PREDICTION OF TROPOSPHERIC OZONE AT RIO DE JANEIRO CITY FROM PRIMARY POLLUENTS AND METEOROLOGICAL FACTORS USING ARTIFICIAL NEURAL NETWORKS (ANN) AND SUPPORT VECTOR MACHINE (SVM) REGRESSION

[1] G. C. G. de OLIVEIRA, A. S. LUNA, M. L. L. PAREDES e S. M. CORRÊA

[1] Universidade do Estado do Rio de Janeiro, Instituto de Química, Programa de Pós-graduação em Engenharia Química
[2] Universidade do Estado do Rio de Janeiro, Faculdade de Tecnologia
E-mail para contato: asluna@uerj.br

ABSTRACT - Tropospheric ozone is well known as an extremely important factor due to its strong environmental impact. The air quality depends on emissions, meteorology and topography. NO_2, NO, NOx, CO, O_3, scalar wind speed, solar radiation, temperature, and relative humidity (HUM) were monitored. These data sets were collected by the mobile station monitoring located at PUC-Rio and UERJ between 2011 and 2012, from the Secretary of the Environment of Rio de Janeiro. This study aimed at the prediction of O_3 from primary pollutants and meteorological factors. The obtained results of ANN and SVM regression techniques were acceptable to the dataset UERJ presenting coefficient of determination (R^2) for validation, 0.9122 and 0.9152, and the square root of the mean square error of cross validation (RMECV) 7.66 and 7.85, respectively. Regarding the data set of PUC-Rio, the two techniques gave inferior results.

1. INTRODUCTION

Tropospheric ozone can have a negative impact on the environment and public health when present in the lower atmosphere, in sufficient quantities. Regulations have been introduced to set limits on the emissions of pollutants in such a way that they cannot exceed prescribed maximum values (EPA, 1999).

One of the first published article dedicated to analyze the seasonal behavior of tropospheric ozone and its precursors was carried out in the São Paulo Metropolitan Area during 1987. The air pollutant measurements were related to both daily total insolation, and the number of hours of insolation measured. The total number of sunshine hours was highly correlated to mean hourly ozone concentration values as expected (Massambani and Andrade, 1994). However, if one wants to trace and predict ozone, one must know the conditions that contribute to its formation. Besides, ozone concentrations are strongly connected to meteorological conditions. Land–sea breezes also affect ozone concentrations at coastal sites. Therefore, it is necessary to develop a model that encompasses the description and understanding relationships between ozone concentrations and the many variables that cause or inhibit ozone production to predict ozone concentrations (Abdul-Wahab and Al-Alawi, 2002). One group of researchers

developed a model to predict the ozone concentration based on the meteorological variables using multiple linear regression (MLR). The observed and predicted values of daily ozone showed a correlation coefficient of 0.69 (Souza et al., 2014). Recently, a group of researchers (Wilson et al., 2012) analyzed trends in ozone levels in the European troposphere from 1996 to 2005. They indicated that average levels have been increasing despite reductions in pollutants which impact ozone formation. However, they also identified year-by-year variations, caused by climate and weather events, and suggested they could be masking the impact of emission reductions on long-term ozone trends. This study confirmed that the relationship between ozone and its precursors is complicated. It is caused by the fact that meteorological and chemical processes can interact over a remarkably wide range of temporal and spatial scales (Adame et al., 2008). For this reason, the use of statistical tools provides a sound framework for the analysis of such data.

