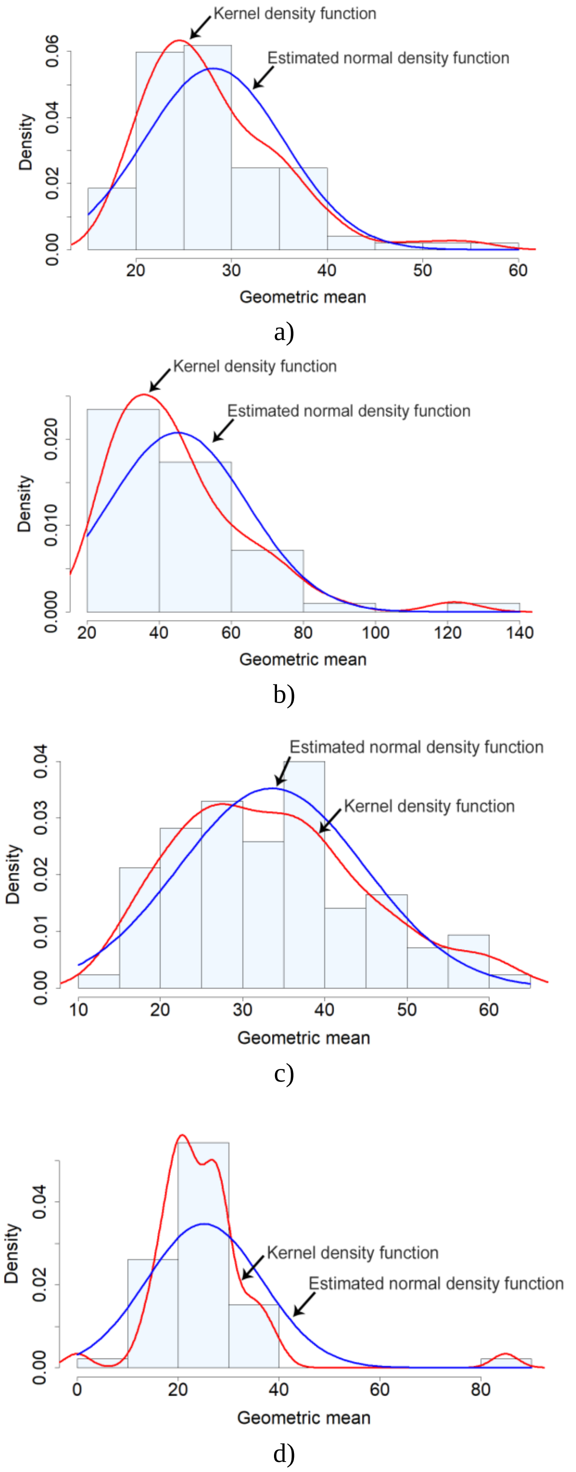

Fig. 4. Histograms of PM10 monthly concentration for stations 2, 6, 9, and 12, along with the estimated normal density functions and the kernel density functions.

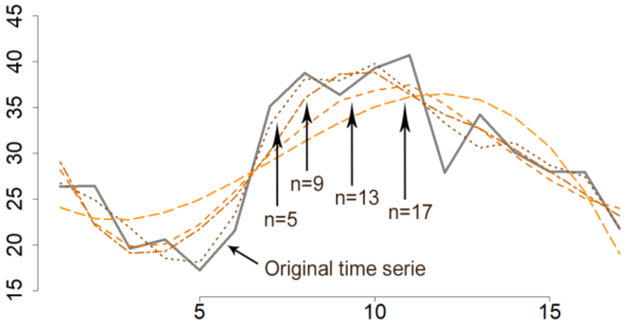

Figure 5. Effect of window size (n) variation on the Sgolay filter.

