Predicting Surface Ozone Concentrations using Support Vector Regression

Hossam Faris [1], Nazeeh Ghatasheh [2], Ali Rodan [1], Mua'ad Abu-Faraj [3]

[1.] King Abdullah II School for Information Technology, The University of Jordan, Amman, Jordan
[2.] Department of Business Information Technology, Faculty of Information Technology and Systems, The University of Jordan, Aqaba, Jordan
[3.] Department of Computer Information Systems, Faculty of Information Technology and Systems, The University of Jordan, Aqaba, Jordan
hossam.faris@ju.edu.jo

Abstract: Ozone layer diminution has been one of the major environmental problems so far. Such problem calls for a reliable monitoring mechanism to aid the strategic long-term remedy. However, it is challenging to develop a reliable prediction model due to the complexity of the relationships among the main attributes involved. Therefore, the causal attributes to the problem require an innovative modeling scheme. In this study we will investigate the application of support vector Regression (SVR) for predicting the surface Ozone concentrations. Several SVR models were developed using different kernel functions. The developed prediction models are based on limited number of input attributes which are atmospheric temperature, relative humidity and Nitrogen-dioxide. Apart from the complexity of the adopted approach, models are evaluated and compared using different measurement criteria.
[Faris H., Ghatasheh N., Rodan A., Abu-Faraj M. **Predicting Surface Ozone Concentrations using Support Vector Regression.** *Life Sci J* 2014;11(6):126-131]. (ISSN:1097-8135). http://www.lifesciencesite.com. 18

Keywords: Ozone; Genetic Programming; Neural networks; Support Vector Machines; Modeling; Prediction

1. Introduction

Ozone (O3) is one of the most important components of the Earth's atmosphere and the most well-known photochemical oxidants. This oxidant can be found in different layers of the Earth atmosphere (lower "troposphere" and the upper "stratosphere" layers). The "stratosphere" which is called the Ozone layer is found and resides at a range between 10 to 50 kilometers of altitude above the earth's surface.

Ozone layer acts as an ultraviolet light filter by absorbing more than 97 percent of the Sun's radiations that range between 200-315 nanometer of wavelength, where these radiations can affect and harm humans and the biological life in the earth. Therefore, preserving the thermal properties of the atmosphere makes the Ozone a crucial and vital issue [1]. Ozone becomes a threat when a number of complex chemical interactions occur between Nitrogen Oxides (NO_x) and Volatile Organic Compounds (VOC). These NO_x and VOC interactions turn the surface ozone into a photochemical pollutant [2]. On the other hand, Ozone is called non-methane hydrocarbons, in the presence of Ultraviolet Radiation (UV). It is important to mention that several emissions by human activities play a major role in producing the compounds that interact with the Ozone. Due to these reasons, a reliable prediction system for the concentrations of the Ozone and the inducing factors is highly required in order to trigger warnings before they reach critical levels. In previous literature, several approaches were investigated and proposed by researchers and scientists for predicting pollutants and Ozone levels including a wide range of machine learning and statistical approaches.

Moreover, climate change and air quality can also be affected by Ozone as mentioned in many previous studies [3,4]. This effect is considered to be the main cause for several health problems based on the exposure level [5], such health problems are related to the human cardiovascular or respiratory system as irritation, asthma, lung malfunctioning, and inflammations. Other health problems can include difficulty in breathing, eye irritation, cough and discomfort. Depending on the Ozone concentration level, exposure duration and frequency, the dangers vary from having chronic disease to affecting the mortality, or become a serious death factor. Moreover, an indirect threat to human life is caused by the Ozone effects on the plants. The exposure to Ozone causes several types of damage to the crops and trees including the interference with plant functions, growth, and the natural life cycle. Over 90 percent of the problems to the plants are claimed to be cause by the Ozone [6]. It is clear that Ozone is a major and serious pollutant that has significant effects on humans, animals, plants and also it may affects other natural or artificial resources.

