Developing Neural Networks to Investigate Relationships Between Air Quality and Quality of Life Indicators

Kyriaki Kitikidou and Lazaros Iliadis
*Democritus University of Thrace,
Department of Forestry and
Management of the Environment
and Natural Resources, Orestiada
Greece*

1. Introduction

Quality of life (QOL) is an integral outcome measure in the management of diseases. It can be used to assess the results of different management methods, in relation to disease complications and in fine-tuning management methods (Koller & Lorenz, 2003). Quantitative analysis of quality of life across countries, and the construction of summary indices for such analyses have been of interest for some time (Slottje et al., 1991). Most early work focused on largely single dimensional analysis based on such indicators as per capita GDP, the literacy rate, and mortality rates. Maasoumi (1998) and others called for a multidimensional quantitative study of welfare and quality of life. The argument is that welfare is made up of several distinct dimensions, which cannot all be monetized, and heterogeneity complications are best accommodated in multidimensional analysis. Hirschberg et al. (1991) and Hirschberg et al. (1998) identified similar indicators, and collected them into distinct clusters which could represent the dimensions worthy of distinct treatment in multidimensional frameworks.

In this research effort we have considered the role of air quality indicators in the context of economic and welfare life quality indicators, using artificial neural networks (ANN). Therefore in this presentation we have obtained the key variables (life expectancy, healthy life years, infant mortality, Gross Domestic Product (GDP) and GDP growth rate) and developed a neural network model to predict the air quality outcomes (emissions of sulphur and nitrogen oxides). Sustainability and quality of life indicators have been proposed recently by Flynn et al. (2002) and life quality indices have been used to estimate willingness to pay (Pandey & Nathwani, 2004). The innovative part of this research effort lies in the use of a soft computing machine learning approach like the ANN to predict air quality. In this way, we introduce the reader to a technique that allows the comparison of various attributes that impact the quality of life in a meaningful way.

2. Materials and methods

It is well known that the quality of the air in a locale influences the health of the population and ultimately affects other dimensions of that population's welfare and its economy. As a simple example, in cities where pollution levels rise significantly in the summer, worker absenteeism rates rise commensurately and productivity is adversely impacted. Other dimensions of the economy are influenced on "high pollution days" as well. For example, when outdoor leisure activity is restricted this may have serious consequences for the service sector of the economy (Bresnahan et al., 1997). In this chapter, we have introduced two measures of environmental quality or air quality as quality of life factors. A feature of these indices is the fact that these types of pollution are created by some of the very activities that define economic development. The two factors under investigation here are sulfur oxides (SOx) and nitrogen oxides (NOx) (million tones of SO_2 and NO_2 equivalent, respectively). Sulphur oxides, including sulphur dioxide and sulphur trioxide, are reported as sulphur dioxide equivalent, while nitrogen oxides, including nitric oxide and nitrogen dioxide, are reported as nitrogen dioxide equivalent. They are both produced as byproducts of fuel consumption as in case of the generation of electricity. Vehicle engines also produce a large proportion of NOx. SOx is primarily produced when high sulphur content coal is burned which is usually in large-scale industrial processes and power generation. Thus, the ratio of these emissions to the population is an indication of pollution control.

The following attributes of QOL have been used:

- Life expectancy at birth: The mean number of years that a newborn child can expect to live if subjected throughout his life to the current mortality conditions (age specific probabilities of dying).
- Healthy life years: The indicator Healthy Life Years (HLY) at birth measures the number of years that a person at birth is still expected to live in a healthy condition. HLY is a health expectancy indicator which combines information on mortality and morbidity. The data required are the age-specific prevalence (proportions) of the population in healthy and unhealthy conditions and age-specific mortality information. A healthy condition is defined by the absence of limitations in functioning/disability. The indicator is also called disability-free life expectancy (DFLE). Life expectancy at birth is defined as the mean number of years still to be lived by a person at birth, if subjected throughout the rest of his or her life to the current mortality conditions (WHO, 2010).
- Infant mortality: The ratio of the number of deaths of children under one year of age during the year to the number of live births in that year. The value is expressed per 1000 live births.
    - Gross Domestic Product (GDP) per capita: GDP is a measure of the economic activity, defined as the value of all goods and services produced less the value of any goods or services used in their creation. These amounts are expressed in PPS (Purchasing Power Standards), i.e. a common currency that eliminates the differences in price levels between countries allowing meaningful volume comparisons of GDP between countries.

- GDP growth rate: The calculation of the annual growth rate of GDP volume is intended to allow comparisons of the dynamics of economic development both over time and between economies of different sizes. For measuring the growth rate of GDP in terms of volumes, the GDP at current prices are valued in the prices of the previous year and the thus computed volume changes are imposed on the level of a reference year; this is called a chain-linked series. Accordingly, price movements will not inflate the growth rate.