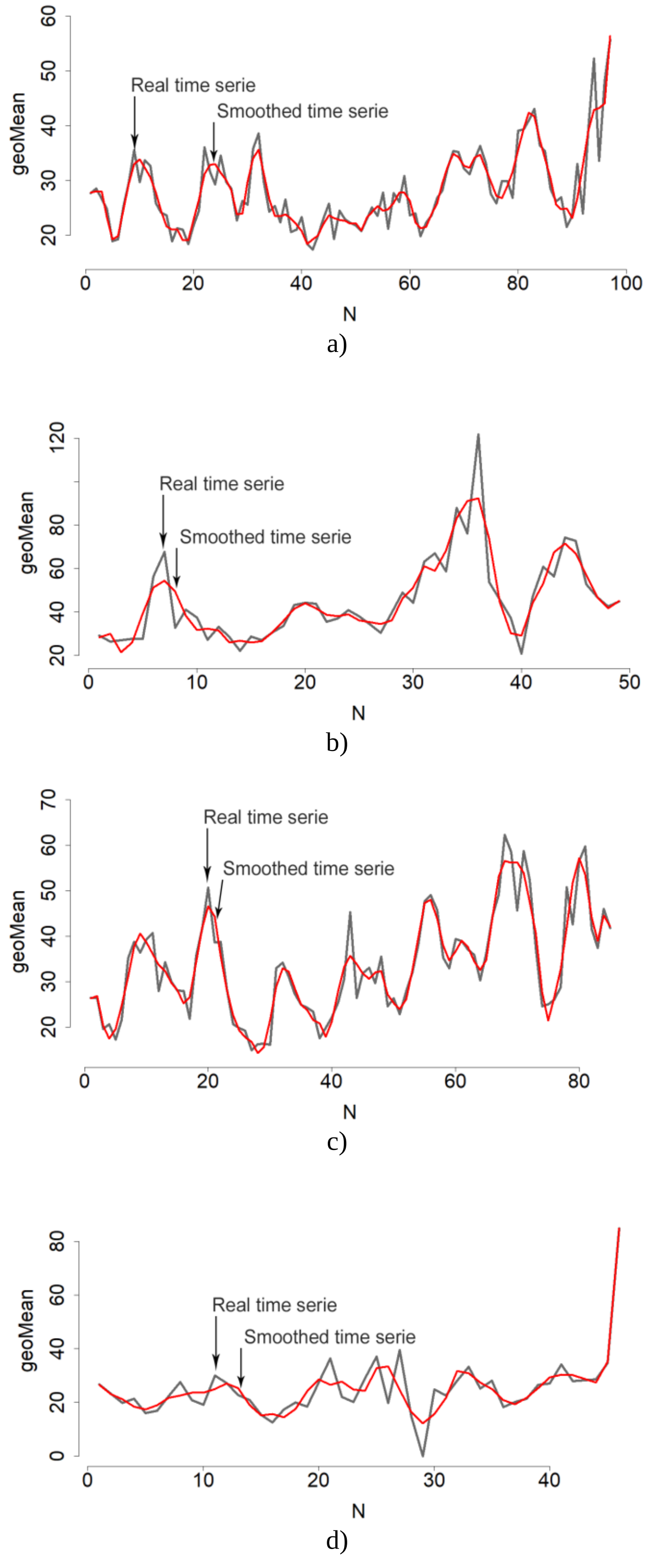

Fig. 6. Real and smoothed data for stations a) station id 2, b) station id 6, c) station id 9 and d) station id 12.

Contrary to *ε-SVR*, *ν-SVR* controls the amount of data employed for each support vector. Nevertheless, *ε-SVR* controls error by penalizing values bigger than ε 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 *ν*-SVR regression. It is important to note that *ν*-SVR used significantly more support vectors than *ε*-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 there is 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 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 preprocecing data for *SVR* model is due to 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 PM_{10} 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 4.
ACCURACY RESULTS FOR EACH KERNEL AND REGRESSION TYPE. THE PARAMETERS FOR EACH EXPERIMENT WERE SET UP BASED ON ALGORITHM 1.

