# Prediction of Missing Data for Ozone Concentrations Using Support Vector Machines and Radial Basis Neural Networks

Biljana Mileva-Boshkoska and Mile Stankovski
Department of Automatics, Faculty for Electrotechnics and Information Technologies,
Karposh 2, bb, 1000 Skopje, Republic of Macedonia
E-mail: biljanamb|milestk@feit.ukim.edu.mk

**Keywords:** air modelling, support vector machines, neural networks

**Received:** July 6, 2007

*In this paper we present results from prediction of data for ozone ($O_3$) concentrations in ambient air by using the modelling techniques of support vector machines (SVM) and radial basis neural networks (RBF NN). The predictions are performed for two short periods of time: for 24 hours and for one week in August and in December in 2005, in Skopje, Macedonia. The built SVM models use different kinds of kernels: polynomial and Gaussian kernels and the best values of the free parameters of SVM kernels are chosen by examining a range of values for each of the free parameters. This is the first attempt in Macedonia for prediction of concentrations of any air parameters in the ambient air.*

*Povzetek: Podana je analiza ravni ozona v Makedoniji z metodami strojnega učenja.*

## 1 Introduction

In the process of EU integration, Republic of Macedonia had to harmonize environmental legislation with European one. According to the new Macedonian legislation for air quality (Law on ambient air quality, Official Gazette of Republic of Macedonia, no 67/2004) the country is obliged to perform continuous monitoring of the ambient air throughout the whole territory of the country. For that reason, in Macedonia were installed fifteen automatic monitoring stations for gathering data for the air quality. However, mainly due to financial reasons, and technical problems in the maintenance of the monitoring stations, the data sets from the monitoring stations are not complete. According to the EU directives and Macedonian legislation, the country must fulfill 90% of the yearly measurements for the air quality on the measuring spots during one year. In order to fulfill the gaps in the data sets for air quality, we decided to use appropriate mathematical modeling technique, as a method that is allowed to be used by the EU directives.

In this paper we present the results obtained from filling in the existing gaps of the measured hourly data for the levels of ozone ($O_3$) in the ambient air for two short periods of time in the municipality of Karposh III, in Skopje, Republic of Macedonia. We process the two data sets for August and December, 2005 and we build statistical models for hourly predictions of concentrations for one day and for one week. Solution of the problem had to be generated in a simple manner and the used algorithm had to be applicable for similar problems e.g for prediction of concentrations of other air quality parameters.

One approach for prediction of hourly values is using neural networks for evaluating air parameters concentrations [1], [2], [3], [4]. SVM is another method that started in the late seventies [5], [6] and today is used for ambient air parameters prediction [7], [8], [9] and for time series forecasting [10] in the environmental applications.

For prediction of the $O_3$ levels, we use the modelling techniques based on SVM and Radial Basis Function (RBF) NN.

## 2 Overview of the whole process

Prediction of levels of O3 in ambient air is a complex process that consists of the following phases (Figure 1):

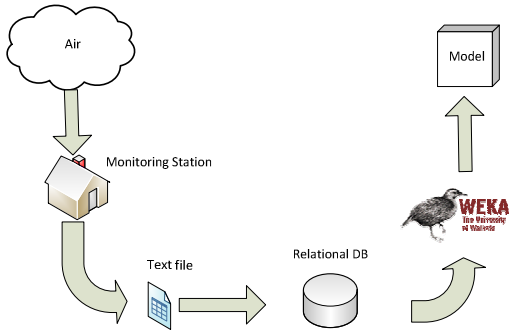


**Figure 1** Overview of the whole process of prediction

- Measurement of the levels of parameters of the ambient air by automatic monitoring station.
- Transmission of the measured data via radio connection from the monitoring station to the textual data base situated in the Ministry for Environment and Physical Planning (MoEPP).
- Data processing and preparation of ARFF files that are recognized by the WEKA software.
- Electing tools (software) for modelling the data.
- Modelling using the software package WEKA.
- Comparison of the received models and choosing the one that gives the best prediction results.

# 3 Used techniques

## 3.1 Support Vector Regression

The concept of a maximum margin hyperplane only applies to classification. However, support vector machine algorithms have been developed for numeric prediction that share many of the properties encountered in the classification case: they produce a model that can usually be expressed in terms of a few support vectors and can be applied to non-linear problems using kernel functions.