Data were extracted for 34 European countries, for the year 2005, from the Eurostat database (Eurostat, 2010). Descriptive statistics for all variables are given in **Table 1**.

Statistics	Emissions of sulphur oxides (million tones of SO_2 equivalent)	Emissions of nitrogen oxides (million tones of NO_2 equivalent)	Infant Mortality	GDP (Purchasing Power Standards, PPS)	GDP Growth Rate	Life Expectancy At Birth (years)	Healthy Life Years
Valid N*	34	34	34	33	33	33	27
Missing**	0	0	0	1	1	1	7
Mean	0.503	0.372	5.721	95.921	4.206	77.535	60.448
Std. Deviation	0.648	0.482	4.227	46.620	2.521	3.244	5.443
Min	0.00	0.00	2.30	28.50	0.70	70.94	50.10
Max	2.37	1.63	23.60	254.50	10.60	81.54	69.30

*Number of observations (countries) for each variable.
**Number of countries that didn't had available data.

Table 1. Descriptive statistics for all variables used in the analysis.

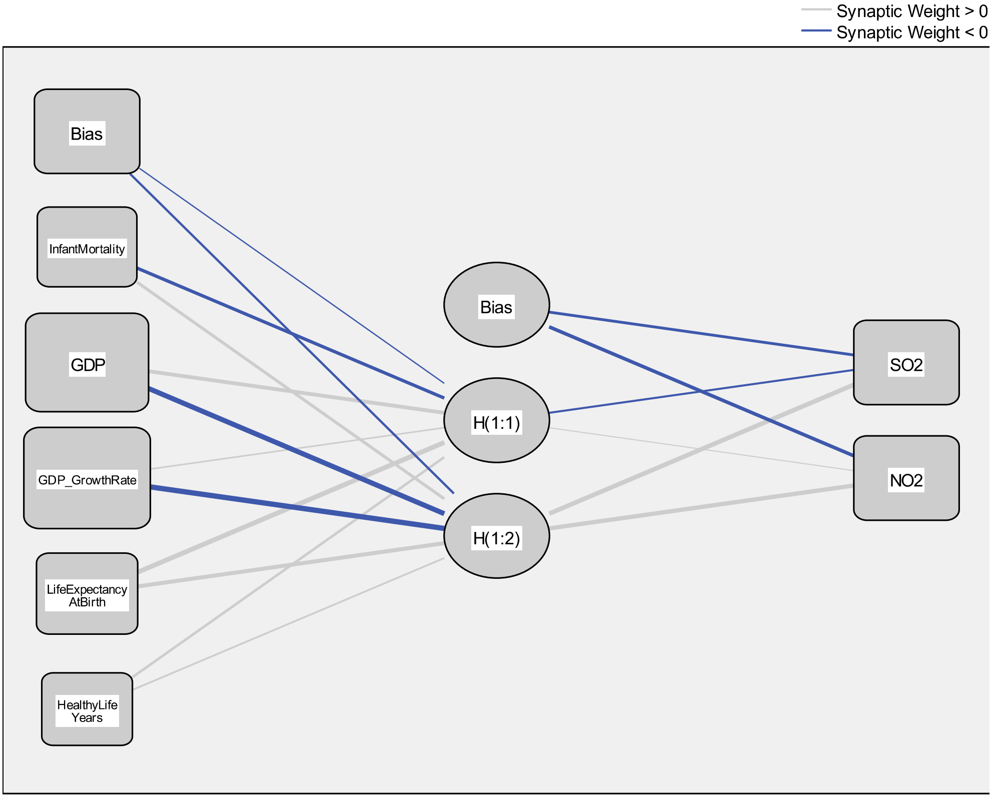
For the performance of the analyses, multi-layer perceptron (MLP) and radial-basis function (RBF) network models were developed under the SPSS v.19 statistical package (IBM, 2010). We specified that the relative number of cases assigned to the training:testing:holdout samples should be 6:2:1. This assigned 2/3 of the cases to training, 2/9 to testing, and 1/9 to holdout. For the MLP network we employed the back propagation (BP) optimization algorithm. As it is well known in BP the weighted sum of inputs and bias term are passed to the activation level through the transfer function to produce the output (Bishop, 1995; Fine, 1999; Haykin, 1998; Ripley, 1996). The sigmoid transfer function was employed (Callan, 1999; Kecman, 2001), due to the fact that the algorithm requires a response function with a continuous, single valued with first derivative existence (Picton, 2000).

Before using the input data records to the ANN a normalization process took place so that the values with wide range do not prevail over the rest. The autoscaling approach was applied. This method outputs a zero mean and unit variance of any descriptor variable (Dogra, Shaillay, 2010). Thus, each feature's values were normalized based on the following equation:

$$Z_i=(X_i-\mu_i)/\sigma_i$$

where X_i was the *i*th parameter, Z_i was the scaled variable following a normal distribution and σ_i, μ_i were the standard deviation and the mean value of the *i*th parameter.

