Jul 1st, 12:00 AM

Studying and predicting quality of life atmospheric parameters with the aid of computational intelligence methods

Kostas Karatzas

D. Voukantsis

iEMSs 2008: International Congress on Environmental Modelling and Software
Integrating Sciences and Information Technology for Environmental Assessment and Decision Making
4th Biennial Meeting of iEMSs, http://www.iemss.org/iemss2008/index.php?n=Main.Proceedings
M. Sànchez-Marrè, J. Béjar, J. Comas, A. Rizzoli and G. Guariso (Eds.)
International Environmental Modelling and Software Society (iEMSs), 2008

Studying and predicting quality of life atmospheric parameters with the aid of computational intelligence methods

K D. Karatzas and D. Voukantsis
Department of Mechanical Engineering, Informatics Applications and Systems Group, Aristotle University, 54124 Thessaloniki, Greece
(kkara@eng.auth.gr, voukas@isag.eng.auth.gr)

Abstract: Air quality management is among the most challenging problems in terms of analysis and modelling. Air quality modelling and forecasting is directly affected by the highly nonlinear relationships between pollutants and weather, while in many cases there is insufficient domain knowledge due to the influences of local conditions. As atmospheric quality has an impact on the quality of life of millions of people, the ability to reveal interrelationships between parameters that influence environmental decision making is very important. In addition, forecasting of such parameters for the purpose of early warning and health risk prevention is of paramount importance for sensitive parts of the population. In the present paper a number of Computational Intelligence methods are presented with the purpose to the investigate atmospheric quality parameters, towards better understanding of the interrelationships between pollutants, and with the aim to improve forecasting of critical values. For this reason, the use of Fast Fourier Transformation for the construction of Periodograms is firstly presented, followed by the application of Principal Component Analysis. Then, Self Organizing Maps, a method based on the Neural Networks approach, is investigated and applied, for knowledge extraction and atmospheric parameter analysis. Last, an Artificial Neural Network based on the multi-linear perceptron (MLP) model is presented in order to construct prediction models. Results indicate a number of important features within the data investigated, and reveal hidden interrelations, thus providing valuable information for the understanding and the explanation of environmental problems, and for the support of environmental policy and decision making in both long and short terms. It is also demonstrated that the performance of forecasting models justify their selection for early warning information services.

***Keywords**:* Computational intelligence, periodograms, principal component analysis, self organising maps, artificial neural networks, air quality, atmospheric environment, quality of life.

1. INTRODUCTION

A knowledge domain is described on the basis of nominal, categorical or arithmetic values of parameters that serve as the basis for information creation, as they are being processed with the aid of various computational methods, tools or human judgment. In the case of the environmental engineering domain, these data have (in the majority of cases) the form of time stamped records that formulate a multivariate time series within the spatial and temporal scale of the phenomenon of interest.

Although air pollution is “interwoven” to the atmospheric environment, pollutants (i.e. harmful agents) behave differently in various spatial and temporal scales. Nevertheless, it has been agreed upon by the scientific community that certain pollutants should be monitored in locations that are representative of urban spatial forms like city centres, high traffic areas, residential areas, suburban or rural areas. Thus, the quality of the atmospheric environment may be described with the aid of hourly concentration values of various

pollutants like Ozone (O3), Nitrogen Oxides (NO, NO2 etc), Sulphur Dioxide (SO2), Particulate Matter with mean aerodynamic diameter of various scales (coarse, PM10, PM2.5, ultra fine), and other pollutants. In addition, a number of meteorological parameters influence the quality of air and play an important role in our understanding concerning the life cycle of atmospheric pollution. These are parameters like wind speed and wind direction, air temperature, relative humidity, etc.

