# ANALYSIS OF A NEURO-FUZZY APPROACH OF AIR POLLUTION: BUILDING A CASE STUDY

## Ciprian-Daniel NEAGU[1], Daniela NEAGU[2], Lucian GEORGESCU[3], Vasile PALADE[1]

[1]*Department of Computer Science & Eng., "Dunarea de Jos" University of Galati*
[2]*Department of Fishing and Aquaculture, "Dunarea de Jos" University of Galati*
[3]*Department of Chemistry, "Dunarea de Jos" University of Galati*

Abstract: This work illustrates the necessity of an Artificial Intelligence (AI)-based approach of air quality in urban and industrial areas. Some related results of Artificial Neural Networks (ANNs) and Fuzzy Logic (FL) for environmental data are considered: ANNs are proposed to the problem of short-term predicting of air pollutant concentrations in urban/industrial areas, with a special focus in the south-eastern Romania. The problems of designing a database about air quality in an urban/industrial area are discussed. First results confirm ANNs as an improvement of classical models and show the utility of ANNs in a well built air monitoring center.

Keywords: artificial neural networks, fuzzy logic, air quality prediction.

## 1. INTRODUCTION

Since the beginning of the industrial age, concentrations of greenhouse gases (GHG) have been increasing substantially. Combustion of fossil energy, land uses changes and, in recent years, the use of fluorocarbons, are the main activities responsible for this increase. Actual concentrations are about 25 % greater than those at the beginning of the Industrial Revolution. If current trends continue, concentrations will double from pre-industrial levels before the end of this century and, if unchecked, continue to rise thereafter (IPCC, 1998). A warming trend of about 0.5°C has been detected over the last 150 years (Houghton, 1994). In particular, the 80s and early 90's of the XX century were among the warmest on the record, with 1995 heading the higher values ever registered on temperature (MetOffice, 1997).

Air pollution became a priority issue for the authorities (WHO 1987) and, at the same time, subject of many research studies. A critical factor, contributing to the creation of the urban smog in urban/ industrial areas, is the high population density associated with intense polluting anthropogenic activities: plants, chemical factories, industrial vehicles. The proposed case study will focus on the approaches and proposals to improve and adapt the study of southeastern Romanian area to the new techniques applied in air quality prediction. In the majority part of recent national and international documents about air pollution, the region of south Romania is listed as affected from industrial effluents (EEA 2000, MAPPM 2000).

## 2. PREMISES OF STUDY

One of the main objectives of this work was to evaluate the local area input scenario on the air quality as an urban/industrial region in Romania, where the weather affects the air quality as well as local industry output (MAPPM 2000). The chosen approach was based on both, meteorological parameters study, and photochemical pollutants measurements. The neuro-fuzzy performed models have to estimate the photochemical pollutant $NO_2$. Relevant factors in such case are the local

This paper was recomended for publication by S.Bumbaru

topography, poor ventilation, the frequent appearance of river breeze and strong northern winds, and finally the high levels of solar irradiation with the subsequent formation of photochemical pollutants from the transformation of the emitted oxides of nitrogen and hydrocarbons. The main pollution sources in the city area are automobile traffic, industry and central heating. On a yearly basis, traffic is considered to be the main source of carbon monoxide (CO), nitrogen oxides ($NO_X$), black smoke (BS) and hydrocarbons, while industry is the main source for sulfur dioxide ($SO_2$) and suspended particles (MAPPM 2000). But if only the cold period of the year (November-April) is considered, central heating becomes the most important pollution source of $SO_2$, while it has the same importance in smoke emissions as car traffic, which is clearly the most important source of smoke for the rest of the year. In all the listed reports, south-eastern urban/ industrial area is listed as extremely polluted in photochemical compounds and suspended particles.

Photochemical air pollution is a major environmental problem in contemporary large cities. This is a phenomenon involving a series of chemical reactions, triggered by solar radiation. Nitric oxide and nitrogen dioxides ($NO_x$), ozone ($O_3$), and the energy that comes through a specific range of the spectrum of solar radiation are the elements of the photochemical reaction. The photochemical pollutants $NO_2$ and $O_3$ have significantly increased the last few years.

### 3. RELATED WORK

In order to foreseen the regional impacts of changes on climate system it is necessary to apply numerical algorithms to estimate the relative area climate variability. The horizontal resolution of present coupled atmosphere-urban/industrial models is still too coarse to capture the effects of local and regional forcing in areas of complex surface shape and to provide information suitable for many impact assessment studies (IPCC, 1996). Downsizing the spatial resolution of the models thus constitutes a major challenge for climate modelers and different techniques were developed (IPCC, 1998): (i) empirical approaches, (ii) semi-empirical approaches, and (iii) modeling technique. Within the modeling, just in the last years, Artificial Intelligence techniques began to be applied. Until now, several studies have been made and research papers have been published discussing the role the artificial intelligence tools could play in predicting photochemical pollution. In (Lee 1995), prediction of the atmospheric ozone concentrations using neural networks was proposed. A machine learning approach was used in (Sucar et al, 1997) to predict ozone pollution in Mexico City, and in (Avouris 1995, Neagu et al 2001) for applied on Athens city. Studies on the usage of neural networks for short-term air pollution have been presented in (Boznar

1997, Sucar et al, 1997). In (Avouris et al., 1997) was made a comparative study using three algorithms coming from three active trends of AI: a case-based reasoning adapted algorithm, a fully connected multi-layer perceptron (MLP), and a common inductive decision tree approach using the Oblique Classifier (OC1, Murthy et al., 1994).