Another study was carried out to compare the artificial neural networks (ANNs) with multiple linear regression models to predict the next day's maximum hour ozone concentration in the Athens basin. Results based on an extensive array of forecast quality measures indicate that the ANNs provides better estimates of ozone concentrations at the monitoring sites, whilst the more commonly used linear models are less useful for accurately forecasting high-ozone concentrations (Chaloulakou et al., 2003). Lately, another study was carried out to predict ozone concentration in the São Paulo Metropolitan Area using neural network based on average values of meteorological variables in the morning and afternoon periods. The model presented good performance as a prediction tool for the maximum ozone concentration (Borges et al., 2012). On the other hand, the support vector machine (SVM) can be used for time-series prediction and has been reported to perform well by some promising results. One group of researchers developed an online SVM model to predict air pollutant levels in an advancing time-series based on the monitored air pollutant database in Hong Kong downtown area (Wang et al., 2008). In another application, the SVM was used to the prediction of hourly ozone values in Madrid urban area. Using the modified SVM-r, based on reductions of the SVM-r hyper-parameters search space, they explore different influences, which may alter the ozone forecast, such as former ozone measurements in a given station, measures in neighbors stations, and the impact of meteorological variables. The prediction tool based on SVM-r was flexible enough to incorporate any other prediction variable, such as city models, or traffic patterns, which may enhance the prediction obtained with the SVM-r (Ortiz-García et al., 2010).

The objective of this study was to get parsimonious prediction models (i.e., models that depend on as few variables as needed) for ozone as a function of other ambient air concentration data and meteorological parameters as predictor variables, using the support vector machine (SVM-r) and artificial neural network regression (ANN-r) models.

2. THEORETICAL BACKGROUND

2.1 Support vector of machine regression (SVM-R)

The support vector machine (SVM) is a machine learning technique developed by Vapnik, which increasingly is gaining ground in many areas of knowledge. Originally this technique was developed for pattern recognition problems. The model consists of a number of support vectors (primarily samples selected from the calibration set) and non-linear model coefficients which define the non-linear mapping of variables in the input x-block. The model allows prediction of the continuous y-block variable.

SVM is implemented using the LIBSVM package which provides both epsilon-support vector regression (epsilon-SVR) and nu-support vector regression (nu-SVR). The original SVM formulations for Regression (SVR) used parameters C [0, inf) and epsilon[0, inf) to apply a penalty to the optimization for points which were not accurately predicted. An alternative version of both SVM regression was developed where the epsilon penalty parameter was replaced by an alternative parameter, nu [0,1], which applies a slightly different penalty. It is because nu represents an upper bound on the fraction of training samples, which are errors (badly predicted), and a lower bound on the fraction of samples which are support vectors. Some users feel nu is more intuitive to use than C or epsilon. Epsilon or nu are just different versions of the penalty parameter (Eigenvector Documentation). Recently, with the inclusion of loss functions in its structure, the SVM has expanded the usefulness of its use in the area of nonlinear regression and time series forecasts (Lu and Wang, 2005). A group of researchers used SVM method to estimate highly nonlinear source–receptor relationships between precursor emission and pollutant concentrations. They applied the SVM model for the resolution of the multi-objective air quality control problem in the Avilés industrial urban area in Spain (Suárez Sánchez et al. 2011). This method was used to predict the ozone concentration in this study.

2.2 Artificial neural networks regression (ANN-Regression)

The ANN technique analogously to the human nervous system has nodes in one or more layers, and connections, called synapses link these. According to Fiorin et al. (2011), this method is capable of storing knowledge, and its use covers problems of adjustment functions, pattern recognition, predictive modeling and other applications in different areas. This method has a high capacity for self-organization and temporal processing that enables to solve various problems of high complexity. In this context, the multilayer perceptron, one type of artificial neural network, has been shown to be a valuable tool for prediction, function approximation and classification. The benefits of the multilayer perceptron approach were particularly evident in applications where a complete theoretical model cannot be built, and especially when dealing with non-linear systems (Gardner and Dorling, 1998). According to Abdul-Wahab and Al-Alawi (2002), ANN-based models have the potential of describing nonlinear relationships such as those which control the production of O_3. Moreover, this technique was successfully used by these authors to predict the ozone concentration using meteorological and chemical data.

