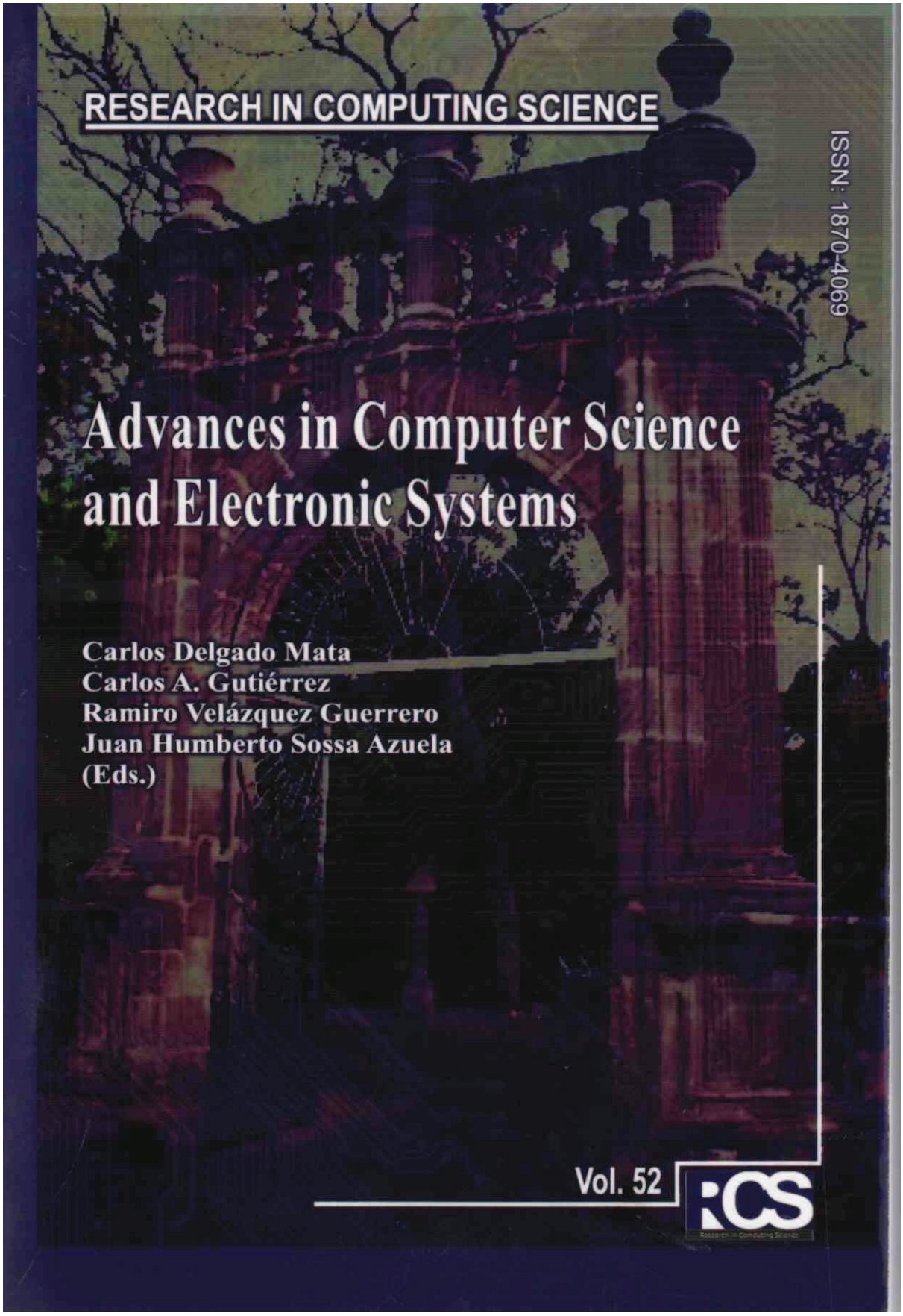

RESEARCH IN COMPUTING SCIENCE
ISSN: 1870-4069
Advances in Computer Science and Electronic Systems
Carlos Delgado Mata
Carlos A. Gutiérrez
Ramiro Velázquez Guerrero
Juan Humberto Sossa Azuela
(Eds.)
Vol. 52
RCS