These networks were trained in an iterative process. A single hidden sub layer architecture was followed in order to reduce the complexity of the network, and increase the computational efficiency (Haykin, 1998). Two units were chosen in the hidden layer. The schematic representation of the neural network is given in **Fig. 1**.

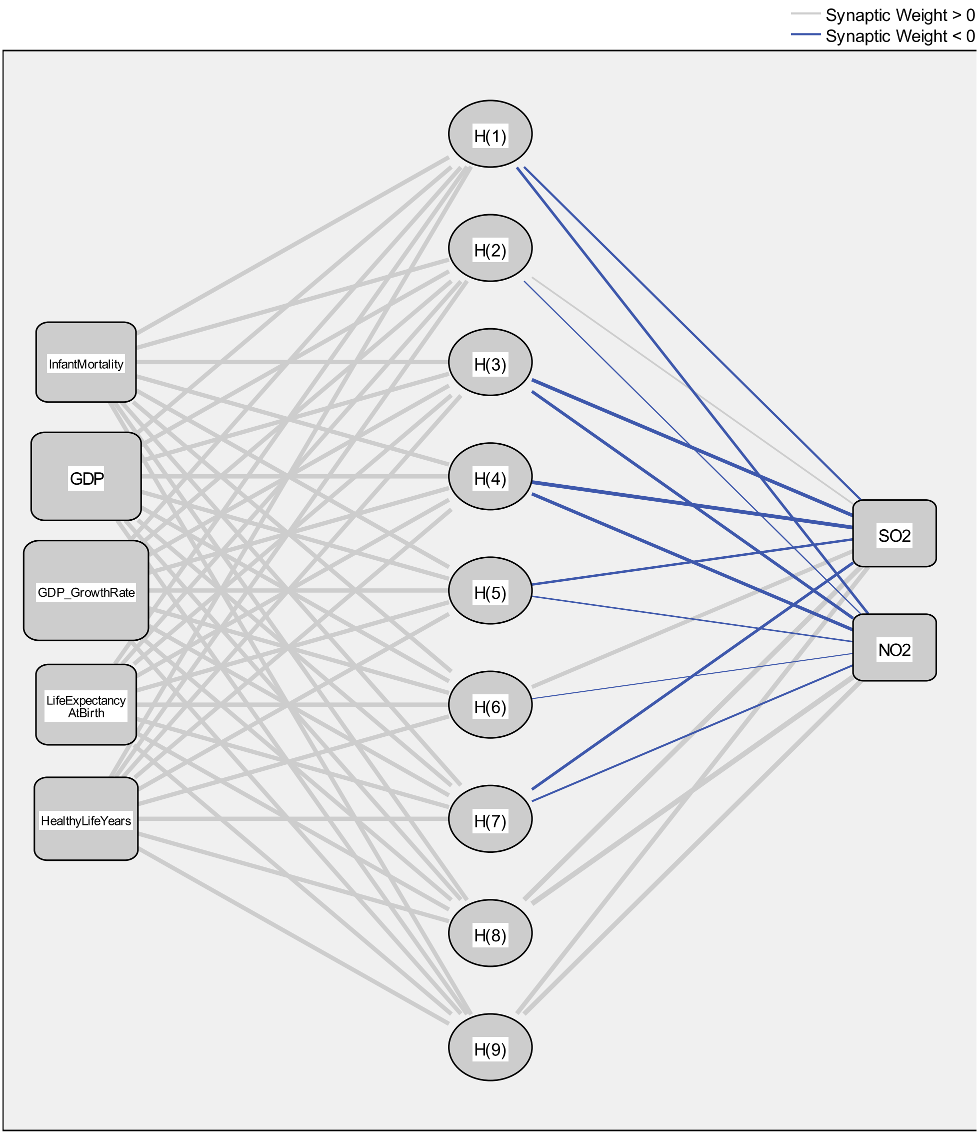


Hidden layer activation function: Sigmoid

Output layer activation function: Identity

Fig. 1. Multi-layer perceptron network structure.

As regards the RBF network (Bishop, 1995; Haykin, 1998; Ripley, 1996; Tao, 1993; Uykan et al., 2000), the architecture that was developed included nine neurons in the hidden layer. The transfer functions (hidden layer activation functions and output function) determine the output by depicting the result of the distance function (Bors & Pitas, 2001; Iliadis, 2007). The schematic representation of the neural network with transfer functions is given in **Fig. 2.**



Hidden layer activation function: Softmax

Output layer activation function: Identity

Fig. 2. Radial-basis function network structure.