		Linear				Polynomial				Radial				Sigmoid			
		Coarse		Fine		Coarse		Fine		Coarse		Fine		Coarse		Fine	
E. Id	Tipo	perf	vs	perf	vs	perf	vs	perf	vs	perf	vs	perf	vs	perf	vs	perf	vs
2	ε-SVR	0.7933	29	0.6089	22	0.7500	30	0.5639	22	0.3366	16	0.3328	27	1.9692	53	0.6260	17
	ν-SVR	0.7524	51	0.2789	58	0.7713	51	0.4341	93	0.2989	53	0.2785	57	2.1572	50	0.5429	97
6	ε-SVR	0.5571	15	0.4010	12	0.8034	16	0.4002	14	0.1042	37	0.0568	26	0.7038	11	0.4327	9
	ν-SVR	0.5742	26	0.2158	25	0.8297	26	0.4177	41	0.0979	27	0.0567	32	0.9975	27	0.4558	36
9	ε-SVR	0.6132	20	0.4652	16	0.7169	49	0.4230	27	0.4094	23	0.4142	23	0.6906	23	0.4609	21
	ν-SVR	0.6226	44	0.4330	61	0.7329	44	0.3945	85	0.4671	46	0.3804	85	0.8592	44	0.4251	61
12	ε-SVR	0.8250	9	0.1491	9	0.8616	28	0.7932	27	0.0240	46	0.0179	36	1.080	7	0.8732	9
	ν-SVR	0.9506	24	0.7310	29	0.7423	25	0.7334	36	0.0218	29	0.0137	46	1.1359	24	0.7232	38

TABLE 5.
ERROR ESTIMATION FOR THE BEST MODEL FOR EACH TIME-SERIES.

E. id	MAE	RMSE	MAPE(%)	MAAPE(Rad)	IA
2	0.365307	0.467860	298.3705	0.923807	0.919576
6	0.100931	0.151759	40.1036	0.874319	0.993306
9	0.462434	0.554829	267.6361	0.881412	0.904340
12	0.065141	0.103595	43.1614	0.896402	0.993000

TABLE 6.
ERROR MEASURE FOR ONE-STEP (MONTH) FORECASTING.

	Raw		Smooth		Smooth-Norm	
E. id	RMSE	MAPE(%)	RMSE	MAPE(%)	RMSE	MAPE(%)
2	5.07270	9.10091	2.28103	6.28736	0.30978	6.55967
6	28.47664	63.37646	0.68135	4.03060	0.03393	2.17593
9	19.48254	80.62483	18.45204	51.55000	0.73111	1.75134
12	56.14660	66.21061	2.78011	5.98025	0.23700	3.78349

TABLE 7.
ERROR COMPARISON.