2. COMPUTATIONAL INTELLIGENCE FOR ATMOSPHERIC PARAMETER INVESTIGATION

The understanding of the relationships, dependencies, profiles and behaviour of parameters describing the quality of the atmospheric environment is of paramount importance for anyone interested to improve quality of life, especially in urban areas. Moreover, the ability to forecast such parameters, and especially those values that affect human health, is critical in all health prevention systems, Although many AQ modelling methods have been applied in this field, Computational Intelligence (CI) provides with an arsenal of scientific approaches that may be applied in order to better understand (and forecast) the behaviour of parameters of interest. Each urban area has its own physical, geographical and meteorological characteristics. Thus knowledge gained in one area may enrich our knowledge for another area and improve our understanding of environmental pressures. On this basis, a number of methods will be described in the next chapters aiming at demonstrating the advantages resulting from the usage of computational intelligence for atmospheric quality study and forecasting.

2.1 Fourier Transformation and Periodogram Construction

The discrete Fourier transform of a stationary discrete time series x(n) is a set of N-discrete harmonics X(k) according to eq. 1.

$$X(k) = \sum_{n=0}^{N-1} x(n) e^{-j 2\pi k n / N} \qquad (1)$$

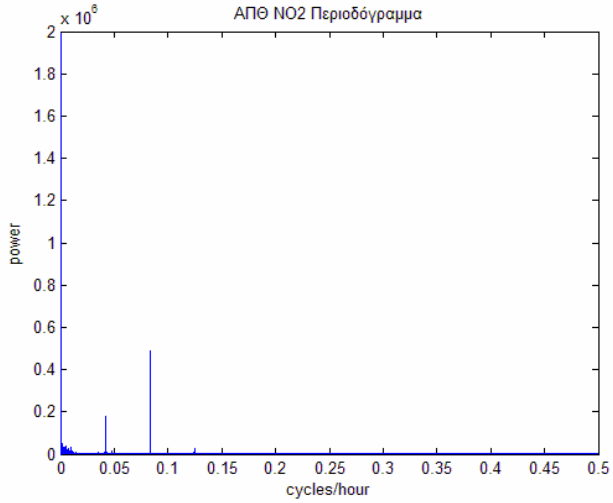

Figure 1. Periodogram of nitrogen dioxide time series (AUTh station, Thessaloniki)

This transformation represents the frequency response of x(n). X(k) is complex. The magnitude of X(k) squared, a real value, is called strength and a diagram of all the strength harmonics is called periodogram. When x(n) is real then the periodogram is symmetric and only half of the harmonics are needed [Karatzas et al., 2007]. Here x(n) is the hourly concentration values of pollutants. The method is demonstrated via its application on data coming from an urban station located in the city of Thessaloniki, Greece, in the campus of

the Aristotle University. These data describe NO2 hourly concentrations for the years 2001-3 (Figure 1). Thessaloniki is a city of 1 million inhabitants and more than 400 000 vehicles, suffering from the asphyxiating pressures of urban development. Thus, it is common than atmospheric pollutants reach high values, and the investigation of their behaviour is of high importance for the authorities responsible for pollution abatement. The specific periodogram reveals that the main periodicity of NO2 is approx 0.08 cycles/hour, i.e. 12 hours. This means that this pollutant has a 12 hour cycle, thus its existence should be attributed to a source having the same periodicity, i.e. urban traffic, a finding indicating the influence of vehicle circulation on air quality in the city centre.

2.2 Principal Component Analysis

Principal component analysis (PCA) is a computational intelligence method originating from multivariate statistical analysis that allows for the identification of the major drives within a certain multidimensional data set and thus may be applied for data compression, identifying patterns in data, and expressing the data in such a way as to highlight their similarities and differences. Since patterns in data can be hard to find in high dimensional data, where the luxury of graphical representation is not available, PCA is a powerful tool for such an analysis. The other main advantage of PCA is that once these patterns have been identified the data can be compressed, using dimensionality reduction, without significant loss of information [Smith, 2002]. In order to demonstrate the effectiveness of the method, PCA was applied for atmospheric datasets derived from the air quality monitoring station located at Aristotle University, in Thessaloniki, Greece (same with the one used before). The software package employed for this purpose was MATLAB that has a built-in function for applying the method to any given dataset. Data for the period 2001-2003 were used. The data matrix contained measurements for O3, NO2, temperature, humidity, wind speed and wind direction transformed according to eq. 2 into the two new variables, namely sinwd and coswd [Karatzas and Kaltsatos, 2007].