Research in Computing Science

Research in Computing Science es una publicación trimestral, de circulación internacional, editada por el Centro de Investigación en Computación del IPN, para dar a conocer los avances de investigación científica y desarrollo tecnológico de la comunidad científica internacional. **Volumen 52**, abril, 2011. Tiraje: 500 ejemplares. *Certificado de Reserva de Derechos al Uso Exclusivo del Título* No. 04-2004-062613250000-102, expedido por el Instituto Nacional de Derecho de Autor. *Certificado de Licitud de Título* No. 12897, *Certificado de licitud de Contenido* No. 10470, expedidos por la Comisión Calificadora de Publicaciones y Revistas Ilustradas. El contenido de los artículos es responsabilidad exclusiva de sus respectivos autores.

Impreso en la Ciudad de México, en los Talleres Gráficos del IPN – Dirección de Publicaciones, Tres Guerras 27, Centro Histórico, México, D.F. Distribuida por el Centro de Investigación en Computación, Av. Juan de Dios Bátiz S/N, Esq. Av. Miguel Othón de Mendizábal, Col. Nueva Industrial Vallejo, C.P. 07738, México, D.F. Tel. 57 29 60 00, ext. 56571.

Editor Responsable: *Juan Humberto Sossa Azuela*, RFC *SOAJ560723*

Research in Computing Science *is published by the Center for Computing Research of IPN.* ***Volume 52****, april, 2011. Printing 500.*

Printed in Mexico City, april, 2011, in the IPN Graphic Workshop – Publication Office.

Volume 52

Volumen 52

Modeling PMx Trends Contaminants by using Support Vector Machines

Artemio Sotomayor-Olmedo[1], Marco Aceves-Fernandez[1], Efren Gorrostieta-Hurtado[1], Carlos Pedraza-Ortega[1], Emilio Vargas-Soto[1], Ruth Rico-Hernandez[1]

[1]UniversidadAutonoma de Queretaro –Facultad de Informatica, Av. De las Ciencias. S/N, 76230 Queretaro, Mexico
{artemio.sotomayor, marco.aceves, efren.gorrostieta, carlos.pedraza,rico}@uaq.mx

Abstract. Monitoring, modeling and forecasting of air quality parameters are important topics in environmental and health research due to their impact caused by exposing to airborne particles in urban environments. The aim of this article is to show that forecast of daily airborne pollution using support vector machines (SVM) and Kernel functions such as Polynomial, Gaussian, and Spline are feasible. Results are presented using data measurements of Particulate Matter of aerodynamical size on the order of 10 and 2.5 micrograms (PMx) in London-Bloomsbury at south England.

Keywords: Particulate matter, Support Vector Machines, Kernel techniques, PMx, airborne pollution, forecast.

1 Introduction

In recent times, urban air pollution has been a growing problem especially for urban communities. Size, shape and chemical properties govern the lifetime of particles in the atmosphere and the site of deposition within the respiratory tract. Health effects differ upon the size of airborne particulates. In this contribution, PM10 (particles less or equal than 10 micrometers) and PM2.5 (particles less or equal than 2.5 micrometers) are considered due to its effect on human health, according to several authors [1-6]. This is the primary reason this research has been done; to monitor, and model the levels and spread of PMx in urban environments. In previous contributions, it has been shown that forecast of concentration levels of PM10 may be possible by using other techniques such as neural networks and various fuzzy clustering algorithms [7-8]. However, even though these works have shown that is feasible to accurately model the non-linear behavior of the system, a more robust model is needed with an enhanced method to reduce the error between the raw data and the model. For this reason, support vector machines (SVM) are chosen for this work. In this appraisal, the modeling will be carried out using support vector machines working in regression mode. Support vector machines are a recent statistical learning technique, based on machine learning and generalization theories, it implies an idea and could be considered as a method to minimize the risk [9]. Also, a

Advances in Computer Science and Electronic Systems
Research in Computing Science 52, 2011, pp. 114-122

Received 23/01/11
Accepted 20/02/11
Final version 07/03/11

generalization capability makes possible their application to modeling dynamical and non-linear data sets.