3. Results – Discussion

From the MLP analysis, 19 cases (70.4%) were assigned to the training sample, 2 (7.4%) to the testing sample, and 6 (22.2%) to the holdout sample. The choice of the records was done in a random manner. The whole effort targeted in the development of an ANN that would have the ability to generalize as much as possible. The seven data records which were excluded from the analysis were countries that did not had available data on Healthy Life Years. Two units were chosen in the hidden layer.

Table 2 displays information about the results of training and applying the MLP network to the holdout sample. Sum-of-squares error is displayed because the output layer has scale-dependent variables. This is the error function that the network tries to minimize during training. One consecutive step with no decrease in error was used as stopping rule. The relative error for each scale-dependent variable is the ratio of the sum-of-squares error for the dependent variable to the sum-of-squares error for the "null" model, in which the mean value of the dependent variable is used as the predicted value for each case. There appears to be more error in the predictions of emissions of sulphur oxides than in emissions of nitrogen oxides, in the training and holdout samples.

The average overall relative errors are fairly constant across the training (0.779), testing (0.615), and holdout (0.584) samples, which give us some confidence that the model is not overtrained and that the error in future cases, scored by the network will be close to the error reported in this table

Training	Sum of Squares Error		14.029
	Average Overall Relative Error		0.779
	Relative Error for Scale Dependents	Emissions of sulphur oxides (million tones of SO_2 equivalent)	0.821
		Emissions of nitrogen oxides (million tones of NO_2 equivalent)	0.738
Testing	Sum of Squares Error		0.009
	Average Overall Relative Error		0.615
	Relative Error for Scale Dependents	Emissions of sulphur oxides (million tones of SO_2 equivalent)	0.390
		Emissions of nitrogen oxides (million tones of NO_2 equivalent)	0.902
Holdout	Average Overall Relative Error		0.584
	Relative Error for Scale Dependents	Emissions of sulphur oxides (million tones of SO_2 equivalent)	0.603
		Emissions of nitrogen oxides (million tones of NO_2 equivalent)	0.568

Table 2. MLP Model Summary.

In the following **Table 3** parameter estimates for input and output layer, with their corresponding biases, are given.

Predictor		Predicted			
		Hidden Layer 1		Output Layer	
		H(1:1)	H(1:2)	SO_2	NO_2
Input Layer	(Bias)	-0.119	-0.537		
	Infant Mortality	-0.805	0.752		
	GDP	1.033	-3.377		
	GDP Growth Rate	0.318	-3.767		
	Life Expectancy At Birth	1.646	1.226		
	Healthy Life Years	0.567	0.358		
Hidden Layer 1	(Bias)			-0.635	-0.877
	H(1:1)			-0.518	0.116
	H(1:2)			1.396	1.395

Table 3. MLP Parameter Estimates.

Linear regression between observed and predicted values ($\mathrm{SO}_2 = a + b\hat{\mathrm{SO}}_2 + error$, $\mathrm{NO}_2 = a + b\hat{\mathrm{NO}}_2 + error$) showed that the MLP network does a reasonably good job of predicting emissions of sulphur and nitrogen oxides. Ideally, linear regression parameters *a* and *b* should have values 0 and 1, respectively, while values of the observed-by-predicted chart should lie roughly along a straight line. Linear regression gave results for the two output variables $\mathrm{SO}_2 = 0.114 + 0.918\hat{\mathrm{SO}}_2 + error$ (**Fig. 3**) and $\mathrm{NO}_2 = 0.005 + 1.049\hat{\mathrm{NO}}_2 + error$ (**Fig. 4**), respectively. There appears to be more error in the predictions of emissions of sulphur oxides than in emissions of nitrogen oxides, something that we also pointed out in Table 2. **Figs 3 and 4** actually seem to suggest that the largest errors of the ANN are overestimations of the target values.

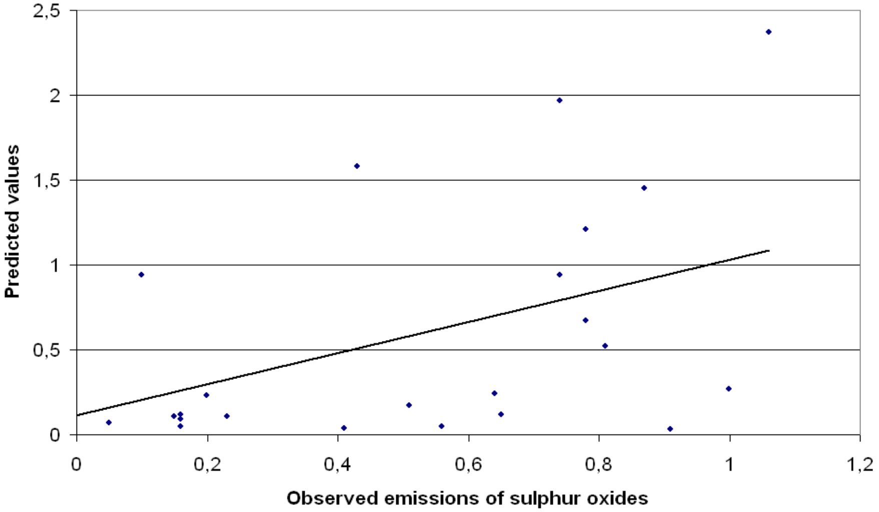


Fig. 3. Linear regression of observed values for emissions of sulphur oxides by predicted values of MLP.

