CORRELATION OF AIR POLLUTION AND METEOROLOGICAL DATA USING NEURAL NETWORKS

Theodora Slini, Kostas Karatzas and Nicolas Moussiopoulos
Aristotle University, Laboratory of Heat Transfer and Environmental Engineering,
Box 483, GR-54124 Thessaloniki, Greece

INTRODUCTION

The attempt to improve the effectiveness and the operational ability of classic statistical methods and discover perspectives of modern and sophisticated approaches such as neural networks, for air quality forecasting, is the basic motive behind the present research.

Linear regression methods have been applied for decades and are well known and understood (*Millionis, A.E. and T.D. Davies,* 1994; *Robeson, S.M. and D.G. Steyn,* 1990; *Ryan, W.F.* 1995; *Shi, J. P. and R.M. Harrison,* 1997*)*. However, there are numerous environmental processes that exhibit significant non-linear behaviour. Advances in the field of Artificial Neural Networks (ANN) in the late 1980s popularised non-linear regression techniques like Multi-layer Perceptons (MLP) and self-organising maps (SOM). It is shown that Neural Networks (NN) can be trained to successfully approximate virtually any smooth, measurable function (*Hornik, K., M. Stinchcombe and H.White,* 1989). NN are highly adaptive to non-parametric data distributions and, whilst other statistical methodologies require a set of assumptions to be fulfilled, the former make no prior hypotheses about the relationships between the variables. NN are also less sensitive to error term assumptions and they can tolerate noise, chaotic components and heavy tails better than most of the others methods. Other advantages include greater fault tolerance, robustness, and adaptability especially compared to expert systems, due to the large number of interconnected processing elements that can be trained to learn new patterns (*Lippman, R.P.*, 1987). These features provide NN the potential to model complex non-linear phenomenon like air pollution (*Kolhmainen, M., H. Martikainen and J. Ruuskanen*, 2001; *Perez, P. and A. Trier,* 2001; *Chelani, A.B., D.G. Gajghate and M.Z. Hasan*, 2002).

THE DATA

In the present paper, MLP models were developed including a number of air quality and meteorological parameters for the Greater Athens Basin, Greece. More specifically, a 2-year long (1995-96) data set is being used, consisting of hourly air pollutant concentrations and meteorological information, as resulting from the operation of the corresponding monitoring networks in the city of Athens. This time period was selected on the basis of data homogeneity availability and completeness. Information on the air pollution monitoring network of Athens can be found in *Directorate of Air and Noise Pollution*, 1997.

In particular, the data set collection of the current study consists of 2-year long hourly values for the concentration levels of carbon monoxide (in mg/m^3), nitrogen dioxide and ozone (in $\mu g/m^3$), as well as hourly data for air temperature and soil temperature at 15cm below the ground surface (^oC), solar radiation (Wm^{-2}), wind speed (ms^{-1}) and direction (rad), pressure (mbar), and relative humidity.

The meteorological observations orogonate from the meteorological observation station of the National Observatory of Athens for the same time period, and they are expected to be of high quality and consistence. The O_3 air quality information used in the analysis were collected at the monitoring stations of Patision, Marousi and Liosia, located nearly at the centre, at the northwest and northeast suburbs of the city of Athens respectively. Regarding the NO_2 and CO air quality

information, the monitoring stations of Patision and Athinas were used, situated nearly at the centre of the city. The selection of the monitoring stations was based on the frequency of high concentration levels of the respective pollutant observed in the whole Athens Basin. In addition to the available variables presented in the first part, variables describing the month and the day were created, so as to reveal the potential influence of those factors to the behaviour of the examined air pollutants. Pollution episodes are highly correlated to the spatial and temporal distribution of emissions. The dependence of the emissions on the city activities induces generally low pollution levels during weekends, even if the prevailed weather conditions are unfavourable. Similarly, the majority of pollution episodes take place during the summer months (April to September) (*Ziomas, I., D. Melas, C. Zerefos, A. Paliatsos and A. Bais* , 1995).

During the selection process of the influence factors, variables with the largest partial correlation coefficient with pollutant formation were selected. The transformation of the variables was effective and led to better results, in some cases. For example, the sine and cosine functions of the wind direction were created, leading to minor improvements of the models. However, the index WD= 1+sin(θ+π/4), suggested by *Melas, D., I. Kioutsioukis, and I. Ziomas* (2000) and adapted by *Chelani, A.B., D.G. Gajghate and M.Z. Hasan*, (2002), provided better results in most of the cases.

NEURAL NETWORKS

Neural Networks use a complex combination of weights and functions to convert input variables into an output (prediction). The MLP consists of a system of simple processing interconnected elements called neurons, cells or nodes. Each of the various inputs to the network is multiplied by a connection weight. These products are simply summed, fed through a transfer function to generate a result, and then output. This is a gradient descent algorithm that is normally used to train a MLP network. Errors in the output of this procedure are assumed to be due to all processing elements and connections, and these errors are reduced by propagating the output error backward to the connections in the previous layer.