One of the crucial Ozone issues to consider is the seasonal patterns which differentiate it from other pollutants. Ozone concentrations are highly correlated to the seasons, having the highest level in summer. For that reason, this seasonal pattern should

be considered in order to have an effective prediction model. In this work, only May to July time period is considered, further details regarding the study period are highlighted in the Area of study and data description section.

This study proposes the application of Support Vector Regression (SVR) models for predicting the surface Ozone concentrations. A number of SVR models are developed using different kernel functions. Parameters of the models are also tuned in order to maximize prediction accuracy. Moreover, a comparison of the SVR results against Multilayer Perceptron Neural Network with back-propagation learning is conducted and presented in the experimental results section.

2. Related Work

Machine Learning techniques have been applied in different domains including environmental fields to overcome complex and dynamic issues. Some of these techniques are inspired from nature and biological systems (i.e; Artificial Neural Network, Genetic Algorithms, and Particle Swarm Intelligence). Surface Ozone prediction is one of the issues that has been tackled by Machine Learning approaches. Artificial Neural Networks (ANNs) are one of the most Machine learning approaches used in this domain. ANNs have been proven to be effective and accurate models which led to better results compared to the statistical-based models (i.e. linear regression and Autoregressive Moving Average (ARMA)) [7–9, 13, 19]. In previous studies [1,3-4,10-11,12-14,19], ANN models were used to empirically predict the surface Ozone concentrations. In [1,4,10-11, 19], authors used Multilayer Perceptron Neural Networks (MLP-NN) with back-propagation for training, while in [3, 12–14] authors applied several hybrid approaches for training different types of ANN. Despite the superior performance of ANN based approaches over the statistical ones, they are prone to a number of drawbacks like performance unpredictability and inconsistency. Moreover, Genetic Programming (GP) which is another bio-inspired modeling approach was used in [18] to predict the Particulate Matters (PM) pollutant relying on meteorological variables. Based on the results, authors in [18] show that GP has promising performance compared to other approaches tackling same problem.

Furthermore, Fuzzy based approaches were also used for predicting surface Ozone levels. A fuzzy-based rule generation was used in [16] to predict Ozone levels over the upper part of Austria. The fuzzy models revealed that the NO_x and VOC have high relation to the concentrations of the surface Ozone.

Authors in [15, 17] used Support Vector Regression (SVR) based models to predict the Ozone pollution. For instance in [17] author used both ANN and SVR with Radial Basis Function (RBF) kernel to develop a model for predicting surface Ozone concentration without any parameter optimization technique. However, SVR parameters and the kernel function used in the model selection can significantly affect the model performance and its accuracy.

3. Support Vector Regression

The ideas of Support Vector Regression (SVR) were developed by Vladimir Vapnik and his co-workers at AT&T Bell Laboratories. The basic idea of SVR is to map the training data into higher dimensional space using a nonlinear mapping function and then perform linear regression in that space [22,23]. SVR has many advantages over other traditional classification and prediction techniques such as that the solution of the problem relies on a small subset of the dataset which gives SVR a great computational advantage. SVR seeks the minimization of the upper bound of generalization error instead of minimizing the training error. Figure.1 shows the optimal hyperplane in SVR that separate two different datasets, where the vectors near the hyperplane are called support vectors.

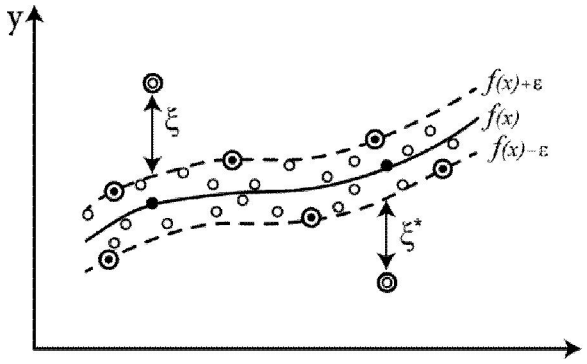

Figure 1. Optimal hyperplane in SVR

Suppose we have a data set $\{x_i, y_i\}_{i=1,..,n}$ where the input vector $x_i \in \Re^d$ and the actual $y_i \in \Re$. The modeling objective of SVR is to find the linear decision function represented in the following equation:

$$f(x) = \langle \mathrm{w}, \emptyset_{\mathrm{i}}(\mathrm{x}) \rangle + b$$

where w and b are the weight vector and a constant respectively, which have to be estimated from the data set. $\emptyset$ is a non linear mapping function