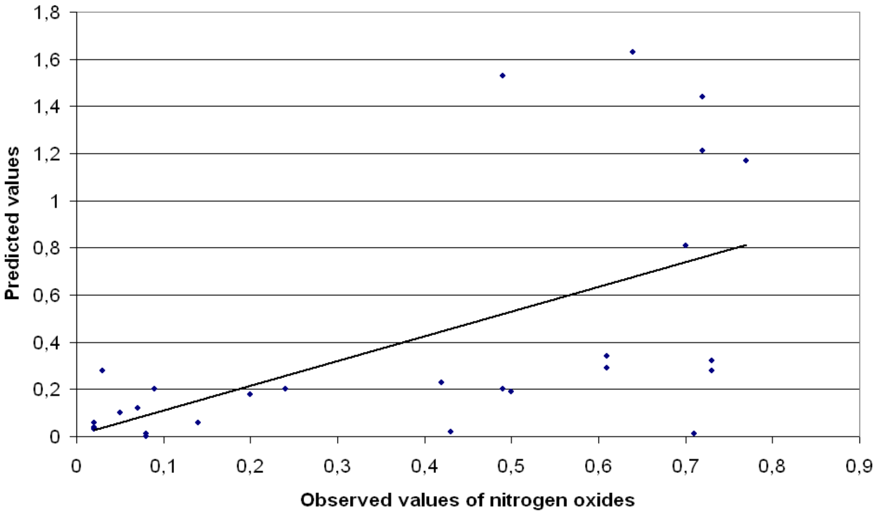


Fig. 4. Linear regression of observed values for emissions of nitrogen oxides by predicted values of MLP.

The importance of an independent variable is a measure of how much the network's model-predicted value changes for different values of the independent variable. A sensitivity analysis to compute the importance of each predictor is applied. The importance chart (**Fig. 5**) shows

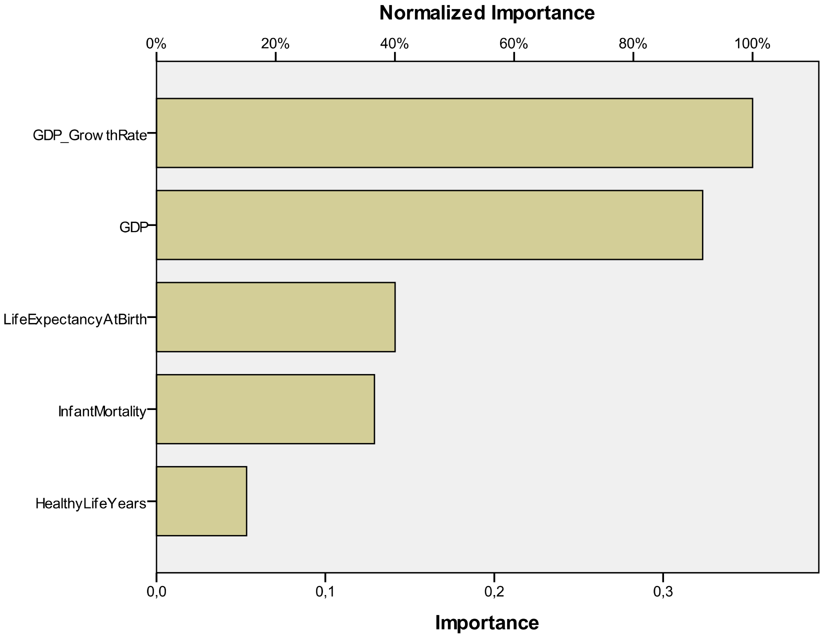


Fig. 5. MLP independent variable importance chart.

that the results are dominated by GDP growth rate and GDP (strictly economical QOL indicators), followed distantly by other predictors.

From the RBF analysis, 19 cases (70.4%) were assigned to the training sample, 1 (3.7%) to the testing sample, and 7 (25.9%) to the holdout sample. The seven data records which were excluded from the MLP analysis were excluded from the RBF analysis also, for the same reason.

Table 4 displays the corresponding information from the RBF network. There appears to be more error in the predictions of emissions of sulphur oxides than in emissions of nitrogen oxides, in the training and holdout samples.

The difference between the average overall relative errors of the training (0.132), and holdout (1.325) samples, must be due to the small data set available, which naturally limits the possible degree of complexity of the model (Dendek & Mańdziuk, 2008).

