A multi-layer perceptron neural network to predict air quality through indicators of life quality and welfare

Kyriaki Kitikidou[1], Lazaros Iliadis[1]

[1]Democritus University of Thrace, Department of Forestry and Management of the Environment and Natural Resources, Pandazidou 193, 68200, Orestiada, Greece
<kkitikid@fmenr.duth.gr>

Abstract. This paper considers the similarity between two measures of air pollution/quality control, on the one hand, and widely used indicators of life quality and welfare, on the other. We have developed a multi-layer perceptron neural network system which is trained to predict the measurements of air quality (emissions of sulphur and nitrogen oxides), using Eurostat data for 34 countries. We used life expectancy, healthy life years, infant mortality, Gross Domestic Product (GDP) and GDP growth rate as a set of inputs. Results were dominated by GDP growth rate and GDP. Obtaining accurate estimates of air quality measures can help in deciding on distinct dimensions to be considered in multidimensional studies of welfare and quality of life.

Keywords: Air quality, Economic welfare, Artificial Neural Networks, Multi-Layer Perceptron, Quality Of Life

1 Introduction

Quantitative analysis of quality of life (QOL) across countries, and the construction of summary indices for such analyses have been of interest for some time [15]. 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) [11] 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) [8] and Hirschberg et al. (1998) [9] 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) [6] and life quality indices have been used to estimate willingness to pay [12]. 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.

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 [2]. In this paper, 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). 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 sulfur 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 attributes of quality of life used in this paper are the following:

- 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.
- 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 1 000 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,

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 [4]. Descriptive statistics for all variables are given in **Table 1**.

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

	Emissions of sulphur oxides (million tones of SO_2 equivalent)	Emissions of nitrogen oxides (million tones of NO_2 equivalent)	Infant Mortality	GDP (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

For the performance of the analysis, multi-layer perceptron (MLP) network models were used, employing 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 ([1], [5], [7], [14]). The sigmoid transfer function was employed ([3], [10]), due to the fact that the algorithm requires a response function with a continuous, single valued with first derivative existence [13]. These networks were trained in an iterative process. The number of hidden layers is chosen to be only one to reduce the network complexity, and increase the computational efficiency [7]. The schematic representation of the neural network is given in **Fig. 1**.

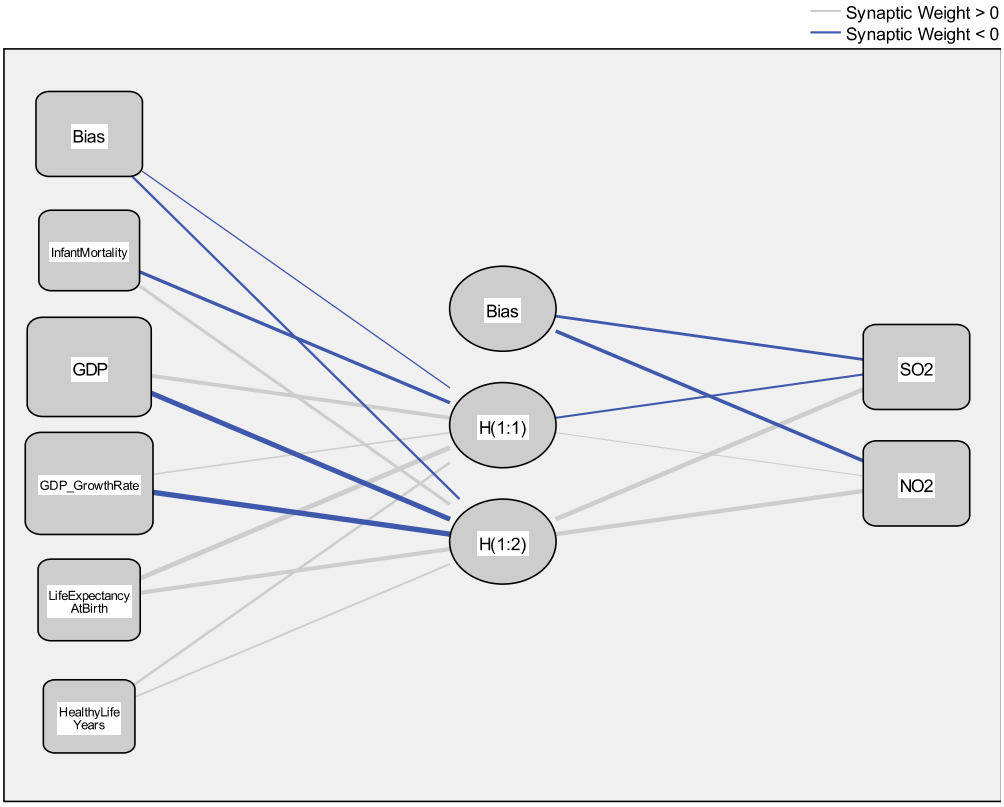

Fig. 1. Multi-layer perceptron network structure.

3 Results-Discussion

From the 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 final 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.

Table 2. Model Summary.

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

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

Table 3. Parameter Estimates.

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

Linear regression between observed and predicted values ($SO_2 = a + b\overset{\Omega}{SO}_2 + error$, $NO_2 = a + b\overset{\Omega}{NO}_2 + error$) showed that the 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 $SO_2 = 0.114 + 0.918\overset{\Omega}{SO}_2 + error$ (**Fig. 2**) and $NO_2 = 0.005 + 1.049\overset{\Omega}{NO}_2 + error$ (**Fig. 3**), 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 2 and 3** actually seem to suggest that the largest errors of the ANN are overestimations of the target values.