2 Support Vector Machines

2.1 Support Vector Machines Background using Regression Mode

The support vector machines (SVM) theory, was developed by Vapnik in 1995, and is applied in many machine-learning applications such as object classification, time series prediction, regression analysis and pattern recognition. Support vector machines (SVM) are based on the principle of structured risk minimization (SRM) [10-11].

In the analysis using SVM, the main idea is to map the original data $\boldsymbol{x}$ into a feature space $\boldsymbol{F}$ with higher dimensionality via non-linear mapping function $\boldsymbol{\phi}$, which is generally unknown, and then carry on linear regression in the feature space [10]. Thus, the regression approximation addresses a problem of estimating function based on a given data set (where x_i represent the input vectors, $d_i$are the desired values), which is produced from the ϕ function. SVM method approximates the function by:

$$y = \sum_{i=1}^{m} w_i \phi_i(x) + b = w\phi(x) + b \tag{1}$$

where $w = [w_1, ..., w_m]$ represent the weights vector, b are the bias coefficients and $\phi(x)=[\phi_1(x),\ldots,\phi_m(x)]$ the basis function vector.

The learning task is transformed to the weights of the network at minimum. The error function is defined through the ε-insensitive loss function, $L_\varepsilon(d,y(x))$ and is given by:

$$L_\varepsilon(d, y(x)) = \begin{cases} |d - y(x)| - \varepsilon & |d - y(x)| \geq \varepsilon \\ 0 & others \end{cases} \tag{2}$$

The solution of the so defined optimization problem is solved by the introduction of the Lagrange multipliers α_i, α_i^* (where i=1,2,…,k) responsible for the functional constraints defined in Eq. 2. The minimization of the Lagrange function has been changed to the dual problem [9]:

$$\phi(\alpha, \alpha^*) = \left[\sum_{i=1}^{k} d_i(\alpha_i - \alpha_i^*) - \varepsilon \sum_{i=1}^{k} (\alpha_i - \alpha_i^*) - \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} (\alpha_i, \alpha_i^*)(\alpha_j, \alpha_j^*) K(x_i, x_j) \right] \tag{3}$$

With constraints:

$$\sum_{i=1}^{k} (\alpha_i, \alpha_i^*) = 0, \tag{4}$$

$$0 \le \alpha_i \le C, 0 \le \alpha_i^* \le C$$

Where C is a regularized constant that determines the trade-off between the training risk and the model uniformity.

According to the nature of quadratic programming, only those data corresponding to non-zero $(\alpha_i - \alpha_i^*)$pairs can be referred to support vectors (nsv). In Eq. 3 $K(x_i, x_j)=\phi(x_i)*\phi(x_j)$ is the inner product kernel which satisfy Mercer's condition [13] that is required for the generation of kernel functions given by:

$$K(x_i, x_j) = \left\langle \phi(x_i), \phi(x_j) \right\rangle \tag{5}$$

Thus the support vectors associates with the desired outputs $y(x)$ and with the input training data x can be defined by:

$$y(x) = \sum_{i=1}^{Nsv} (\alpha_i, \alpha_i^*) K(x, x_i) + b \tag{6}$$

Where x_i are learning vectors. This leads to a SVM architecture (Fig. 1) and are also founded in [9][10][14].