Similar with linear regression, the basic idea here is to find a function that approximates the training points well by minimizing the prediction error. The crucial difference is that all deviations up to a user-specified parameter $\mathbf{x_i}$ are simply discarded. Also, when minimizing the error, the risk of overfitting is reduced by simultaneously trying to maximize the flatness of the function. Another difference is that what is minimized is normally the predictions' absolute error instead of the squared error used in linear regression. A user-specified parameter $\mathbf{x_i}$ defines a tube around the regression function in which errors are ignored.

SVM approximate the learning data set with a function given in a form of:

$$\mathbf{f(x)} = \sum_{\mathbf{i=1}}^{\mathbf{l}} \mathbf{w_i \phi_i(x)} + \mathbf{b} \tag{1}$$

meaning that the original data $\mathbf{x} \rightarrow \boldsymbol{\phi}(\mathbf{x})$ are mapped into high dimensional space and then construct an optimal hyperplane in this space. $\boldsymbol{\phi}(\mathbf{x})$ represents feature of the inputs, while $\mathbf{w_i}$ and b are coefficients. These are estimated by minimizing the risk function [10]:

$$\mathbf{R(f)} = \int \mathbf{c(x, y, f(x))dp(x, y)} \tag{2}$$

where c(x,y,f(x)) is cost function that determines how to penalize estimation errors based on the empirical data X [7]. Given that we do not know the probability measure dp(x,y) we can only use X for estimating a function f that minimizes R[f]. A possible approximation consists in replacing the integration by the empirical estimate to get so called empirical risk function

$$\mathbf{R_{emp}[f]} = \frac{\mathbf{1}}{\mathbf{l}} \sum_{\mathbf{i=1}}^{\mathbf{l}} \mathbf{c(x_i; y_i; f(x_i))} \tag{3}$$

A first attempt would be to find the function $\mathbf{f_0} = \mathbf{argmin_{f \in H} R_{emp}[H]}$) for some hypothesis class H. However if H is very rich, i.e. its capacity is very high as for instance when dealing with few data in very high dimensional spaces, this may be not such a good idea as it will lead to overfitting and thus bad generalization properties. Hence one should add a capacity control term, which in the SV case results to be $\left|\left|\mathbf{w}\right|\right|^{\mathbf{2}}$, which leads to regularized risk function

$$\mathbf{R_{reg}} = \mathbf{R_{emp}[f]} + \frac{\boldsymbol{\lambda}}{\mathbf{2}} \mathbf{\|w\|^2} \tag{4}$$

## 3.2 Kernels

A kernel is essentially a similarity function with certain mathematical properties, and it is possible to define kernel functions over all sorts of structures-for example, sets, strings, trees, and probability distributions.

The choice of kernel $\mathbf{K(x_i, x_j)}$ influences drastically on the performance of the SVMs depending on the problem considered. Several kernels are available for learning sand they have to satisfy the so-called Mercer's condition [9].

The most commonly used kernels are the Gaussian kernel

$$\mathbf{K(x_i, x_j)} = \mathbf{exp}^{-\left(\frac{\left\|\mathbf{x_i - x_j}\right\|^{\mathbf{2}}}{\mathbf{2\sigma^2}}\right)} \tag{5}$$

and the polynomial kernel

$$\mathbf{K(x_i, x_j)} = \mathbf{(x_i x_j + 1)^p} \tag{6}$$

which are also used for the purposes of this research.

# 4 Data processing

The data sets that are used are gathered by the national automatic monitoring network (AMN) by the MoEPP in Republic of Macedonia. As soon as the data are transferred to the central DB in MoEPP they are first validated, that is the missing and the unreal data are marked with -9999. We have picked a small period of time where we do not have missing data, that is the period 1-17 August and 1-17 December 2005. We used two different seasons because we wanted to show the difference of the predicted results from different models depending on the standard deviation of the input data.