$$v1 = \sin(2\pi(v - \min(v))/(\max(v) - \min(v))) \\ v2 = \cos(2\pi(v - \min(v))/(\max(v) - \min(v))) \quad (2)$$

The method's results as well as the correlation coefficients matrix are presented in Table 1 and Table 2 respectively.

Table 1: PCA results, AUTH station, 2001 – 2003.

	PC1	PC2	PC3	PC4
NO_2	-0.3049	-0.4226	0.0413	-0.7218
O_3	0.5345	0.0421	0.1121	0.1121
Temp	0.3773	-0.2596	0.6095	0.0932
Hum	-0.4233	-0.1367	-0.2845	0.5643
WS	0.3644	0.3986	-0.4673	-0.2001
SinWD	-0.3587	0.3469	0.5232	0.1583
CosWD	-0.1959	0.6741	0.2030	-0.2723
%Var	41.8937	20.7019	13.5555	11.1588
Cum. %Var	41.8937	62.5956	76.1511	87.3099
	PC5	PC6	PC7	
NO_2	0.2097	0.3084	0.2582	
O_3	0.2396	0.0062	0.7937	
Temp	-0.2171	0.5344	-0.2782	
Hum	-0.0627	0.5722	0.2672	
WS	0.3485	0.4993	-0.2814	
SinWD	0.6676	0.0297	-0.0749	
CosWD	-0.5295	0.2042	0.2643	
%Var	5.3721	4.6824	2.6355	
Cum. %Var	92.6820	97.3645	100	

Table 2: Correlation coefficient matrix, AUTH station (2001 – 2003).

	NO_2	O_3	Temp	Hum
NO_2	1.00	-0.51	-0.18	0.20
O_3	-0.51	1.00	0.59	-0.62
Temp	-0.18	0.59	1.00	-0.45
Hum	0.20	-0.62	-0.45	1.00
WS	-0.41	0.52	0.04	-0.42
SinWD	0.09	-0.42	-0.26	0.29
CosWD	-0.08	-0.28	-0.31	-0.00

	WS	SinWD	CosWD
NO_2	-0.41	0.09	-0.08
O_3	0.52	-0.42	-0.28
Temp	0.04	-0.26	-0.31
Hum	-0.42	0.29	-0.00
WS	1.00	-0.34	0.08
SinWD	-0.34	1.00	0.48
CosWD	0.08	0.48	1.00

Table 3: Varimax rotation to the first 2 PCs, AUTH station (2001 – 2003).

	PC1	PC2
NO_2	-0.51	-0.12
O_3	0.43	-0.32
Temp	0.12	-0.44
Hum	-0.41	0.17
WS	0.54	0.06
SinWD	-0.05	0.50
CosWD	0.29	0.64

The cumulative variance percentage of each Principal Component (PC) is considered to be a measure of the representativity of the information contained in the initial data set. Each coefficient (with values from -1 to 1), represents the "weight" of each parameter in the formulation of each PC. The PCs are normalized, so that the squares of each PC coefficients sum to one.

The Humphrey – Ilgen parallel analysis indicates that only 2 PCs, explaining 62.59% of the total variance of the data, should be considered for further analysis, while the rest PCs are more likely to represent random variations of the data. The varimax rotation was applied to the coefficients of the first 2 PCs, and the results are presented in Table 3. It is evident that PC1 reveals the antagonistic relationship between NO_2 with O_3, and thus indicating photochemistry influenced by traffic. Furthermore, the strong contribution of Humidity and Wind Speed to PC1, suggests that O_3 formation is mainly expressed by this PC. In contrast, PC2 expresses the production dispersion of O_3, since it is mostly characterized by wind temperature and direction. These findings are supported by the fact that NO_2 is the result of the chemical degradation of O_3 in the atmosphere, while humidity accumulates under conditions of weak solar radiation, the latter acting as a catalyst (and thus being required) in the production of O_3 from primary pollutants.