3. MATERIALS AND METHODS

3.1 Area description

The air pollution data for Rio de Janeiro city used in this study was provided by an automated mobile station, which provides hourly O_3 measurements in two different points of the city. Therefore, the databases were built using data collected by the mobile station, from the Secretary of Environment of Rio de Janeiro city, the first time at Pontifical Catholic University (PUC-Rio) (Latitude 22° 97´S and Longitude 43° 23´W), between July to October 2011, including the seasons of winter and spring. After collecting data at PUC-Rio, the mobile station was parked at Rio de Janeiro State University (UERJ) (Latitude 22° 91´S and Longitude 43° 23´W), between November 2011 to March 2012, during the spring and summer seasons.

3.2 Data collection

The original data were arranged in a matrix form. The following variables were used as columns: content of nitrogen dioxide (NO_2), nitrogen monoxide (NO), nitrogen oxides (NOx), carbon monoxide (CO), and ozone (O_3) in the air, scalar wind speed (SWS), global solar radiation (GSR), temperature (TEMP), moisture content in the air (HUM). The samples were arranged in the rows as were taken during the time in an average hourly base. The databases were created under varying conditions; pollution sources and weather conditions for the two studied localities, at different seasons. So, for the database PUC_2 and UERJ_2 different results are expected. This study will propose models for predicting levels of ozone for database PUC_2 and UERJ_2, separately constructed from data obtained at the mobile station located at PUC-Rio and UERJ, respectively. Additionally, in order to provide a greater challenge to the methods used, the databases have been unified in a single database titled PUC_UERJ_2 where techniques should provide a single prediction model for ozone content in the air, considering the two different databases merged, in order to determine its robustness.

Data collected on rainy days were removed from the databases, since rain lower levels of pollutants. Furthermore, the data obtained on the weekends and holidays need to be treated separately because vehicular source emissions are drastically reduced and, therefore, the concentrations of pollutants are remarkably different from weekdays. The data, obtained at night, must be also removed because the mixture layer of the atmosphere is extremely low, which increases the concentration of some primary pollutants, and the photochemical reaction is also virtually nonexistent. After the first refining of databases, in accordance with the aforementioned considerations, three databases were obtained: PUC_2: 643 samples with 9 variables (643 x 9); UERJ_2: 453 samples with 9 variables (453 x 9); the combined database PUC_UERJ_2: 1096 samples with 9 variables (1096 x 9).

3.3 Software and the methodology of the chemometric analysis

The software Matlab R2008b version 7.7.0 (Mathworks, USA) was used to construct the artificial neural networks (ANN) models and PLS Toolbox 6.2 (Eigenvector Research, USA) was used to construct models of support vector machine (SVM), respectively.

4. RESULTS

4.1. Modeling of ozone concentration in tropospheric levels using support vector of machine regression

Auto-scaling was used to preprocess datasets before the application of the regression technique. The SVM-R was applied selecting a 10-fold cross-validation sub-dataset. The epsilon support vector regression showed the best results when it was compared to Nu support vector regression for the prediction of the contents of O_3. Table 1 compares the figures of merit from the models obtained for different datasets using support vector machines.

Table 1 - Figures of merit obtained for the O_3 prediction using epsilon-SVM regression.

Figures of merit	PUC_2	UERJ_2	PUC_UERJ_2
RMSEC	11.62	5.95	11.95
RMSECV	13.79	7.66	14.47
R^2_{Cal}	0.7834	0.9483	0.8045
R^2_{CV}	0.6915	0.9122	0.7099

The values of RMSEC and RMSECV are errors, respectively, of the calibration and cross-validation. RMSEC values and RMSECV indicate the amount of the error of calibration and prediction; respectively. These values differ from zero to infinity, with zero being the best possible value to be achieved for a model. The calculation of RMSECV is shown in Eq. (1):

$$RMSECV = \sqrt{\frac{1}{N}\sum_{I=1}^{N}(P_i - M_i)^2} \quad (1)$$

where P_i is predicted, the corresponding measured value M_i and N is the number of segments. It is clear that UERJ_2 dataset showed minor errors of calibration and prediction when it was compared with the others datasets. The coefficients of determination of calibration (R^2_{Cal}) and validation (R^2_{CV}) obtained for SVM regression model for UERJ_2 dataset were higher than those of PUC_2 or PUC_UERJ_2 datasets. SVM-regression model performed well for UERJ_2, which seems to be a dataset with simpler relationship between ozone concentration and the other observed variables.