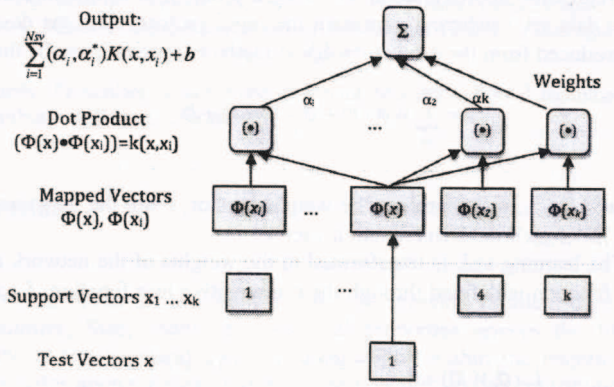


Fig. 1. Support Vector Machine Architecture.

2.2 Kernel functions.

The use of an appropriate kernel is the key feature in support vector applications, since it provide the capability of mapping non-linear data into "feature" spaces that in essence are linear, then an optimization process can be applied as in the linear case. This provides a means to dimensionality the problem properly, but still the results depends of the good selection of a set of training datasets.

The Gaussian kernel function is defined in [11-13] Eq. 7.

$$K(x_i, x_j) = \exp\left(\frac{-\|x_i - x_j\|^2}{2\sigma^2} \right) \tag{7}$$

The Gaussian kernel process delivers an estimate for the reliability of the prediction in the form of the variance of the predictive distribution and the analysis can be used to estimate the evidence in favor of a particular choice of covariance function. The covariance or kernel function can be seen as a model of the data, thus providing a principled method for model selection [13-15].

A polynomial mapping is a widely used method for non-linear modeling [13][15], defined by:

$$K(x_i, x_j) = \langle x_i, x_j \rangle^d \tag{8}$$

Unless the used of equation 8 implies an inherit problem, some Support Vector Machines become zeros, therefore is preferable to rewrite the expression as:

$$K(x_i, x_j) = \left(\langle x_i, x_j \rangle + 1\right)^d \tag{9}$$

In this survey, a Spline kernel is presented as a choice for modeling due to their flexibility. A spline, of order with N knots located at τ_s is expressed by:

$$K(x_i, x_j) = \sum_{r=0}^{\kappa} x_i^r x_j^r + \sum_{s=1}^{\kappa} (x_i - \tau_s)^k + (x_i - \tau_s)^k \tag{10}$$

If $\kappa=1$ and the Spline function is defined as

$$K(x_i, x_j) = 1 + \langle x_i, x_j \rangle + \frac{1}{2}\langle x_i, x_j \rangle \min(x_i, x_j) - \frac{1}{6}\langle x_i, x_j \rangle \min(x_i, x_j)^3 \tag{11}$$

Where the solution is a piecewise cubic.

2.3 General Considerations.

Bias Analysis

The inclusion of a bias within the kernel function generally leads to a more efficient implementation and a slightly better accuracy model. Conversely, the solution achieved with an implicit or explicit bias are not the same. This dichotomy emphasizes the difficulties with the interpretation of generalization in high dimensional feature spaces. In this work the explicit bias approach is used.

Free Parameters

Other important issues in support vector applications are the selection of free parameters such as the coefficient of *C*, the value of error ε it determine the margin within which error is neglected and in the Gaussian kernel function the value of variances σ[15-17].

The Quadratic Programming Problem

The SVM training works flawlessly for not too large data points. However, when the number of data points is large (e.g, over 2,000 data points), the Quadratic Programming (QP), problem becomes extremely difficult to solve with standard QP solvers and methods [13-17]. In the study case of this survey, the number of data points is 365, where each data point represents the daily average ofPMx concentration. Therefore the analysis and solving of the QP problem is not considered in the scope of this survey.

According to [11-13], the insensitive loss function is equal to these slack variables, where the ε-insensitive loss function is defined in equation 2, similarly the quadratic ε-insensitive loss function is defined by

$$L_\varepsilon(d, y(x)) = |d - y(x)|_\varepsilon^2 \tag{12}$$

Figures 2a y 2b, show the form of the linear and quadratic ε-insensitive loss functions. The ε-insensitive loss function is attractive because unlike the quadratic cost function, where all data points will be support vectors, then the solution can be sparse. The quadratic loss function produces a solution which is equivalent to ridge regression.