This regression problem can be formulated as equation 1 to minimize the following regularized risk function.

$$R(C) = \frac{C}{n}\sum_{i=1}^{n} L_\varepsilon(f(x_i), y_i) + \frac{1}{2}\|\mathrm{w}\|^2 \qquad (1)$$

where $L_\varepsilon(f(x)-y)$ is known as ε-intensive loss function and given by equation 2.

$$\begin{cases} |f(x)-y|-\varepsilon, & |f(x)-y| \geq \varepsilon \\ 0 \quad , otherwise \end{cases} \quad (2)$$

Introducing the slack variables ξ_i and ξ_i^* makes the problem in the following constrained form; Minimize equation 3.

$$R(w,\xi_i^*) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}(\xi_i+\xi_i^*) \quad (3)$$

Subject to:

$$\begin{cases} y_i - \langle w, x_i\rangle - b & \leq \varepsilon + \xi_i \\ \langle w, x_i\rangle + b - y_i & \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* & \geq 0 \end{cases} \quad (4)$$

C is a regularized constant greater than 0 to balance between the training error and the model flatness. C represents a penalty for a prediction error that is greater than ε. ξ_i and ξ_i^* are slack variables form the distance from actual values to the corresponding boundary values of ε. The objective of SVR is to minimize ξ_i, ξ_i^* and w^2.

The above optimization with constrain can be converted by means of Lagrangian multipliers to a quadratic programming problem. Therefore, the form of the solution can be given by equation (5)

$$f(x) = \sum_{i=1}^{n}(\alpha_i - \alpha_i^*)K(x_i, x) + b \quad (5)$$

where α_i and α_i^* are Lagrange multipliers, which are subject to the following constraints:

$$\sum_{i=1}^{n}(\alpha_i - \alpha_i^*) = 0 \quad (6)$$

$$0 \leq \alpha_i \leq C \qquad i = 1, \dots, n$$
$$0 \leq \alpha_i^* \leq C \qquad i = 1, \dots, n$$

$K(.)$ is the kernel function and its values is an inner product of two vectors x_i and x_j in the feature space $\emptyset(x_i)$ and $\emptyset(x_j)$ and satisfies the Mercer's condition. Therefore in equation 6,

$$K(x_i, x_j) = \emptyset(x_i).\emptyset(x_j) \quad (7)$$

Kernel functions have a great impact on the performance of SVR. In this work we use the following common four kernel functions which can be given in the following equations:

Linear kernel:

$$K(x_i, x_j) = x_i^T x_j \quad (8)$$

Polynomial kernel:

$$K(x_i, x_j) = (\gamma x_i^T x_j + r)^d, \ \gamma > 0 \quad (9)$$

Radial basis function (RBF) kernel:

$$K(x_i, x_j) = \exp(-\gamma * \|x_i - x_j\|^2), \ \gamma > 0 \quad (10)$$

Sigmoid kernel:

$$K(x_i, x_j) = \tanh(\gamma x_i^T x_j + r) \quad (11)$$

The dual form of the non linear SVR can be given in the following form

$$\emptyset(\alpha_i - \alpha_i^*) = \sum_{i=i}^{n} d_i(\alpha_i - \alpha_i^*) - \varepsilon\sum_{i=i}^{n}(\alpha_i + \alpha_i^*) - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)K(\alpha_i - \alpha_j) \quad (12)$$