	SVR	NN
E. id	RMSE	RMSE
2	0.30978	0.525
6	0.03393	0.198
9	0.73111	0.125
12	0.23700	0.458

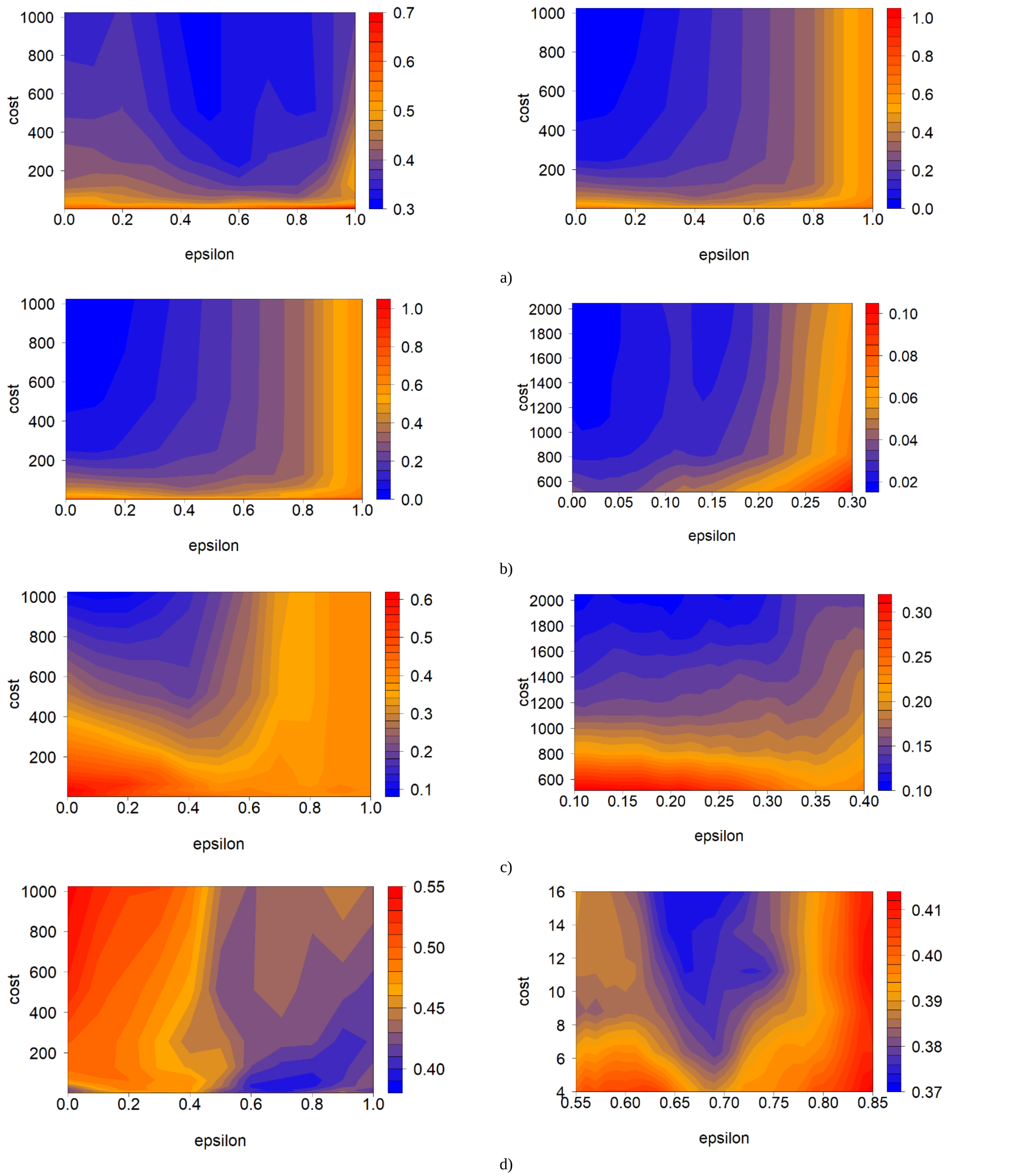

Fig. 7. Color map representation of coarse (left) and fine (right) parameter optimizations. The explored parameters are ε and C for SVR, on four time-series. From top to bottom a), b) , c), and d) correspond to stations 2, 6, 9, and 12 in Table 2, respectively.

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 ν-SVR regressor with a radial kernel function.

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 adquired more regularly and with better measuring devices and methodology.

REFERENCES

[1] B. Ando, S. Baglio, S. Graziani, E. Pecora, and N. Pitrone, “A predictive model for urban air pollution evaluation,” in , IEEE Instrumentation and Measurement Technology Conference, 1997. IMTC/97. Proceedings. Sensing, Processing, Networking, 1997, vol. 2, pp. 1056–1059 vol.2.

[2] A. Clifford, L. Lang, R. Chen, K. J. Anstey, and A. Seaton, “Exposure to air pollution and cognitive functioning across the life course – A systematic literature review,” Environmental Research, vol. 147, pp. 383–398, May 2016.

[3] Z. Deng et al., “Association between air pollution and sperm quality: A systematic review and meta-analysis,” Environmental Pollution, vol. 208, Part B, pp. 663–669, Jan. 2016.

[4] M. Franchini, C. Mengoli, M. Cruciani, C. Bonfanti, and P. M. Mannucci, “Association between particulate air pollution and venous thromboembolism: A systematic literature review,” European Journal of Internal Medicine, vol. 27, pp. 10–13, Jan. 2016.

[5] A. Sureerat, Konglok, and P. Nopparat, "Numerical Computations of Three-dimensional Air-Quality Model with Variations on Atmospheric Stability Classes and Wind Velocities using Fractional Step Method," IAENG International Journal of Applied Mathematics, vol. 46, no.1, pp112-120, 2016.