The predicted versus observed variable plot is useful for model and sample diagnostic. The resulting behavior is to have the points make a line of slope 1 and intercept 0. If the intercept is nonzero, it shows a constant bias between the known and forecast values. A slope not equal to 1 indicates a proportional bias. Figure 1 did not show bias and shows reasonable agreement between predicted and observed values for ozone concentration.

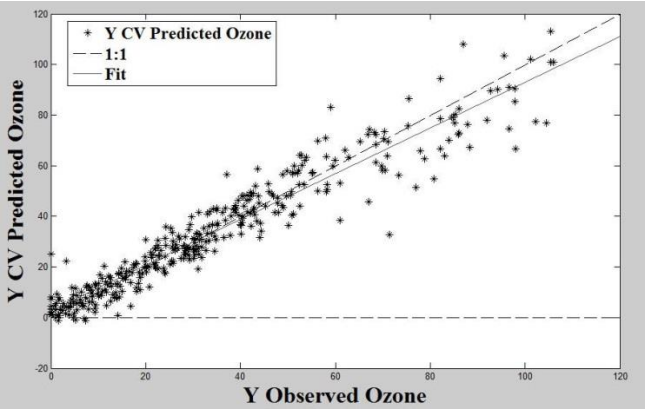

Figure 1 - Predicted versus observed for ozone concentration model using SVM-R for UERJ_2 dataset.

The White test is a statistical test that establishes whether the residual variance of a variable in a regression model is constant; therefore, this is a test for homoscedasticity. This test and an estimator for heteroscedasticity-consistent standard

errors were proposed by Halbert White in 1980. To test for constant variance one undertakes an auxiliary regression analysis: these regresses the squared residuals from the original regression model onto a set of regressors that contain the original regressors, the cross-products of the regressors and the squared regressors. One then inspects the R^2 (coefficient of determination). The Lagrange multiplier (LM) test statistic is the product of the R^2 value and sample size (n):

$$LM = nR^2 \qquad (2)$$

It follows a chi-squared distribution with degrees of freedom equal to the number of estimated parameters (in the auxiliary regression). For all regression models examined, the White test indicated that the residuals variance were constant (p-value <0.05) (White, 1980).

4.2. Modeling of ozone concentration in tropospheric levels using artificial neural networks regression

For the analysis of ANN, some parameters must be present for the system construction. They are the number of network layers, number of nodes in each layer, the relationship between the nodes and the network topology. The number of layers to be chosen depends on the linearity of the data. The multilayer ANN has more than one node between an input and output of the network. The connections between the nodes can be either acyclic (feed-forward) or cyclic (feedback). The connection is acyclic when the output response of a node serves as input to a node of a subsequent layer. The connection is cyclic when the response of the output node serves as input to a node of a single layer or lower layer. When occurs the output feedback in the input layers, the network is said to be recurring (Fiorin *et al.*, 2011).

For all three databases (PUC_2, UERJ_2 and PUC_UERJ_2), the artificial networks were chosen to be acyclic (feed-forward), consisting of eight input nodes, representing variables: NO, NO_x, NO_2, CO, SWS, TEMP, HUM and GSR and an output node, the target variable O_3 in two layers with sigmoidal hidden nodes and output nodes linear. For the hidden layer, 20 nodes were considered which is a parameter arbitrated by the Matlab software used. The networks were trained with training supervision static, with the Levenberg – Marquard method, which is a standard technique used to solve nonlinear least squares curve fitting problems.

In the first stage, the parameterization of networks was done, and in the second stage each dataset was divided into three datasets, which are the training set, with 70% of this data, validation set and test set, each with 15% of the data. The samples, in each dataset, were chosen by the algorithm of Kennard-Stone.