The type of neural network used in this study was the three-layer back-propagation network, consisting of an input layer, a hidden layer and an output layer (*Cobourn, W.G., L. Dolcine, M. French and M.C. Hubbard,* 2000). The learning algorithm used in the present study was Levenberg-Mardquardt back-propagation of Matlab Neural Network toolbox. The transfer functions selected for the layers were sigmoid (eq. 1) for the hidden layer and linear for the output layer (*Gardner, M.W and S.R. Dorling,* 1998; *Kolhmainen, M., H. Martikainen and J. Ruuskanen*, 2001). The S-shaped logistic sigmoid function is bounded between 0 and 1, therefore input and output data should be also normalised in the same range.

$$f(x) = \frac{1}{(1+e^{-x})} \tag{1}$$

$$\hat{z} = \frac{z - z_{\min}}{z_{\max} - z_{\min}} \tag{2}$$

Data were normalized using eq. 2, where $\hat{z}$ is the normalised value and z_{min} and z_{max} are the minimum and the maximum values of z, respectively. The adapted neural network models consisted of 9 nodes in the input parameter, 20 nodes in the hidden layer and one node in the output. The development of the models consisted of two steps. The first step is the training stage, where the network is subjected to a training set of input-output patterns for the year 1995

time series. The second step is the testing stage, where the performance of the network is tested on patterns that have been 'revealed' during the former stage; the year 1996 was used as the testing set.

RESULTS

The lagged observation values (i.e. of the previous hour or day) were completely ignored, as this research is concentrated on the extraction of potential relationship between air quality and meteorological parameters. This is probably one of the reasons for the low forecasting ability of the derived models. The development set includes the first year observations (1995), which correspond to 8760 values (n=8760), while the following 8784 (1996, a leap year) were used as the validation set for the prediction.

The ability of NN models to forecast daily hourly concentration levels o O_3, NO_2 and CO was verified by performing a linear regression between the network response and the target (observed values). Several statistics were used to evaluate errors of prediction results, including the observed and forecasted mean and standard deviation of the pollutant concentration levels, as well as correlation coefficient (r), mean bias error (MBE), error standard deviation (s), the root mean square error (RMSE). The results are summarized at two tables: Table 1 for O_3 and Table 2 for NO_2 and CO.

The analysis of the results shows that the observed and forecasting means are similar for the training sets; nonetheless they are quite different for the testing period. Additionally, the observed standard deviations are greater than the forecasted for all sites, revealing the ineffectiveness of the models to predict high concentration levels of the pollutants examined. Regarding the correlation coefficient, measuring the linearity between observed and predicted values, showed similarities in the relative performances of the models for the training tests of all sites, except Athinas station, which had the lowest concentration (0.45). Marousi for O_3 and Patision for CO had the highest correlation in the range of 0.877. However, the same linear relationship in the testing set is significantly reduced ranging between 0.383 (Patision – O_3) and 0.010 (Athinas –CO, as it was expected).

Table 1. Model forecast statistics between observed and model forecasted ozone at Patision, Liosia and Marousi monitoring stations.

Model Statistics	O_3					
	Patision		Liosia		Marousi	
	1995	1996	1995	1996	1995	1996
mean (obs.)	25.43	28.16	61.73	61.73	63.69	68.71
mean (for.)	25.47	41.96	62.45	59.82	63.68	73.75
st.dev. (obs.)	19.25	20.91	39.27	41.69	45.25	44.48
st.dev. (for.)	16.67	33.88	34.02	36.99	39.69	39.96
r	0.866	0.383	0.866	0.062	0.877	0.177
MBE	-0.000	-0.180	0.000	0.325	0.000	-0.028
s	0.182	0.422	0.123	0.556	0.125	0.304
RMSE	0.182	0.459	0.123	0.555	0.124	0.305

Table 2. Model forecast statistics between observed and model forecasted nitric dioxide and carbon monoxide at Patision and Athinas monitoring stations.

Model tatistics	NO_2				CO			
	Patision		Athinas		Patision		Athinas	
	1995	1996	1995	1996	1995	1996	1995	1996
mean (obs.)	93.51	95.36	51.40	80.24	5.12	4.83	3.27	3.61
mean (for.)	93.38	70.20	51.40	196.4	-0.60	-0.66	17.69	3.11
st.dev.(obs.)	39.78	46.14	32.15	32.49	3.06	2.93	2.41	2.49
st.dev.(for.)	30.57	55.41	14.70	80.31	0.17	0.23	6.55	2.46
r	0.768	0.089	0.450	0.117	0.877	0.177	0.761	0.010
MBE	0.000	0.212	-0.000	-0.579	-0.000	-0.028	0.000	0.038
s	0.133	0.298	0.449	0.614	0.125	0.304	0.120	0.261
RMSE	0.133	0.366	0.533	0.733	0.124	0.305	0.120	0.264

Furthermore, the MBE, expressing the difference between the estimated mean and population mean, equals zero in all training cases, while it is raised up to –0.579 for NO_2 at Athinas station. Generally, the negative values of the MBE indicate the trend of the adapted models to underpredict the observed data and do not effectively capture the extreme values that are of major concern for operational use. The standard deviation is used to present the range of prediction errors based on observed data, reaching high values at Liosia and Athinas stations for O_3 (0.556) and NO_2 (0.614), respectively. Finally, the square roots of the mean of all squared residual between the observed and forecasted (RMSE: error due to model) were fairly high, especially at the abovementioned stations for the same pollutants, indicating the capability of further improvement of the models.