4. Area of Study and Data Description

Data used in the work were collected and measure for Chenbagaramanputhur area by R. Samuel Selvaraj and others in [20]. The area is located in the countryside in Kanyaku- maridistrict; it is about 12 kilometers from Nagercoil town. The operating temperature range is from 5° C to 50°C, relative humidity limits are 5% and 95%. Similar kind of NO_2 sensor has been used for nitrogen dioxide measurement, the gas sensitive semiconductor (GSS) type sensor is described in (www.aeroqual.com). The sampling was carried out for three months from May 2009 to July 2009. For Ozone, seven readings were taken per day starting from 530h to 2330h with three-hour interval. For NO_2, only two readings were taken one at daytime and the other at night time. Furthermore, the data set, used in this work, carried out for three months from May 2009 to July 2009. For the ozone, seven readings were taken per with three-hour interval. For the NO_2, only two readings were taken each day, one at daytime and the other at night time. The model inputs and Output are given in Table I. The forecasting horizon of the model is three-hours ahead. Figure 2 shows the collected measurements to develop the model as given in [20].

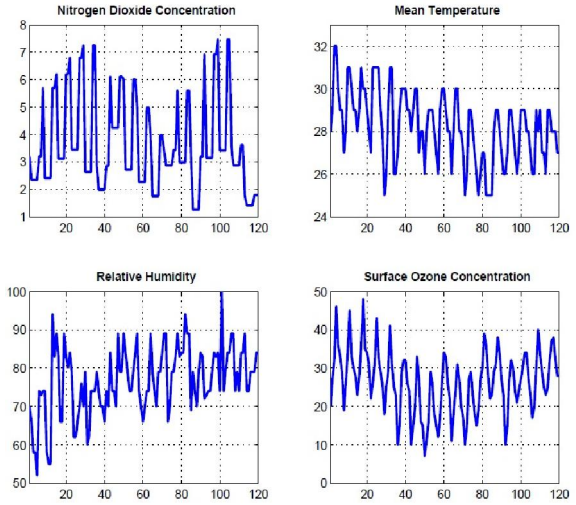

Figure 2. Training and Testing data set [20]

5. Evaluation Criteria

The performance of the developed SVR model is assessed using Root Mean Squares Error (RMSE) and Mean Absolute Error (MAE) to measure how close the measured values are to the predicted values. RMSE and MAE are computed as follows:

1) Root Mean Squares Error (RMSE):

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y - \hat{y})^2} \qquad (13)$$

2) Mean Absolute Error (MAE):

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y - \hat{y}|^2 \qquad (14)$$

where y and $\hat{y}$ are the actual and the estimated Ozone measurements based on proposed models and n is the number of measurements used in the experiments.

6. Experimental Results

In our experiments, four SVR models are developed using four different kernels; linear, polynomial; RBF, and Sigmoid. We used LIBSVM which is a library for Support Vector Machine [21]. 90 samples of the data were used for training the models while the rest were used for testing.

For SVR, there are number of parameters which should be tuned in order to maximize the accuracy of the prediction models. These parameters include the cost C and the parameters γ and d of the kernels. Grid search was used to find the best values for C and γ in Sigmoid and RBF kernels. The best obtained results for training and testing cases for each SVR model are shown in Table 1 along with the best parameters' values. SVR with Sigmoid kernel showed the best results for training and testing cases. Actual against predicted values obtained by the latter model for training and testing cases are shown in Figure 3.

Moreover, in Table 1, SVR prediction results are compared with those obtained previously for the same dataset using a Multilayer Perceptron Neural Network (MLP-NN) in [17]. SVR outperformed MLP-NN in means of RMSE and MAE performance measurements.

Although SVR achieved good prediction results compared to MLP-NN, there is still considerable level of error which can be attributed to the limited number of input variables and the simple instruments used for measuring those variables.

Table 1. RMSE and MAE values for SVR with different kernel functions compared with MLP-NN values from [17].

	SVR *Linear kernel*		**SVR** *Polynomial kernel* γ=0.1 d=2		**SVR** *Sigmoid kernel* γ =0.3 C=1		**SVR** *RBF kernel* γ =0.12 C=1		**MLP-NN** epochs=2000 $\alpha = 0.2$ Momentum=0.3 Neurons in hidden layer=2	
	Training	Testing	Training	Testing	Training	Testing	Training	Testing	Training	Testing
RMSE	7.023	5.559	8.730	6.037	7.331	**5.431**	7.379	5.495	6.836	7.155
MAE	5.413	4.359	6.905	4.833	5.667	**4.1843**	5.716	4.241	5.415	5.817

*Bold means best results obtained.