Training	Sum of Squares Error		2.372
	Average Overall Relative Error		0.132
	Relative Error for Scale Dependents	Emissions of sulphur oxides (million tones of SO_2 equivalent)	0.161
		Emissions of nitrogen oxides (million tones of NO_2 equivalent)	0.103
Testing	Sum of Squares Error		0.081
	Average Overall Relative Error		[a]
	Relative Error for Scale Dependents	Emissions of sulphur oxides (million tones of SO_2 equivalent)	[a]
		Emissions of nitrogen oxides (million tones of NO_2 equivalent)	[a]
Holdout	Average Overall Relative Error		1.325
	Relative Error for Scale Dependents	Emissions of sulphur oxides (million tones of SO_2 equivalent)	1.347
		Emissions of nitrogen oxides (million tones of NO_2 equivalent)	1.267

[a] Cannot be computed. The dependent variable may be constant in the training sample.

Table 4. RBF Model Summary.

In **Table 5** parameter estimates for input and output layer are given for the RBF network.

		Predicted										
		Hidden layer									Output layer	
Predictor		H(1)	H(2)	H(3)	H(4)	H(5)	H(6)	H(7)	H(8)	H(9)	SO_2	NO_2
Input Layer	Infant Mortality	1.708	1.517	-1.064	-1.279	-0.562	-0.491	-0.276	-0.204	-0.132		
	GDP	-1.092	-0.986	3.098	0.451	0.667	-0.714	-0.101	-0.076	0.161		
	GDP Growth Rate	1.572	-0.164	0.575	1.390	-0.448	0.924	-0.544	-0.720	-1.212		
	Life Expectancy At Birth	-1.710	-1.640	0.500	1.169	0.578	-0.611	0.211	0.820	0.461		
	Healthy Life Years	-1.245	-1.223	0.497	1.161	1.123	-0.111	-0.346	0.868	-0.831		
Hidden Unit Width		0,606	0.363	0.363	0.363	0.645	0.363	0.576	0.363	0.363		
Hidden Layer	H(1)										-0.552	-0.668
	H(2)										0.463	-0.395
	H(3)										-0.813	-0.773
	H(4)										-0.833	-0.795
	H(5)										-0.617	-0.401
	H(6)										0.970	-0.253
	H(7)										-0.718	-0.547
	H(8)										3.116	3.429
	H(9)										2.698	2.790

Table 5. RBF Parameter Estimates.

Linear regression between observed and predicted values ($\mathrm{SO}_2 = a + b\,\hat{\mathrm{SO}}_2 + error$, $\mathrm{NO}_2 = a + b\,\hat{\mathrm{NO}}_2 + error$) showed that the RBF network does also a reasonably good job of predicting emissions of sulphur and nitrogen oxides. Linear regression gave results for the two output variables $\mathrm{SO}_2 = -0.0114 + 0.8583\,\hat{\mathrm{SO}}_2 + error$ (**Fig. 6**) and $\mathrm{NO}_2 = -0.026 + 0.7932\,\hat{\mathrm{NO}}_2 + error$ (**Fig. 7**), respectively. In this case, it is difficult to see if there is more error in the predictions of emissions of sulphur or nitrogen oxides.

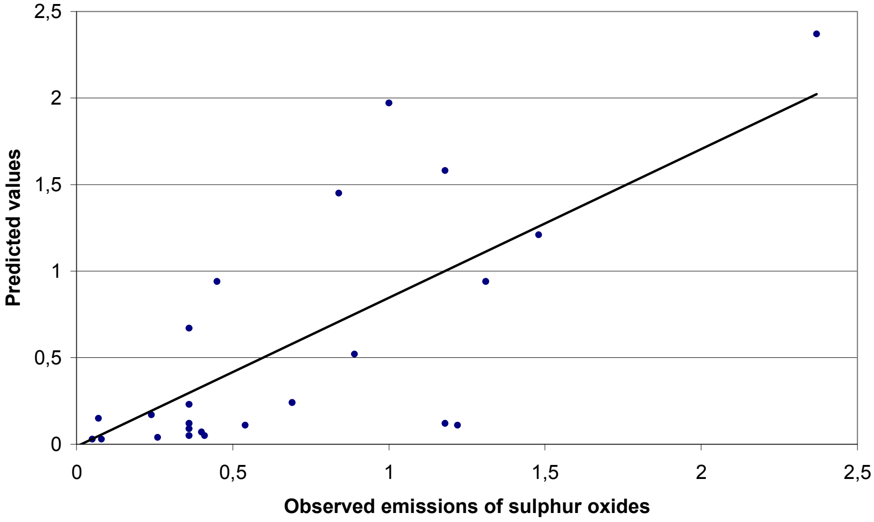


Fig. 6. Linear regression of observed values for emissions of sulphur oxides by predicted values of RBF.