Figure 2 shows the graphs of predicted versus measured data for a set training, validation, testing and complete set for UERJ_2 dataset. The dotted lines in Figure 4 represent the line where the predicted values are equal to the measured (Y = T) and the solid lines represent the model. The slopes of the lines of the model are influenced by the distribution of the samples, the smaller the angle between the two lines, the closer are the predicted values to the measured ones. The higher the R-value, the smaller is the distance between the observed and predicted values. Again, as for the SVM technique, the ANN achieved the best results for the database UERJ_2.

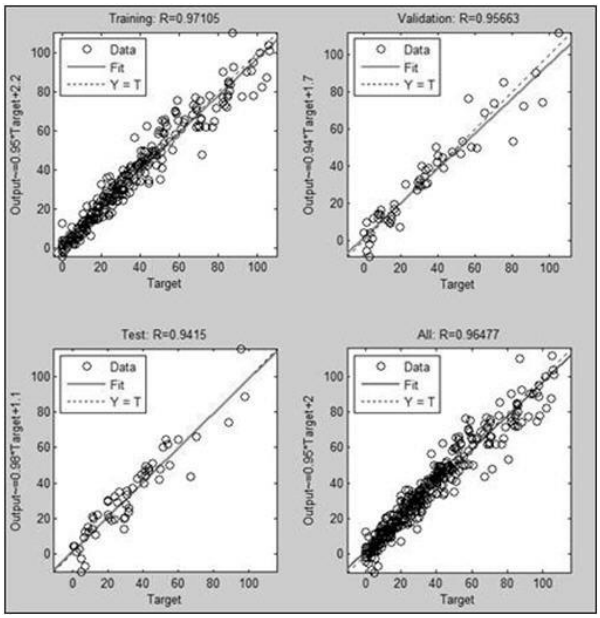

Figure 2 - Predicted versus measured data for training, validation, testing, and complete using ANN regression for UERJ_2 dataset.

Table 2 showed the figures of merit, for all datasets obtained for models using ANN regression method. Once more, this approach offers the best results when applied to UERJ_2 dataset as well as occurred with SVM regression technique.

Table 2 - Figures of merit obtained from ANN regression model applied to datasets.

Figures of merit	PUC_2	UERJ_2	PUC_UERJ_2
RMSEC	12.37	6.32	11.91
RMSECV	12.95	7.85	14.57
RMSEP	16.51	8.10	15.27
$R^2_{Training}$	0.7310	0.9429	0.8013
$R^2_{Validation}$	0.7543	0.9151	0.7205
R^2_{Test}	0.6527	0.8864	0.6583

Agreeing with the technique SVM, ANN also had more success in modeling predictions of O_3 to the UERJ_2 dataset with low RMSEC and RMSECV and R^2_{Cal} and R^2_{CV} high values. The results for PUC_2 were slightly worse again; however, the results were acceptable for PUC_UERJ_2 dataset as already mentioned for the SVM method.

Both techniques presented models less suitable for the database PUC_2, which seems to describe the complex relationship between ozone concentration and the other monitored variables. These effects can be attributed to its geographical location of the site located between the sea and the lagoon. Its region is formed of high atmospheric turbulence, which strongly influences the variation of pollutant concentrations. Therefore, the dataset has high variability of the data because there are sudden changes in the concentrations of pollutants, making it difficult to predict concentrations of O_3 as a function of the monitored variables.

5. CONCLUSION

The artificial neural network and support vector machine techniques were used to estimate the complex relationship between ozone and other variables based on ambient air monitoring measurements. The results provide insight into the dependence of ozone concentrations on other key pollutant concentrations and meteorological conditions. It was found that the models' predictions and the actual observations were consistent. The relative importance of the various input variables was also examined. The results also indicated the dependence of ozone concentrations on the other pollutants and meteorological conditions. Mainly, ozone concentration was negatively correlated with CO, NO, NO_x as expected since these pollutants are known precursors of ozone, indicating that a rise in ozone concentration follows a drop in the levels of these variables. Similarly, the correlation between ozone and HUM is also negative due to the reaction between water and ozone, leading to the oxidative specie OH.