The first phase is parsing of data and their storage in a relational data base. We convert the validated data into ARFF format that is recognized by the WEKA software that is used for the process of prediction of the $O_3$ levels.
In order to build models for prediction of $O_3$ levels, as input parameters we use the hourly data for the levels of $NO_2$, $O_3$, temperature and humidity for 10 days in a raw. The output function is following:

$$\boldsymbol{O_3(t) = f(NO_2(t-z) + O_3(t-z) + NO_2(t) + temp(t-z) + hum(t-z))} \quad (1)$$

We built eight different models for prediction of $O_3$ levels for $t - z$ hours, where $z = 1,2, \ldots, 8$.

For prediction of $O_3$ levels, first we build three types of models from which two are based on SVM, while the third one is based on RBF NN. In order to build the first two models, we use the existing functions in WEKA: SVMreg with polynomial kernel, where p=1 and SVMreg with RBF kernel, known as SVM with Gaussian kernel. For building the third model we use the function RBF with neural network which is also implemented in Weka. The three functions are used both for prediction of levels of $O_3$ for 24 hours and for one week. That way we get two groups of models. In the first group belong models for prediction of levels for 24 hours and in the second group belong models for prediction of levels for one week.
In order to decrease the total processing time for training the SVM we used the tool Explorer from WEKA that enabled us to distribute the whole process of learning of the model on three computers controlled by one "master" computer.

The results from the obtained models are compared. As a measure for deviation of the predicted results from the measured one we use the mean absolute error given by:

$$\boldsymbol{MAE = \frac{1}{n}\sum_{i=1}^{n} |\alpha_i - p_i|} \quad (2)$$

# 5 Results from modelling

When modelling with SVM, first we choose the best values of the free parameters of the kernels: $C$ (factor of penalty, Figure 2), $\varepsilon$ (Figure 3) and $\gamma$,which is connected to the speed parameter $\sigma$ with the relation of $\gamma = \sqrt{\left(\frac{1}{\sigma^2}\right)}$ (Figure 4). To choose the values of the free parameters is the main difficulty when modelling with SVM. Taking into consideration that there are no general rules on determination of the values of the free parameters, it is necessary to determine the influence of the chosen value of the free parameter on the resulting error on the predicted results from the model. In this paper we use MAE for assessment of the deviation between the original measured data and the predicted data. In general, the smaller MAE, the better results the built models achieve.

Figure 2 presents the variations of MAE from the parameter $C$. The graph shows that the parameter $C$ has very small influence on MAE and it is sensitive only on very small values for $C$, for example when $C \leq 0.001$. When increasing the values of $C$, the value of MAE steeply decreases until $C$ recives values $C \geq 0.5$ when again parameter $C$ makes very small influence on MAE. In general, in order one to guarantee a stable learning process, the value of the parameter C has to receive large values, for example C=100, as it is the case in this paper.

Figure 3 presents the variations of MAE from the values of the parameter $\varepsilon$. Parameter $\varepsilon$, like parameter C has small influence on the performances of the model for prediction of the ozone concentrations. The values of MAE are almost constant for values of the parameter $\varepsilon < 10^{-2}$ and $\varepsilon > 0.5$. In the models in which we use SVM the value of $\varepsilon$ is should be small. In this research, we set the value of $\varepsilon$ to 0.1.

Theoretically, the value of the speed parameter $\sigma$ influences a lot on the prediction performances of the model. Very small ($\sigma \to 0$) or very large ($\sigma \to \infty$) values of $\sigma$ may lead to bad prediction results. If $\sigma \to 0$, all training data become support vectors. In that case, when an unknown data occur as input at the SVM model, the SVM model will not be able to provide good predication results. From the other side, if $\sigma \to \infty$, all training data will be considered as one point and the SVM model may produce same results for any new input data to the model. Therefore, these two extreme cases should be avoided. We should note that both $\sigma \to \infty$ and $\sigma \to 0$ represent two approximate processes. In real applications, if $\sigma \ll \|x_i - x_j\|$ and $\sigma \gg \|x_i - x_j\|$ the extreme cases mentioned above will occur. Figure 4 presents the variations of MAE from the values of the parameter $\sigma$. Results in the **Error! Reference source not found.** show that MAE is large, when $\sigma$ is small (for example $\sigma = 0.001$), than it decreases with increasing of $\sigma$ and it reaches minimum for values of $\sigma$ around 1. Figure 4 shows that MAE fluctuates when $\gamma$ is in the range of [0.9, 1.1]; then it increases with increasing of $\gamma$, and finally it has tendency to become constant after $\gamma$ reaches values $\gamma \geq 100$. For that reason, in practical applications only parameter $\gamma$ (or $\sigma$) of the Gaussian kernel function has to be determined, while the two parameters $C$ and $\varepsilon$ may be set in advance by experience. In this application we set the value of $\gamma$ to 0.5.