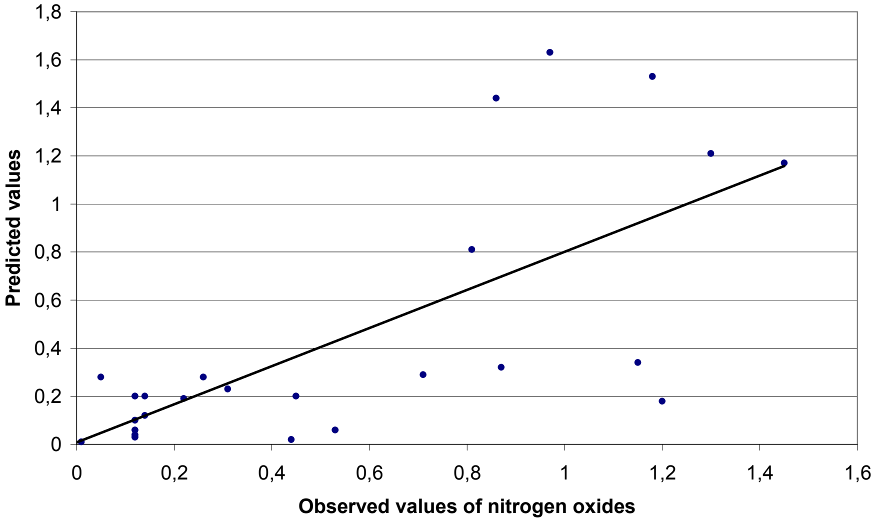


Fig. 7. Linear regression of observed values for emissions of nitrogen oxides by predicted values of RBF.

Finally, the importance chart for the RBF network (**Fig. 8**) shows that, once again, GDP growth rate and GDP are the most important predictors of sulphur and nitrogen oxides emissions.

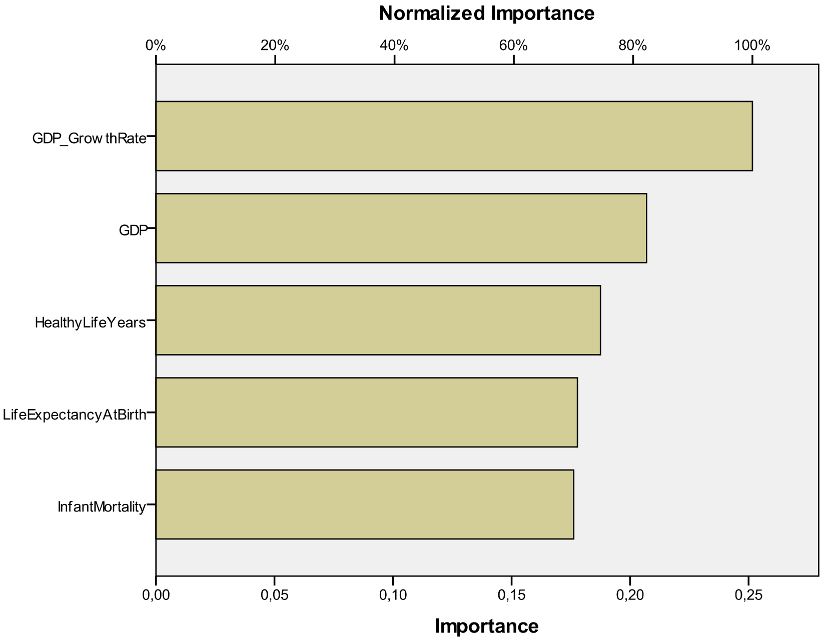


Fig. 8. RBF Independent variable importance chart.

4. Conclusions

The multi-layer perceptron and radial-basis function neural network models, that were trained to predict air quality indicators, using life quality and welfare indicators, appear to perform reasonably well. Unlike traditional statistical methods, the neural network models provide dynamic output as further data is fed to it, while they do not require performing and analyzing sophisticated statistical methods (Narasinga Rao et al., 2010).

Results showed that GDP growth rate and GDP influenced mainly air quality predictions, while life expectancy, infant mortality and healthy life years followed distantly. One possible way to ameliorate performance of the network would be to create multiple networks. One network would predict the country result, perhaps simply whether the country increased emissions or not, and then separate networks would predict emissions conditional on whether the country increased emissions. We could then combine the network results to likely obtain better predictions. Note also that neural network is open ended; as more data is given to the model, the prediction would become more reliable. Overall, we find that predictors that include economic indices may be employed by investigators to represent dimensions of air quality that include, as well as go beyond, these simple indices.

5. References

Bishop, C. (1995). *Neural Networks for Pattern Recognition, 3rd ed*. Oxford University Press, Oxford.