[6] T. Taj, K. Jakobsson, E. Stroh, and A. Oudin, “Air pollution is associated with primary health care visits for asthma in Sweden: A case-crossover design with a distributed lag non-linear model,” Spatial and Spatio-temporal Epidemiology, vol. 17, pp. 37–44, May 2016.

[7] Á. Gómez-Losada, J. C. M. Pires, and R. Pino-Mejías, “Characterization of background air pollution exposure in urban environments using a metric based on Hidden Markov Models,” Atmospheric Environment, vol. 127, pp. 255–261, Feb. 2016.

[8] J. Sunyer, X. Basagaña, J. Belmonte, and J. Antó, “Effect of nitrogen dioxide and ozone on the risk of dying in patients with severe asthma -- Sunyer et al. 57 (8): 687 -- Thorax,” 2002. [Online]. Available: http://thorax.bmj.com/content/57/8/687. [Accessed: 27-May-2016].

[9] S. Vedal, M. Brauer, R. White, and J. Petkau, “Air pollution and daily mortality in a city with low levels of pollution.,” Environ Health Perspect, vol. 111, no. 1, pp. 45–52, Jan. 2003.

[10] Ministerio de ambiente, vivienda y desarrollo territorial, “Resolución 601.” Ministerio de Ambiente, Vivienda y Desarrollo Territorial, Mar-2010.

[11] R. Chen et al., “Beyond PM2.5: The role of ultrafine particles on adverse health effects of air pollution,” Biochimica et Biophysica Acta (BBA) - General Subjects, 2016.

[12] F. J. Kelly and J. C. Fussell, “Size, source and chemical composition as determinants of toxicity attributable to ambient particulate matter,” Atmospheric Environment, vol. 60, pp. 504–526, Dec. 2012.

[13] World Health Organization, Ed., Air quality guidelines for Europe, 2nd ed. Copenhagen: World Health Organization, Regional Office for Europe, 2000.

[14] World Health Organization, Ed., Air Quality Guidelines-Global Update 2005. Particulate matter, ozone, nitrogen dioxide and sulfur dioxide, 1 ed. Copenhagen: World Health Organization, Regional Office for Europe, 2005.

[15] Ministerio de ambiente, vivienda y desarrollo territorial, Resolución 601. 2006.

[16] Ministerio de ambiente, vivienda y desarrollo territorial, Resolución 610. 2010.

[17] IDEAM, Informe del Estado de la Calidad del Aire en Colombia 2007 - 2010. Bogotá, D. C.,: Publicación aprobada por el Comité de Comunicaciones y Publicaciones del IDEAM, 2012.

[18] E. Golub, G. Sanchez-Martinez, I. Klytchnikova, C. M. Molina, and J. C. Belausteguigoitia, “Environmental health costs in Colombia : the changes from 2002 to 2010,” The World Bank, 92956, Jun. 2014.

[19] Corpamag, “Sistema de vigilancia de la calidad del aire - SVCA - informe de resultados diciembre 2015,” Corporación Autonoma Regional del Magdalena, CORPAMAG, Santa Marta, Magdalena, Dec. 2015.

[20] Corpamag, “SVCA,” Sistema de Vigilancia de la Calidad del Aire - SVCA- del departamento del Magdalena. [Online]. Available: http://www.corpamag.gov.co/index.php/es/informacion-ambiental/aire. [Accessed: 22-Jun-2016].

[21] “Aire.” [Online]. Available: http://www.corpamag.gov.co/index.php/es/informacion-ambiental/aire. [Accessed: 30-Jun-2016].

[22] H. Liu, S. Shah, and W. Jiang, “On-line outlier detection and data cleaning,” Computers & Chemical Engineering, vol. 28, no. 9, pp. 1635–1647, Aug. 2004.

[23] C. Chen and L.-M. Liu, “Joint Estimation of Model Parameters and Outlier Effects in Time Series,” Journal of the American Statistical Association, vol. 88, no. 421, pp. 284–297, 1993.