Once the best values for the free parameters C, $\gamma$ (or σ) and ε are determined, the final step is to produce the models for prediction of the missing data for $O_3$.

In this paper, we have calculated the best values for parameters C, $\varepsilon$ and $\gamma$ for z=3. We have used the same values later in order to predict results for $z \neq 3$ ($z = 1,..,8$) in which cases the free parameters are not optimal. Although in those cases we do not use the optimal values for the free parameters, still the results obtained from models built with SVM are better than those from the models built with RBF NN.

Figures 5 – 8 show the results from the modelling. Each figure show the distribution of the original data and the distribution of predicted data obtained from three different models built with polynomial and Gaussian kernel and with RBF NN.

In August 2005, the data of the levels of $O_3$ are very close to each other i.e. the standard deviation is very small. Therefore the three models give similar results for prediction of $O_3$ levels (Figure 5 and Figure 6).

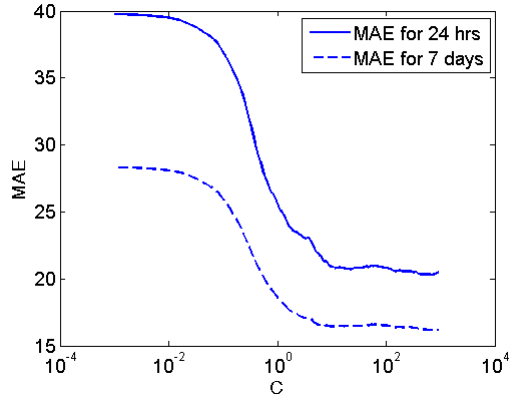


**Figure 2:** Variations of MAE from the parameter C for prediction of $O_3$ levels for 24h and for 7 days for August, 2005

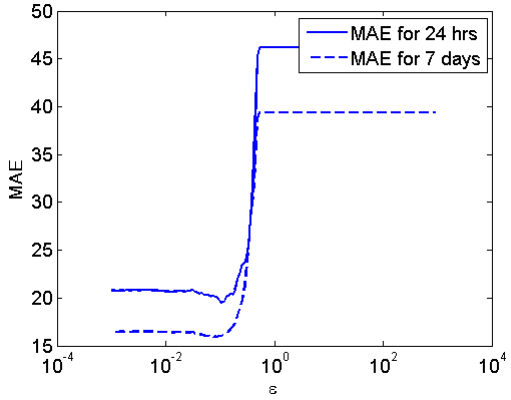


**Figure 3:** Variations of MAE from the parameter **ε** for prediction of $O_3$ levels for 24h and for 7 days for August, 2005

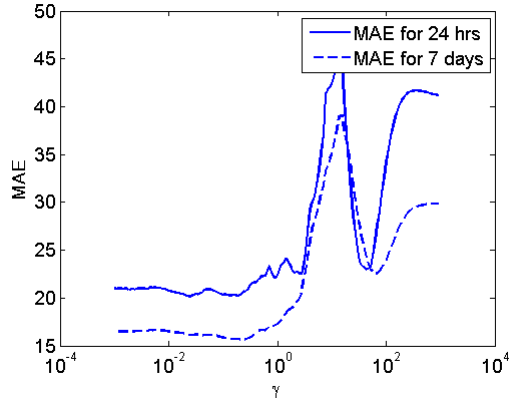


**Figure 4:** Variations of MAE from the parameter **γ** for prediction of $O_3$ levels for 24h and for 7 days for August, 2005