2.3 Self Organising Maps

The self-organizing map (SOM) is also referred to as Kohonen Network [Kohonen, 1997], and is a subtype of artificial neural networks. SOM is based on competitive learning, which runs in an unsupervised manner, aiming at selecting the so called winning neuron that best matches a vector of the input space (Figure 2). In this way, "a continuous input space of activation patterns is mapped onto a discrete output space of neurons by a process of competition of the neurons in the network" [Haykin, 1999]. This makes SOM one of best methods for modelling a knowledge domain with the aim to reveal topological interrelations and hidden knowledge, via the visualization of the network's neurons.

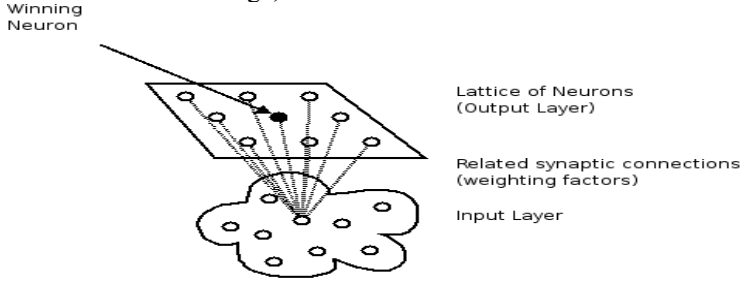

Figure 2. The generalised Kohonen Network. The input space is mapped on the output layer via the winning neurons. The topological relationship of the latter reveal dependencies and relationships between the parameters of the input space, on the basis of the features mapped on the output layer.

SOM is capable of learning from complex, multi-dimensional data without specification of the output. The resulting nonlinear classification consists of clusters that can be interpreted via visual inspection. The methods' unsupervised learning algorithm involves a self-organizing process to identify the weight factors in the network, reflecting the main features of the input data as a whole. In that process the input data is mapped onto a lower dimensional (usually two-dimensional) map of output nodes with little or no knowledge of the data structure being required. The output nodes (neurons) represent groups of entities with similar properties, revealing possible clusters in the input data. It should be noted that, although the method is unsupervised in learning, the number of the output nodes and configuration of the output map (number of nodes included, etc), need to be specified before the learning process [Karatzas and Kaltsatos, 2007b].

The air quality and meteorological data used for the demonstration of SOM capabilities, originate from the following monitoring stations in Thessaloniki, Greece: Aristotle University (city centre, within the University Campus), Kalamaria (urban stations in the east side of the city), Eleftherio-Kordelio (urban stations in the west side of the city) and Sindos (industrial area, located in the west of the city). As a first step to apply the SOM method, input variables were be normalized to unit variance, since Euclidean distance is employed as the error metric, and thus, if the input variables are not normalized, the mapping realized by the SOM may be dictated by some variable which has a much larger variance than the others. Then, the aim of the demonstration was to investigate air quality in one specific location (the Aristotle University- AUTh monitoring station).