[24] A. Savitzky and M. J. E. Golay, “Smoothing and Differentiation of Data by Simplified Least Squares Procedures.,” Anal. Chem., vol. 36, no. 8, pp. 1627–1639, Jul. 1964.

[25] R. Valencia, G. Sanchez, and I. Diaz, “A general regression neural network for modeling the behavior of PM10 concentration level in Santa Marta, Colombia,” Journal of Engineering and Applied Sciences, vol. 11, no. 11, pp. 7085–7092, Jun. 2016.

[26] Kewalee Suebyat, and Nopparat Pochai, "A Numerical Simulation of a Three-dimensional Air Quality Model in an Area Under a Bangkok Sky Train Platform Using an Explicit Finite Difference Scheme," IAENG International Journal of Applied Mathematics, vol. 47, no.4, pp471-476, 2017.

[27] C.-C. Chang and C.-J. Lin, “Training V-support Vector Regression: Theory and Algorithms,” Neural Comput., vol. 14, no. 8, pp. 1959–1977, Aug. 2002.

[28] B. Schölkopf, A. J. Smola, R. C. Williamson, and P. L. Bartlett, “New Support Vector Algorithms,” Neural Comput., vol. 12, no. 5, pp. 1207–1245, May 2000.

[29] B. Schölkopf, P. Bartlett, A. Smola, and R. Williamson, Shrinking the Tube: A New Support Vector Regression Algorithm. 1999.

[30] P. Tsirikoglou, S. Abraham, F. Contino, C. Lacor, and G. Ghorbaniasl, “A hyperparameters selection technique for support vector regression models,” Applied Soft Computing, vol. 61, pp. 139–148, Dec. 2017.

[31] S. M. H. Bamakan, H. Wang, and A. Z. Ravasan, “Parameters Optimization for Nonparallel Support Vector Machine by Particle Swarm Optimization,” Procedia Computer Science, vol. 91, pp. 482–491, Jan. 2016.

[32] V. Cherkassky and Y. Ma, “Practical Selection of SVM Parameters and Noise Estimation for SVM Regression,” Neural Netw., vol. 17, no. 1, pp. 113–126, Jan. 2004.

[33] C. Hsu, C. Chang, and C. Lin, A practical guide to support vector classification. 2010.

[34] W. Zhao, T. Tao, and E. Zio, “System reliability prediction by support vector regression with analytic selection and genetic algorithm parameters selection,” Applied Soft Computing, vol. 30, pp. 792–802, May 2015.

[35] K.-Y. Chen, “Forecasting systems reliability based on support vector regression with genetic algorithms,” Reliability Engineering & System Safety, vol. 92, no. 4, pp. 423–432, Apr. 2007.

[36] H. Shafizadeh-Moghadam, A. Tayyebi, M. Ahmadlou, M. R. Delavar, and M. Hasanlou, “Integration of genetic algorithm and multiple kernel support vector regression for modeling urban growth,” Computers, Environment and Urban Systems, vol. 65, pp. 28–40, Sep. 2017.

[37] J. Wang, L. Li, D. Niu, and Z. Tan, “An annual load forecasting model based on support vector regression with differential evolution algorithm,” Applied Energy, vol. 94, pp. 65–70, Jun. 2012.

[38] W. Zhao, T. Tao, E. Zio, and W. Wang, “A Novel Hybrid Method of Parameters Tuning in Support Vector Regression for Reliability Prediction: Particle Swarm Optimization Combined With Analytical Selection,” IEEE Transactions on Reliability, vol. 65, no. 3, pp. 1393–1405, Sep. 2016.

[39] S. Kim and H. Kim, “A new metric of absolute percentage error for intermittent demand forecasts,” International Journal of Forecasting, vol. 32, no. 3, pp. 669–679, Jul. 2016.

[40] C. J. Willmott, “On the validation of models,” Physical Geography, vol. 2, no. 2, pp. 184–194, 1981.

[41] S. F. Crone, J. Guajardo, and R. Weber, The Impact of Preprocessing on Support Vector Regression and Neural Networks in Time Series Prediction. 2006.