This study allows implying that both chemometric techniques can be used in modeling and predicting the ground-level concentrations of ozone, with the determination coefficient (R^2) up to about 0.95. Clearly, this study has indicated the potential of chemometric tools application for capturing the non-linear interactions between ozone and other factors and for the identification of the relative importance of these factors. Artificial neural network and support vector machine modeling, therefore, provide an easy way of modeling and analysis of air pollutants and could be used in conjunction with other methods. Also, the obtained results support the fact that variability patterns are usually associated with the interaction of regional-level meteorological.

6. REFERENCES

ABDUL-WAHAB, S. A.; AL-ALAWI, S. M. Assessment and prediction of tropospheric ozone concentration levels using artificial neural networks. *Environ. Modell. Softw.* 17, 219-228, 2002.

BORGES, A. S.; ANDRADE, M. F.; GUARDANI, R. Ground-level ozone prediction using a neural network model based on meteorological variables and applied to the metropolitan area of São Paulo. *Int. J. Environ. Pollut*. 49(1-2), 1 – 15.

CHALOULAKOU, A.; SAISANA, M.; SPYRELLIS, N. Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. *Sci. Total Environ.* 313, 1-13, 2003.

EPA. *Guideline for developing an ozone forecasting program.* USA, Environmental Protection Agency, EPA-454/R-99-009, 1999.

FIORIN, D. V.; MARTINS, F. R.; SCHUCH, N. J.; PEREIRA, E. B. Aplicações de redes neurais e previsões de disponibilidade de recursos energéticos solares. *Rev. Bras. Ens. Fis.* 33, 1309-1320, 2011.

GARDNER, M. W.; DORLING, S. R. Artificial neural networks (the multilayer perceptron) – a review of applications in the atmospheric sciences. *Atmos. Environ.* 32, 2627-2636, 1998.

LU, W.; WANG, W. Potential assessment of the "support vector machine" method in forecasting ambient air pollutant trends. *Chemosphere* 59, 693-701, 2005.

MASSAMBANI, O.; ANDRADE, F. Seasonal behavior of tropospheric ozone in the São Paulo Metropolitan Area. *Atmos. Environ*. 28(19), 3165-3169, 1994.

ORTIZ-GARCÍA, E. G.; SALCEDO-SANZ, S.; PÉREZ-BELLIDO, Á. M.; PORTILLA-FIGUERAS, J. A.; PRIETO, L. Prediction of hourly O_3

concentrations using support vector regression algorithms. *Atmos. Environ.* 44, 4481-4488, 2010.

SÁNCHEZ SUÁREZ, A.; GARCÍA NIETO, P. J.; RIESGO FERNÁNDEZ, P.; DEL COZ DÍAZ, J. J.; IGLESIAS-RODRÍGUEZ, F. J. Application of an SVM-based regression model to the air quality study at local scale in the Avilés urban area (Spain). *Math. Comput. Model.* 54, 1453-1466, 2011.

SOUZA, A.; ARISTONES, F.; PAVÃO, H. G.; FERNANDES, W. A. Development of short-term ozone prediction tool in Campo Grande-MS-Brazil area based on meteorological variables. *Open J. Air Pollut.* 3, 42-51, 2014.

SVM Support Vector Machine for regression, Eigenvector Documentation, see http://wiki.eigenvector.com/index.php?title=SVM_Function_Settings.

WANG, W.; MEN, C.; LU, W. Online prediction model based on support vector machine. *Neurocomputing* 71, 550-558, 2008.

WILSON, R.C.; FLEMING, Z. L.; MONKS, P. S.; CLAIN, G.; HENNE, S.; KONOVALOV, I. B.; SZOPA, S.; MENUT, L. Have primary emission reduction measures reduced ozone across Europe? An analysis of European rural background ozone trends 1996-2005. *Atmos. Chem. Phys.*, 12, 437-454, 2012.

WHITE, H. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica* 48 (4), 817-838, 1980.