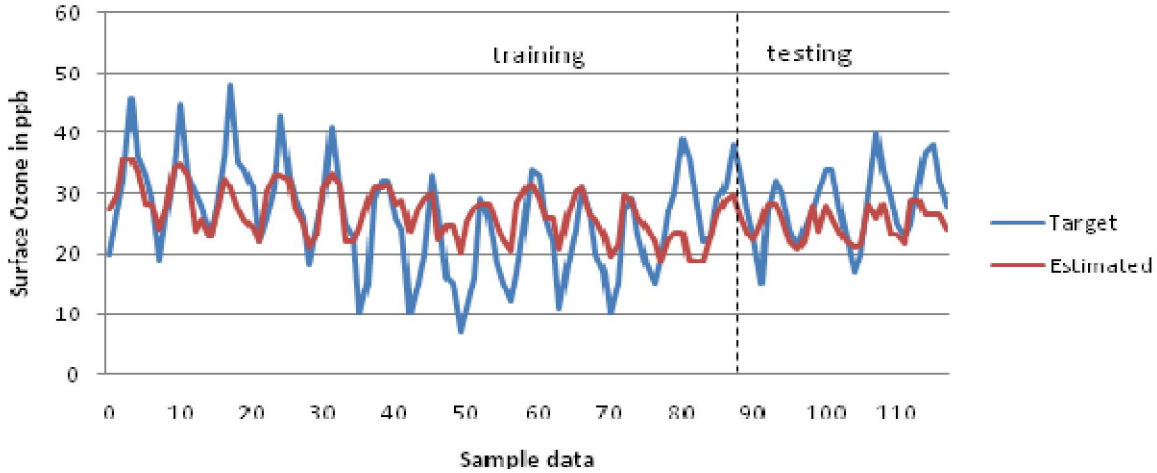

Figure 3. Actual and Estimated measurements for the surface Ozone concentration using SVR with sigmoid kernel for training and testing cases

7. Conclusion

In this work we investigated the application of Support Vector Regression (SVR) in predicting surface Ozone concentrations based on limited number of measured attributes which are atmospheric temperature, relative humidity and Nitrogen-dioxide. Several SVR models were developed using different kernel functions. SVR models were evaluated and compared with results obtained previously by Multilayer Perceptron Neural Network (MLP-NN) [17]. SVR model with Sigmoid kernel function showed higher predictive power and outperformed MLP-NN model results.

Corresponding Author:
Dr. Hossam Faris
King Abdullah II School for Information Technology
The University of Jordan
Amman, Jordan
E-mail: hossam.faris@ju.edu.jo