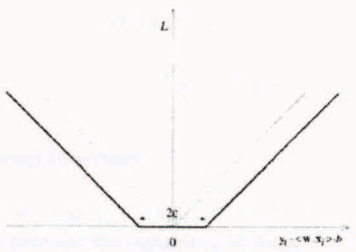


Fig. 2a. The linear ε-insensitive loss for zero and non-zero ε

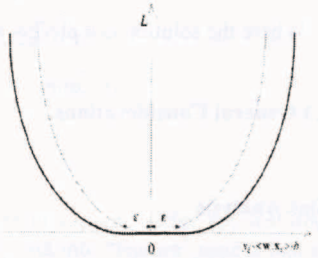
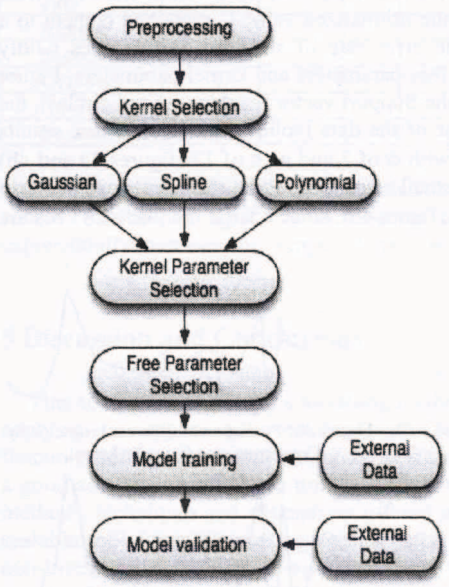


Fig. 2b. The quadratic ε-insensitive loss for zero and non-zero ε

Fig. 2. The ε-insensitive loss functions for zero and non-zero ε

3 Methodology

The proposed Methodology have been taken from [3-4], such this works provides the general steps to make pollutants modeling and predictions by using SVM working in regression mode. In this survey Gaussian, Polynomial and Spline kernel functions are used [3][4][13][14][15], a Gaussian distribution provides a natural representation of the system behavior [13][15]. The aim of this survey is to show the relations between kernel Gaussian, Polynomial and Spline kernels and the obtained SVM models. In order to perform an appropriate design, train, and testing of SVM this article describes a generic methodology based in a review of [3-4]. See Fig 3.



(a) Preprocessing of the input data by selecting the most relevant features, scaling the data in the range [−1, 1], and checking for possible outliers.

(b) Selecting an appropriate kernel function that determines the hypothesis space of the decision and regression function.

(c) Selecting the parameters of the kernel function, in polynomial kernels the degree for polynomials and the variances of the Gaussian kernels respectively.

(d) Choosing the penalty factor C and the desired accuracy by defining the ε-insensitive loss function.

(e) If required, solving the QP problem in l for classification problem and 2l variables in the case of regression problems.

(f) Validating the model obtained on some previously, during the training, unseen test data, and if not pleased iterate between steps (c) (or, eventually b) and (e).

Fig 3. Diagram of the proposed Model Solution

The fundamental reason for considering SVM working in regression mode as an approach for PMx modeling is the non-linear aspect of the application. There is no predetermined heuristic for the choice of free parameters and design for the SVM, many applications appear to be specific, in order to improve the SVM performance trough the automatic adjustment of free parameters. Using SVM on real time applications appear to be rather complex since of the computational demands of the deriving results.