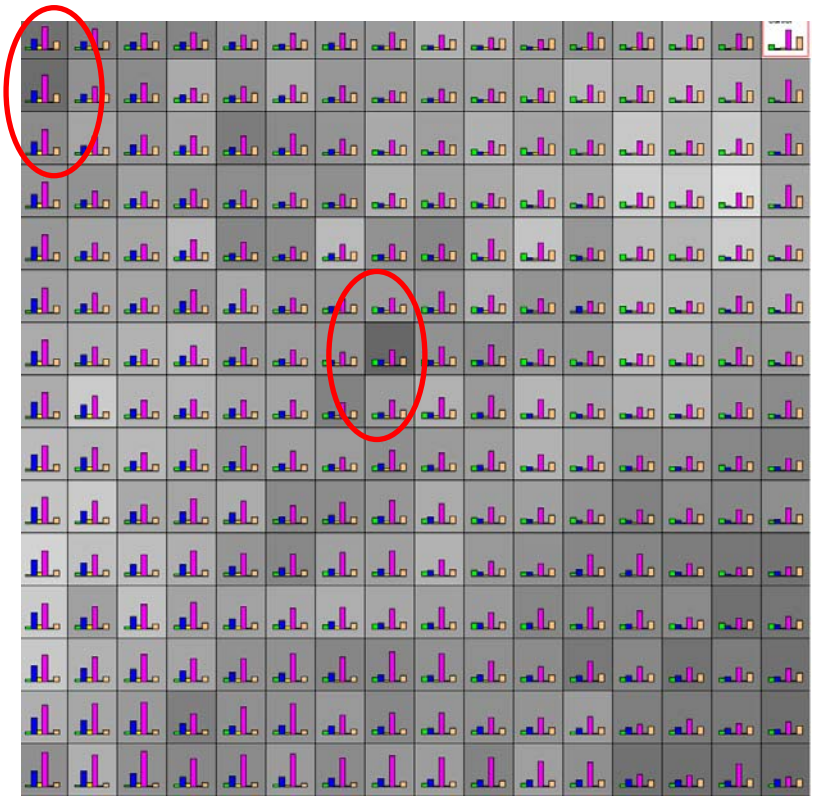

Figure 3. The resulting Self Organising Map for the AUTh station. Each square represents a neuron for which the frequency of occurrence is analogous to the darkness of the grey tone, and the values of the parameters constituting the neuron is analogous to the height of the relevant bars. These parameters (from left to right) are: NO_2, O_3, T, WD, WS, RH.

Figure 3 reveals two dark grey zones, which may be interpreted as clusters of neurons possessing s strong topological relationship. The cluster on the upper left reveals that high NO2 values (the 2nd data bar) are associated with low O_3 (1st data bar), a finding that is expected, due to the photochemical interconnection of these two pollutants. This cluster also reveals that high NO_2 occurs when WD (4th data bar) is high, i.e from 270-360 deg., and WS (5th) is low, indicating that there might be some local transportation from the

western parts of the city. This may be attributed to the sitting of major industrial combustion sources in the western parts of the area. The cluster in the centre reveals that WD (4th data bar) of values around 180 deg (i.e. south) are associated with high RH values, indicating that those are the winds transferring wet air over the city centre. This may be related to the fact that the local sea breeze circulation supports such air mass interactions between land and sea surfaces.

2.4 ANN models for modelling and forecasting ozone concentrations in Thessaloniki, Greece

In addition to the SOM method, Artificial Neural Networks were applied in order to demonstrate forecasting capabilities concerning hourly ozone concentration levels. Ozone is a photochemical pollutant that is being created as a result of the chemical imbalance in the atmosphere caused by pollutants like nitrogen dioxides and hydrocarbons. These pollutants enter the atmosphere mainly as a result of anthropogenic activities, and have a high and direct impact to the quality of life of citizens, especially those belonging to the so called sensitive part of the population.

The network architecture chosen here was the Multi-Layer Perceptron (MLP) with one hidden layer: the network consisted of one input layer, one hidden layer and the output layer. MLP models are widely applied in predicting air pollutant concentrations since they can capture the highly nonlinear relationship between the variables [Gardner and Dorling, 1999; Kolehmainen et al., 2001; Yi and Prybutok, 1996]. The datasets used include hourly concentration values for O_3, NO_2, and hourly records of temperature, humidity, wind speed and the transformed wind direction, as all these parameters were proven to be of importance for describing the investigated datasets via the PCA method previously presented. The choice of one hidden layer was made after several tests with different network structures since it gave us lower error values and smaller convergence times. Table 3 compares the forecasting performance statistical indexes for two different models. Model A uses NO_2, temperature, humidity, wind speed and transformed wind direction as prediction variables, while model B uses the one hour lagged O_3 concentration as an additional prediction variable