References

1. S. Stephen Rajkumar Inbanathan, O. Mahendran, R. Samuel Selvaraj and R. Jayalakshmi, "Modelling of surface ozone using artificial neural network in an urban area,"International Journal of Engineering Science and Technology, vol. 3, no. 2, pp. 1173-1177, 2011.
2. S. Chattopadhyay and G. Bandyopadhyay, "Artificial neural network with backpropagation learning to predict mean monthly total ozone inarosa, switzerland,"International Journal of Remote Sensing, vol. 28, no. 20, pp. 4471–4482, 2007.
3. P. H´ajek and V. Olej, "Ozone prediction on the basis of neural networks, support vector regression and methods with uncertainty," Ecological Informatics, vol. 12, no. 0, pp. 31–42, 2012.
4. A. Elkamel, S. Abdul-Wahab, W. Bouhamra, and E. Alper, "Measurement and prediction of ozone levels around a heavily industrialized area: a neural network approach,"Advances in Environmental Research, vol. 5, no. 1, pp. 47–59, 2001.
5. WHO, "Health Aspects of Air Pollution with Particulate Matter, Ozone and Nitrogen Dioxide," tech. rep., WHO, 2003.
6. B. S. Felzer, T. Cronin, J. M. Reilly, J. M. Melillo, and X. Wang, "Impacts of ozone on trees and crops," Comptes Rendus Geoscience, vol. 339, no. 11-12, pp. 784–798, 2007.
7. O. Pastor-B´arcenas, E. Soria-Olivas, J. D. Mart'ın-Guerrero, G. Camps-Valls, J. L. Carrasco-Rodr'ıguez and S. del Valle-Tasc´on, "Unbiased sensitivity analysis and pruning techniques in neural networks for surface ozone modelling," Ecological Modelling, vol. 182, no. 2, pp. 149–158, 2005.
8. E. Agirre, A. Anta, and L. J. R. Barron, "Forecasting ozone levels using artificial neural networks," Forecasting Models, pp. 208–218, 2010.
9. V. R. Prybutok, J. Yi and D. Mitchell, "Comparison of neural network models with ARIMA and regression models for prediction of houston's daily maximum ozone concentrations," European Journal of Operational Research, vol. 122, no. 1, pp. 31–40, 2000.
10. J. Yi and V. R. Prybutok, "A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area," Environmental Pollution, vol. 92, iss. 3, pp. 349-357, 1996.
11. S. Abdul-Wahab and S. Al-Alawi, "Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks," Environmental Modelling & Software, vol. 17, no. 3, pp. 219–228, 2002.
12. W. Wang, W. Lu, X. Wang, and A. Y. Leung, "Prediction of maximum daily ozone level using combined neural network and statistical characteristics," Environment International, vol. 29, no. 5, pp. 555–562, 2003.
13. G. Spellman, "An application of artificial neural networks to the prediction of surface ozone concentrations in the united kingdom," Applied Geography, vol. 19, no. 2, pp. 123–136, 1999.
14. A. Coman, A. Ionescu and Y. Candau, "Hourly ozone prediction for a 24-h horizon using neural networks," Environmental Modelling & Software, vol. 23, no. 12, pp. 1407–1421, 2008.
15. S. Canu and A. Rakotomamonjy, "Ozone peak and pollution forecasting using support vectors", IFAC workshop on environmental modelling. International Federation of Automatic Control (IFAC): Yokohama, 2001.
16. M. Ryoke, Y. Nakamori and C. Heyes, "A simplified ozone model based on fuzzy rules generation," European Journal of Operations Research, vol. 122, pp. 00–07, 2000.
17. M. Alkasassbeh,. "Predicting of Surface Ozone Using Artificial Neural Networks and Support Vector Machines," International Journal of Advanced Science & Technology, vol. 55, pp. 1-11, 2013.
18. H. Faris, M. Alkasassbeh, N. Ghatasheh and O. Harfoushi, "PM10 prediction using Genetic

Programming: A Case Study in Salt, Jordan," Life Sci. J., vol. 11, no. 2, pp. 86-92, 2014.
19. J. Gómez-Sanchis, J. D. Martín-Guerrero, E. Soria-Olivas, J. Vila-Francés, J. L. Carrasco and S. Valle-Tascón, "Neural networks for analysing the relevance of input variables in the prediction of tropospheric ozone concentration," Atmospheric Environment, vol. 40, iss. 32, pp. 6173-6180, 2006.
20. R. S. Selvaraj, K. Elampari, R. GAYATHRI and S. J. JEYAKUMAR, "A neural network model for short term prediction of surface ozone at tropical city," International Journal of Engineering Science and Technology, vol. 2, no. 10, pp. 5306–5312, 2010.
21. C.-C. Chang and C.-J. Lin. "Libsvm: A library for support vector machines," ACM Trans. Intell. Syst. Technol., vol. 2, no. 3, pp. 27:1{27:27, May 2011. [Online]. Available: http://doi.acm.org/10.1145/1961189.1961199
22. V. Vapnik. The Nature of Statistical Learning Theory. Springer, New York, 1995.
23. V. Vapnik. "An overview of statistical learning theory." IEEE Transactions on Neural Networks, vol. 10, no. 5, pp. 988–999, 1999.

3/23/2014