4 Experimental Results

The Support Vector methodology can be applied in the case of regression, maintaining all the main features that characterize the system behavior. A Support Vector Machine in a kernel-induced feature space learns a non-linear function while the capability of the system is controlled by a set of parameters that do not depend on the dimensionality of the space. In this section, a set of results and simulations is presented. This is carried out by using the proposed regression SVM model approach with Gaussian, Polynomial and Spline kernel functions and standard nonlinear data sets of PMx. During 2009, simulations were carried out using the proposed SVM model. The σ values were modified to 1 and 2. Likewise, the ε values were modified to 7, 11 and 13. For every case study, the normalized value C remained content to a value of 100.Also is observed that the error rate of standard SVM varies wildly depending on different values of SVM free-parameters and kernel parameters. Figure 4 shows a summary of the results with the Support vector machine (in red circles), the raw data (black cross) and the behavior of the data (solid black line). These results show that the best results are obtained with σ of 2 and an ε of 13 (figures 4a and 4b) due to the small number of SVMs and small error rate, whilst the worst-case scenario is obtained with a σ of 1 and an ε of 13 (figure 4d), since a large number of SVMs are obtained.

Figure 4a: Prediction of PMx concentration in January using a Polynomial kernel.

Figure 4b: Prediction of PMx concentration in October using a Gaussian kernel.

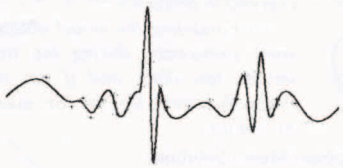


Figure 4c: Prediction of PMx concentration in January using Polynomial kernel.

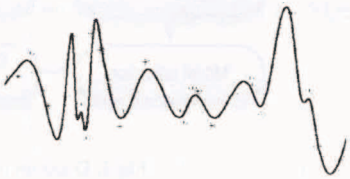


Figure 4d: Prediction of PMx concentration in October using Gaussian kernel.

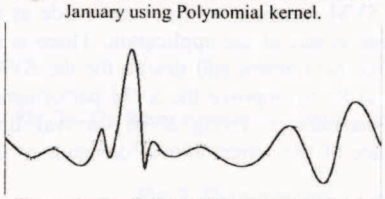


Figure 4e: Prediction of PMx concentration in January using Polynomial kernel.

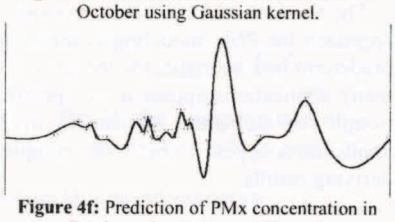


Figure 4f: Prediction of PMx concentration in October using Polynomial kernel.

Figure 4: Prediction of PMx concentration

Table 1. Summary Results

Samples 2007.	Polynomial kernel		Gaussian kernel		Spline kernel	
	No. of SVM	*Incorrect Forecasting*	*No. of SVM*	*Incorrect Forecasting*	*No. of SVM*	*Incorrect Forecasting*
January	17	3	26	1	17	3
February	6	5	18	3	15	5
March	12	4	28	1	24	3
April	12	4	25	0	22	2
May	8	3	23	2	19	3
June	13	1	20	1	15	2
July	18	1	26	1	17	0
August	11	2	24	1	20	3
September	16	1	25	2	20	4
October	18	0	28	1	23	3
November	17	1	24	2	21	5
December	15	3	26	2	24	6

From these results, it can be concluded that for this case study a σ of 1 gives a similar number of SVMs with respects to the number of data points. This exponentially increases the computational cost, making it unfeasible to calculate it.

5 Discussion and Conclusions

This survey has presented a modeling method of the daily atmospheric pollution by applying the support vector machine with Gaussian, Polynomial and Spline kernels functions working in regression mode. The application of SVM has enabled to obtain a good accuracy in modeling pollutant concentration of both PM10 and PM2.5. The methods, techniques and alternatives offered in the SVM field provides a flexible and scalable tool for implementing sophisticated solutions with implied dynamical and non-linear data. It is noteworthy to point that the SVM guarantees this global minimum solution and a good feature of generalization. Furthermore, implementing other kernel functions such as wavelet and hybrid functions may be implemented for future contributions.

Acknowledgement

The authors would like to thank the Air Quality Archive hosted by AEA Energy & Environment, on behalf of the United Kingdom Department for Environment, Food & Rural Affairs and the Devolved Administrations (DEFRA). Also, the authors would like to acknowledge the financial support of the Mexican government via SEP-PROMEP /103-5/09/4100 Project.