Bors, A., Pitas, I. (2001). Radial Basis function networks In: Howlett, R., Jain, L (eds.). *Recent Developments in Theory and Applications in Robust RBF Networks*, 125-153 Heidelberg, NY, Physica-Verlag.

Bresnahan, B., Mark, D., Shelby, G. (1997). Averting behavior and urban air pollution. *Land Economics* 73, 340–357.

Callan, R. (1999). *The Essence of Neural Networks*. Prentice Hall, UK .

Dendek, C., Mańdziuk, J. (2008). *Improving Performance of a Binary Classifier by Training Set Selection*. Warsaw University of Technology, Faculty of Mathematics and Information Science, Warsaw, Poland.

Dogra, Shaillay, K. (2010). *Autoscaling*. QSARWorld - A Strand Life Sciences Web Resource. http://www.qsarworld.com/qsar-statistics-autoscaling.php

Eurostat. (2010). http://epp.eurostat.ec.europa.eu.

Fine, T. (1999). *Feedforward Neural Network Methodology, 3rd ed*. Springer-Verlag, New York.

Flynn P., Berry D., Heintz T. (2002). Sustainability & Quality of life indicators: Towards the Integration of Economic, Social and Environmental Measures. *The Journal of Social Health* 1(4), 19-39.

Haykin, S. (1998). *Neural Networks: A Comprehensive Foundation, 2nd ed*. Prentice Hall, UK.

Hirschberg, J., Esfandiar, M., Slottje, D. (1991). Cluster analysis for measuring welfare and quality of life across countries. *Journal of Econometrics* 50, 131–150.

Hirschberg, J., Maasoumi, E., Slottje, D. (1998). *A cluster analysis the quality of life in the United States over time*. Department of Economics research paper #596, University of Melbourne, Parkville, Australia.

IBM. (2010). SPSS Neural Networks 19. SPSS Inc, USA.

Iliadis, L. (2007). *Intelligent Information Systems and Applications in Risk Management*. Stamoulis editions, Thessaloniki, Greece.

Kecman, V. (2001). *Learning and Soft Computing*. MIT Press, London.

Koller, M., Lorenz, W. (2003). Survival of the quality of life concept. *British Journal of Surgery* 90(10), 1175-1177.

Maasoumi, E. (1998). Multidimensional approaches to welfare. In: Silber, L. (ed.). *Income Inequality Measurement: From Theory to Practice*. Kluwer, New York.

Pandey, M., Nathwani, J. (2004). Life quality index for the estimation of social willingness to pay for safety. *Structural Safety* 26(2), 181-199.

Picton, P. (2000). *Neural Networks, 2nd ed*. Palgrave, New York.

Narasinga Rao, M., Sridhar, G., Madhu, K., Appa Rao, A. (2010). A clinical decision support system using multi-layer perceptron neural network to predict quality of life in diabetes. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews* 4, 57–59.

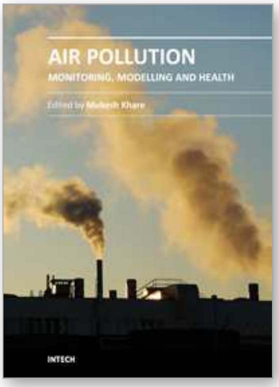
Ripley, B. (1996). *Pattern Recognition and Neural Networks*. Cambridge University Press, Cambridge.

Slottje, D., Scully, G., Hirschberg, J., Hayes, K. (1991). *Measuring the Quality of Life Across Countries: A Multidimensional Analysis*. Westview Press, Boulder, CO.

Tao, K. (1993). A closer look at the radial basis function (RBF) networks. In: Singh, A. (ed.). *Conference Record of the Twenty-Seventh Asilomar Conference on Signals, Systems, and Computers*. IEEE Computational Society Press, Los Alamitos, California .

Uykan, Z., Guzelis, C., Celebi, M., Koivo, H. (2000). *Analysis of input-output clustering for determining centers of RBFN*. IEEE Transactions on Neural Networks 11, 851-858.

WHO: World Health Organization. (2010). http://www.who.int.



Air Pollution - Monitoring, Modelling and Health
Edited by Dr. Mukesh Khare

ISBN 978-953-51-0424-7
Hard cover, 386 pages
Publisher InTech
Published online 23, March, 2012
Published in print edition March, 2012

Air pollution has always been a trans-boundary environmental problem and a matter of global concern for past many years. High concentrations of air pollutants due to numerous anthropogenic activities influence the air quality. There are many books on this subject, but the one in front of you will probably help in filling the gaps existing in the area of air quality monitoring, modelling, exposure, health and control, and can be of great help to graduate students professionals and researchers. The book is divided in two volumes dealing with various monitoring techniques of air pollutants, their predictions and control. It also contains case studies describing the exposure and health implications of air pollutants on living biota in different countries across the globe.

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821