Table 3: Evaluation of two ANN models concerning hourly O_3 concentrations at AUTh, Thessaloniki

	Model A	Model B
MAE	27..326	13.532
RMSE	34.19	18.476
IA	0.814	0.948

It is clear that the Mean Absolute Error (MAE) and the Root Mean Square Error (RMSE) are better for model B. In addition, the Index of Agreement (IA), used for the evaluation of the model's ability to forecast the actual (observed) data, is much higher in the case of model B, indicating the fact that the use of O_3 values of previous hours improves the forecasting performance.

3. CONSLUSIONS AND RECOMENDATIONS

In the present paper a number of Computational Intelligence methods have been applied in order to demonstrate their efficiency for both problem understanding and parameter forecasting concerning air pollution concentration values. The results presented verify the high ability of the methods, and also reveal their unique ability to extract knowledge from the domain of interest and to help us understand and manage environmental problems. On this basis, it would be recommended that such analysis carried out in cases where problems related to the atmospheric environment and quality of life is concerned, in order to better inform policy and decision makers about the constrains of their decisions. In addition, such methods may provide with the necessary scientific competence for services related to public notification, warning and alerts, on the basis of forecasts.

ACKNOWLEDGEMENTS

This paper is related to the work that the authors are conducting in the frame of COST Action ES0602: Towards a European Network on Chemical Weather Forecasting and Information Systems (www.chemicalweather.eu). The data used originate from the freely accessible air quality database (AirBase) of the European Environment Agency (http://air-climate.eionet.europa.eu/databases/airbase) and from the Prefecture of Central Macedonia, Dept. of Environment, Greece, and are related to the project Airthess: Early warning informatics system for air pollution in Thessaloniki (www.airthess.gr).

REFERENCES

Haykin S., *Neural networks a comprehensive foundation*, Pearson Education, 2005

Gardner, M.W., Dorling, S.R., "Neural network modelling and prediction of hourly NOx an NO2 concentrations in urban air in London", *Atmospheric Environment*, vol. 31, 1999, pp. 709–719.

Karatzas K., G. Papadourakis and I. Kyriakidis (2007), Understanding and Forecasting Air Pollution with the Aid of Artificial Intelligence Methods in Athens, Greece. Proceedings of the Int. Workshop on Applications with Artificial Intelligence, 19th IEEE International Conference on Tools with Artificial Intelligence (ICTAI) October 29-31, 2007, Patras, Greece (http://ictai07.ceid.upatras.gr/). Springer Series on Studies in Computational Intelligence, in press.

Karatzas K., and Kaltsatos S. (2007), Air pollution modelling with the aid of computational intelligence methods in Thessaloniki, Greece, Simulation Modelling Practice and Theory, Volume 15, Issue 10, November 2007, Pages 1310-1319

Karatzas K. and Kaltsatos S. (2007b), Modelling and forecasting of air quality with the aid of machine learning methods in Thessaloniki, Greece, Proceedings of the 10th International Conference on Engineering Applications of Neural Networks, (Margaritis N. and Iliadis L., eds.), ISBN 978-960-287-093-8, pp. 419-427

Kohonen T., *Self organizing maps*, Springer, 1995

Kolehmainen M., Martikainen H., Ruuskanen J., "Neural networks and periodic components used in air quality forecasting", *Atmospheric Environment*, vol. 35, 2001, pp. 815-825.

Smith L., "A tutorial on Principal Components Analysis", (February 26, 2002, http://kybele.psych.cornell.edu/~edelman/Psych-465-Spring-2003/PCA-tutorial.pdf, last accessed 20.02.2008).

Yi, J., Prybutok, V.R., (1996), A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area, *Environmental Pollution*, vol. 92, pp. 349–357.