References

1. Lall R, Kendall M, Ito K, Thurston G. D.: Estimation of Historical Annual PM2.5 Exposures for Health Effects Assessment, Atmospheric Environment, Vol. 38, pp. 5217--5226. (2004)
2. Malcolm A.L, Derwent R.G, Maryon R.H.: Modelling the Long-range Transport of Secondary PM10 to the UK, Atmospheric Environment, Vol 34, pp. 881--894. (2000)
3. Vega, E. Reyes, G. Sanchez, et al.: Basic statistics of PM2.5 and PM10 in the atmosphere of Mexico City, The Science of the Total Environment, Vol. 287, pp. 167--176. (2002)
4. Querol X., Alastvey A., Ruiz C. R., et al.: Speciation and origin of PM10 and PM 2.5 in selected European cities, Atmospheric Environment, Vol. 38, pp. 6547--6555. (2004)
5. Harrison R.: Studies of the source apportionment of airborne particulate matter in the United Kingdom, Journal of Aerosol Science, Vol. 31, supplement 1, pp. S106--S107. (2000).
6. F. Wang, D.S. Chen, S.Y. Cheng, J.B. Li, M.J. Li, Z.H. Ren, "Identification of regional atmospheric PM10 transport pathways using HYSPLIT, MM5-CMAQ and synoptic pressure pattern analysis", Environmental Modelling& Software 25 , pp. 927-934 (2010).
7. Collazo-Cuevas J.I., Aceves-Fernandez M.A., Gorrostieta-Hurtado E., Pedraza-Ortega J.C., Sotomayor-Olmedo A.2, Delgado-Rosas M., "Comparison between Fuzzy C-means Clustering and Fuzzy Clustering Subtractive in urban air Pollution", CONIELECOMP 2010, 20th International Conference on Electrical Communications, pp. 174-179.
8. Aceves-Fernández M.A., Sotomayor-Olmedo A., Gorrostieta-Hurtado E., Pedraza-Ortega J.C., Tovar-Arriaga S., Ramos-Arreguin J.M., Performance Assessment of Fuzzy Clustering Models Applied to Urban Airborne Pollution, CONIELECOMP 2011, 21th International Conference on Electrical Communications, Accepted, In press (2011).
9. Lenschow P., Abraham H.J., Kutzner K., Lutz M., Preuß J.D., Reichenbächer, W.: Some ideas about the sources of PM10, Atmospheric Environment, Vol. 35, supplement 1, pp.S23--S33. (2001)
10. Vapnik, V.: The Nature of Statical Learning Theory. Springer-Verlang, New York. (1995).
11. Vapnik, V., Golowich, S., Smola A.: Support method for function approximation regression estimation, and signal processing. Advance in Neural Information Processing System 9.MIT Press, Cambridge, MA. (1997)
12. Schölkfopf B.: Smola A. J.: and Burges C.: Advances in Kernel Methods –Support Vector Learning. Cambridge, M.A.: MIT Press. (1999)
13. Osuna, E., R. Freund, F. Girosi.:Support vector machines: Training and applications. AI Memo 1602, Massachusetts Institute of Technology, Cambridge, MA 44. (1997)
14. Cristianini, N., Shawe-Taylor, J., An introduction to Support Vector Machines and other kernel-based learning methods, Cambridge University Press, Cambridge, UK (2000)
15. S. Osowski and K. Garanty, "Forecasting of the daily meteorological pollution using wavelets and support vector machine," Engineering Applications of Artificial Intelligence, vol. 20, no. 6, pp. 745-755, September (2007)
16. I. Sapankevych and R. Sankar, "Time series prediction using support vector machines: A survey," Computational Intelligence Magazine, IEEE, vol. 4, no. 2, pp. 24-38, (2009)
17. W. Lu and W. Wang, "Potential assessment of the support vector machine method in forecasting ambient air pollutant trends," Chemosphere, vol. 59, no. 5, pp. 693-701, April (2005).