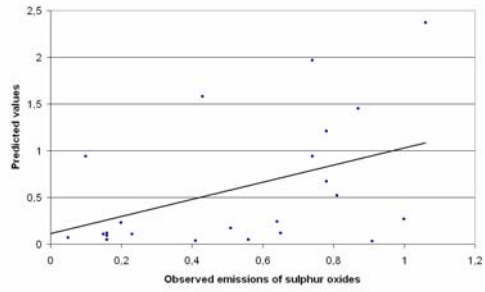

Fig. 2. Linear regression of observed values for emissions of sulphur oxides by predicted values.

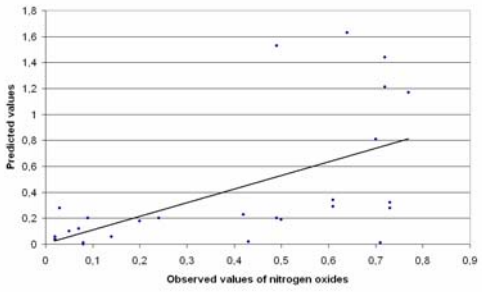

Fig. 3. Linear regression of observed values for emissions of nitrogen oxides by predicted values.

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

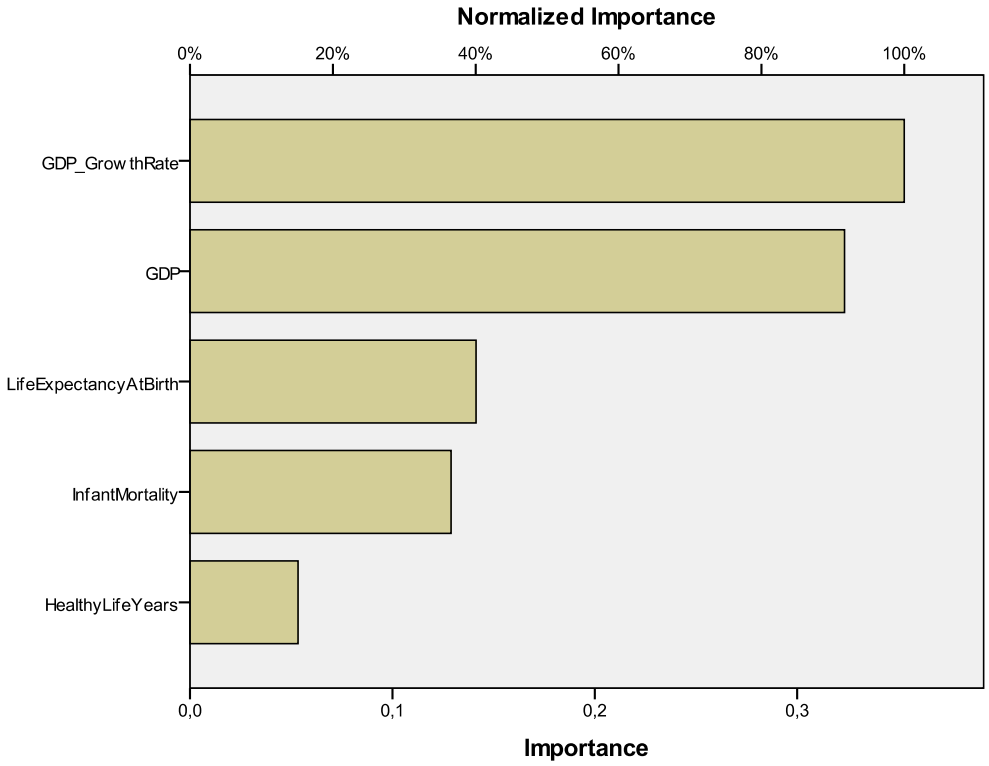

Fig. 4. Independent variable importance chart.

4 Conclusions

The multi-layer perceptron neural network model, that was trained to predict air quality indicators, using life quality and welfare indicators, appears to perform reasonably well. 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.

References

1. Bishop, C.: Neural Networks for Pattern Recognition, 3rd ed. Oxford University Press, Oxford (1995)
2. Bresnahan, B., Mark, D., Shelby, G.: Averting behavior and urban air pollution. Land Economics 73, 340–357 (1997)
3. Callan, R.: The Essence of Neural Networks. Prentice Hall, UK (1999)
4. Eurostat, http://epp.eurostat.ec.europa.eu
5. Fine, T.: Feedforward Neural Network Methodology, 3rd ed. Springer-Verlag, New York (1999)
6. Flynn P., Berry D., Heintz T.: Sustainability & Quality of life indicators: Towards the Integration of Economic, Social and Environmental Measures. The Journal of Social Health 1(4), 19-39 (2002)
7. Haykin, S.: Neural Networks: A Comprehensive Foundation, 2nd ed. Prentice Hall, UK (1998)
8. Hirschberg, J., Esfandiar, M., Slottje, D.: Cluster analysis for measuring welfare and quality of life across countries. Journal of Econometrics 50, 131–150 (1991)
9. Hirschberg, J., Maasoumi, E., Slottje, D.: A cluster analysis the quality of life in the United States over time. Department of Economics research paper #596, University of Melbourne, Parkville, Australia (1998)
10. Kecman, V.: Learning and Soft Computing. MIT Press, London (2001)
11. Maasoumi, E.: Multidimensional approaches to welfare. In: Silber, L. (ed.). Income Inequality Measurement: From Theory to Practice. Kluwer, New York (1998)
12. Pandey, M., Nathwani, J.: Life quality index for the estimation of social willingness to pay for safety. Structural Safety 26(2), 181-199.
13. Picton, P.: Neural Networks, 2nd ed. Palgrave, New York (2000)
14. Ripley, B.: Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge (1996)
15. Slottje, D., Scully, G., Hirschberg, J., Hayes, K.: Measuring the Quality of Life Across Countries: A Multidimensional Analysis. Westview Press, Boulder, CO (1991)