CONCLUSIONS

Overall, it appears that the MLP models developed did not perform very satisfactorily with the current datasets, preventing the extraction of valuable information about the connection of air quality to meteorological variables. The results of the analysis state that NN appear quite complex and less familiar than traditional statistical methods, suggesting that more experimentation is needed. Therefore, implementation of a NN may be considerably more difficult than using a classic linear model. There is a strong relation and dependence on user judgement that can be disastrous to the model fitting. Complex NN may induce spurious correlations between explanatory and response variables, as too simple ones will result in a pure ability to generalize a functional relationship. Moreover, NN are arguably extreme examples of a black-box approach where the adapted models slightly improve in understanding the underlying data generating mechanism (*Balkin, S.D.*, 2000). Likely strategies include using additional variables (for example additional weather elements and synoptic patterns) or changing the nature of the model (for example different number of hidden neurons and layers in the model). It is also worthy to mention that alternative possibilities suggests the weather – air pollutants relationships might be sufficiently complicated that a level of random variation or noise exists (as assumed in some ozone trend studies), which cannot be captured by even a relatively sophisticated empirical-statistical model such as a neural network (*Flaum J.B., S.T. Rao and I.G. Zurbenko,* 1996).

REFERENCES

Balkin, S.D., 2000: A statistical implementation of feedforward neural networks for univariate time series forecasting. A thesis in Business Administration. UMI.

Chelani, A.B., D.G. Gajghate and M.Z. Hasan, 2002: Prediction of Ambient PM_{10} and Toxic Metals Using Artificial Neural Networks. *J. of Air and Waste Management Ass.,* **52**, 805-810.

Cobourn, W.G., L. Dolcine, M. French, and M.C. Hubbard, 2000: A comparison of nonlinear regression and neural network models for ground-level ozone forecasting. *J. of Air and Waste Management Ass.,* **50**, 1999-2009.

Comrie, A.C., 1997: Comparing neural networks and regression models for ozone forecasting. *J. of Air and Waste Management Ass.,* **47**, 653-663.

Directorate of Air and Noise Pollution, 1997: Air pollution in Athens-1996. Ministry of Environment, Planning and Public Works.

Flaum, J.B., S.T. Rao and I.G. Zurbenko, 1996: Moderating the influence of meteorological conditions on ambient ozone concentrations. *J. of Air and Waste Management Ass.,* **46,** 35-46.

Gardner, M.W and S.R. Dorling, 1998: Artificial Neural Networks (The Multilayer Precepton)-A review of applications in the atmospheric sciences. *Atmospheric Environment,* **32**, 14-15, 2627-2636.

Hornik, K., M. Stinchcombe and H.White, 1989: Multilayer feedforward networks are universal approximators. *Neural Networks*, **2**, 359-366.

Kolhmainen, M., H. Martikainen and J. Ruuskanen, 2001: Neural Networks and periodic components used in air quality forecasting. *Atmospheric Environment,* **35**, 815-825.

Lippman, R.P., 1987: An introduction to computing with neural nets. *IEEE ASSP Magazine.*

Melas, D., I. Kioutsioukis, and I. Ziomas, 2000. Neural Network model for predicting peak photochemical pollutant levels. *J. of Air and Waste Management Ass.,* **50**, 495-501.

Millionis, A.E. and T.D. Davies, 1994: Regression and stochastic models for air pollution -I. Review, comments and suggestions. *Atmospheric Environment,* **28**, 17, 2801-2810.

Perez, P. and A. Trier, 2001: Prediction of NO and NO_2 concentrations near a street with heavy traffic in Santiago, Chile. *Atmospheric Environment,* **35**, 1783-1789.

Robeson, S.M. and D.G. Steyn, 1990: Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations. *Atmospheric Environment,* **24B**, 2, 303-312.

Ryan, W.F. 1995: Forecasting severe ozone episodes in the Baltimore metropolitan area. *Atmospheric Environment,* **29**, 17, 2387-2398.

Shi, J. P. and R.M. Harrison 1997: Regression modelling of hourly NO_x and NO_2 concentrations in urban air in London. *Atmospheric Environment,* **31**, 24, 4081-4094.

Ziomas, I., D. Melas, C. Zerefos, A. Paliatsos and A. Bais, 1995: On the relationship between peak ozone levels and meteorological variables. *Fresenius Environmental B.,* **5**, 53-58.