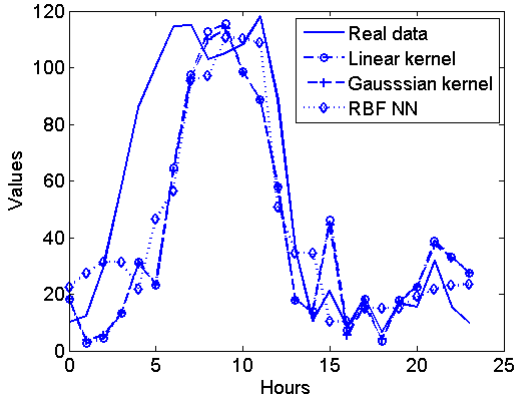


**Figure 5:** Prediction of levels of $O_3$ for 24 hours, for August 2005, for z=3

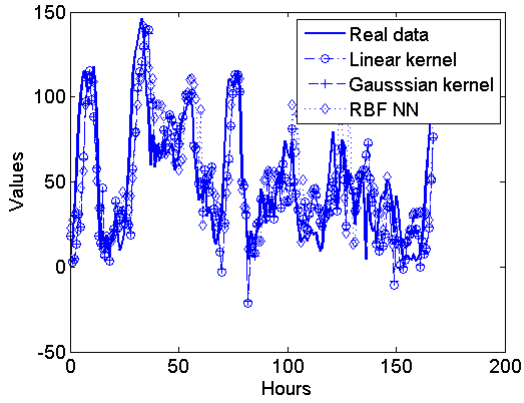


**Figure 6:** Prediction of levels of $O_3$ for one week, for August 2005, for z=3

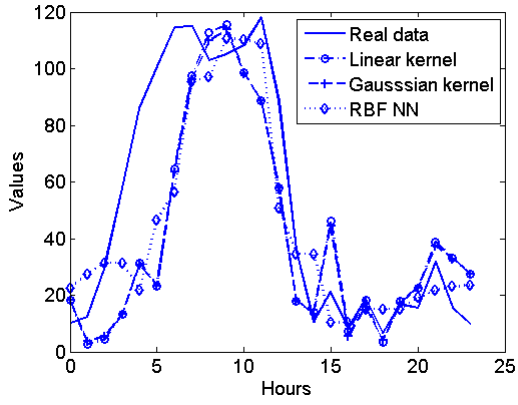


**Figure 7:** Prediction of levels of $O_3$ for 24 hours, for December 2005, for z=3

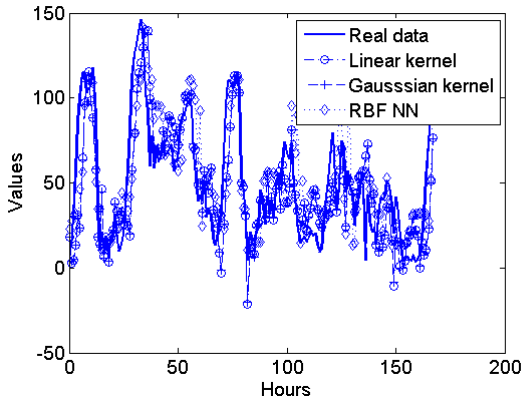


**Figure 8:** Prediction of levels of $O_3$ for one week, for December 2005, for z=3

The input data in December 2005 have big standard deviation. In this case due to the good generalization characteristics of the SVM models, the best prediction results for period of one week are achieved by the model built with SVM with polynomial kernel z = 1, 2, 3 and by SVM with Gaussian kernel for z = 4,..,8. The best results for prediction of 24 hours are achieved by model built with SVM with polynomial kernel for z = 1, 2, 5, 6, 7 and 8 and by SVM with Gaussian kernel for z = 3 and 4.

Figure 5 and 6 show the distribution of original $O_3$ data for the eleventh day and for one week in August, 2005. The same figures show the distribution of the predicted data that are obtained by the three models.

In August, when predicting the $O_3$ levels for one week, the best results are achieved by the model built with SVM with polynomial kernel for $z = 1, 3, 4, 6, 7$ and 8 and by SVM with Gaussian kernel for z = 2 and 5. The best results for prediction of 24 hours, the best results are achieved when using the model built with SVM with polynomial kernel for z = 1, 3, 4, 6 and 7 by SVM with Gaussian kernel for z = 2, 5 and 8.

# 6 Conclusion

The paper describes an attempt to predict the of hourly missing data for $O_3$ concentrations in the ambient air using SVM and RBF NN at the municipality Karposh III, in Skopje, Macedonia.

We developed a complete system for filling the gaps of missing hourly data by predicting the levels of $O_3$.

The built models for prediction of concentrations of ozone are examined for prediction of 8 consequent hourly values. The best results are achieved by the model built with SVM with polynomial kernel for prediction of 24 hours for December and August, 2005. In one case, the best results were achieved by the model built with SVM with Gaussian kernel, for prediction of one week for December, 2005. We should conclude that models built with SVM achieve better results than models built with RBF NN.

Finally we may conclude that SVM models give better results when predicting time series and they offer several advantages before the conventional RBF NN. In this paper we examined the free parameters of Gaussian kernel $C, \epsilon$ and $\sigma$ and we conclude that only parameter $\sigma$ has significant influence on the results from the offered models. Unlike the SVM models, the conventional RBF models parameters like the size of the network, the learning parameter and the training of the network play big role in the performances of the built model. Further on, are a result of the Structural Risk Minimization Principle, models built with SVM provide better prediction results compared with SVM models. Finally, using SVM we overcome the problems of neural networks like overfitting and local minima.

Although it is not possible to use the exact same models to predict the concentrations on the other measurement places in the country, still the presented methodology is general and it may be used for building new models for the other measurement places. The new models will be trained with data measured at the local measurement sites.

Models for prediction of ozone concentrations may be further extended. The developed model for ozone prediction uses data for $NO_2$, $O_3$, temperature and humidity. It may be extended with additional data for NOx, data for emissions from vehicles and other known sources of ozone. Similarly, the models may be extended with additional meteorological parameters.

The developed models are based on real data. In future, the presented methodology could be used for development of models that will take into consideration emissions from large combustion plants or the complexity of terrain where the prediction is performed. The missing data may be fulfilled with the built models, and after that the "new" data sets may be used for further prediction of concentrations of the same or other parameter. In the further research, it is possible to add the additional chemical or time dependence among the parameters , that will lead to new models for prediction.That way, in future, we may improve the use heuristic formula for prediction of ozone concentrations and decrees the MEA.

The experiments showed that the SVM is an appropriate tool for prediction of $O_3$ levels both for summer and winter seasons. The method gives good results and may be used by MoEPP for filling the data gaps for hourly $O_3$ values for short periods of time.

# References

[1] A. Pelliccioni, T. Tirabassi. (2006) Air dispersion model and neural network: A new perspective for integrated models in the simulation of complex situations. s.l. : Environmental Modelling & Software, Vol. 21, pp. 539–546.

[2] Agirre-Basurko, E., Ibarra-Berastegi, G. and Madariaga, I. (2006) Regression and multilayer perceptron-based models to forecast hourly O3 and NO2 levels in the Bilbao area. *4*, Environmental Modelling and Software, Vol. 21.

[3] Kolehmainen M., Martikainen H., Ruuskanen J. () Neural networks and periodic components used in air quality forecasting. *815–825*, 2001 : Atmospheric Environment, Vol. 35.

[4] Wang W., Lu W., Wang X., Leung A.Y.T. (2003) Prediction of maximum daily ozone level using combined neural network and statistical characteristics. s.l. : Environment International, Vol. 1049, pp. 1–8.

[5] Smola, Alexander J. and Scöpholf, Bernhard. (1998) *A tutorial on Support Vector Regression, Statistics and Computing.*

[6] Vapnik, V., Golowich, S. and Smola, A. (1997) Support vector method for function estimation, regression estimation and signal processing. Cambridge : MIT Press, Neural information processing sytems.

[7] Canu, Stephane and Rakotomamonjy, Alian. (2001) Ozone peak and pollution forecasting using support vectors. *IFAC Workshop on environmental modeling.*

[8] Weizhen Lu, Wenjian Wang, Leung, A.Y.T.Siu-Ming Lo Yuen, R.K.K. Zongben Xu, and Huiyuan Fan. (2002) Air pollutant parameter forecasting using support vector machines. *IJCNN*, , s.l. : I Neural Networks. 630–635.

[9] Wei-Zhen, Lu and Wen-Jian, Wang. (2005) Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends. s.l. : Elsevier, Chemosphere , Vol. 59, pp. 693-701.

[10] Cao, Lijuan. (2003) Support vector machines experts for time series Forecasting. s.l. : Elsevier, Neurocomputing, Vol. 31, pp. 321-339.