More related work examples exist dealing with the daily maximum temperature forecasting using machine learning (Abdel et al., 1996), and with a combined set of meteorological, social and industrial factors (Lekkas et al., 1994). In all cases, the neural network approach of the photochemical pollution is restricted to feedforward architectures. More, given the fact that the dangerous peak values of pollutants (high and medium pollution) are rare in all the data sets, comparing with the number of usual low pollution cases, the studies cited above give good results just for low values. The models performance deteriorates in the case of peak and medium values.

The failure in predicting especially medium values of the pollutant by all the related models is due to the fact that the algorithms cannot distinguish clearly low-level cases from the medium and higher ones. It is observed that cases present in the data set are as rare as their pollutant value is higher. This is an important reason to develop a neurosymbolic architecture based on specialized modules in order to combine the experience acquired from learning data sets and the explicit fuzzy rules given by human experts. The main reason is that, our databases have just few and noisy data, so the architectures will be strongly generalizing, without an efficient feedback.

### 4. BUILDING THE CASE STUDY

The specific problem is three-dimensional, and refers to how $NO_2$ pollutant concentrations are evolving in space and time. A similar problem is related to $SO_2$ and CO air pollution, caused, for example, around large thermal power plants. Photochemical pollution has significantly increased during the last few years, so there is a growing concern by the authorities, for management of air pollution. Air quality monitoring organizations, for the specific problem described above, are focused on two main predictive tasks: $NO_2$ and $O_3$ peak, both short term daily prediction as well as long term prediction.

The predictive process done by human experts is taking into account the ultimate meteorological data, as well as measured values of pollutants (NO, $O_3$ and $NO_2$). Therefore, an efficient short-term air pollution prediction model would help the air quality monitoring organization to survey and propose emergency measures. These restraining measures are important to reduce the power of plants or traffic density in the area. The predicted value of pollutants is important both, to estimate the pollution in an area

where there are no measurements, as well as to be distributed to local authorities, and local radio and TV stations, in order to make announcements to the public and take preventive actions.



Fig.1. The urban area of Galati: the end-point of photochemical pollutants prediction.

The specific entries used in the application are similar to those described by (Avouris et al., 1997) as specific to the air pollution monitoring organization in Athens (PERPA), and taken form the Agency of Environmental Protection reports, and (MAPPM 2000). Additional data have been provided by the Romanian National Meteorological Service, called INMH. The inputs relate to ultimate meteorological data (rain, inversion, solar radiation, wind direction, wind speed), and pollutant concentrations (NO, $O_3$, $NO_2$). A part of factors were computed according to the heuristic functions proposed in (Avouris et al., 1997), representing the contribution of each feature to the $NO_2$ evolution. The values are collected once by day, in order to predict the $NO_2$ peak value for the rest of the day, so the current database is rare and noisy.

In order to test the applicability of the neurosymbolic approach to the photochemical pollution problem, we ought to gather and process the data sets first. The cleansing of the data has been a necessary step towards high efficiency, since there is always a high pay off for noise or uncertainty in the data sets. The errors of the input data can be classified into the measurement and sampling errors (Zickus 1999). Systematic measurement errors are caused by faults in the measurement instruments. The systematic sampling errors occur due to influence of specific local micrometeorological conditions, for example of the systematic sampling error can be the wind speed alteration or higher pollutant levels due to closer located emission sources. However, in the present study we are more interested in changes in parameter values, than in absolute levels, since this type of systematic error does not play important role in data analysis. Measurement uncertainties, as quoted by

the device manufacturers, are about 5% of measured value.

The other factor that influences air pollution, e.g. emissions, also is a random element, due to unpredictability of human activities. More, using normalized and fuzzified inputs and outputs and neural representation assures a minimization of the effect of these errors through the entire data set. The pollution database covers a ten-years period (1990 to 1999) and is based on the averaged values taken from (MAPPM 2000) and National Committee for Statistics for the period 1990-1999 (CNS 2000). In order to make a useful dataset from this database, we decided to include in the final data set only the features that were more relevant to $NO_2$ values, and the specific fuzzy sets were proposed in a form as appropriate to technical vocabulary of human experts as possible. We will explain thereafter how each factor has been represented and computed.

### 4.1. The Meteorological Factors

These factors were computed according to how "favorably" each feature contributed to the $NO_2$ episode evolution: each factor has been designed taking into account the results of the statistical analysis on each correspondent attribute, and advises of the field experts. Hence, the meteorological parameters influencing pollutant dispersion are the recorded precipitation temperature inversion, rain factor, and solar radiation levels (gathered by the INMH), wind speed and direction (recorded by APM). The data sets for the meteorological factors have to be normalized before processing, by dividing their values with each factor's maximum value as described in Table 1. The rule according to which the rain factor has been computed (see Table 2), follows:

```
IF RainHeight>0.5 THEN RainToday=1 ELSE
RainToday=0,
```

where 1 stands for rain, 0 stands for lack of rain.

The temperature inversion plays a favorable role to the evolution of pollution, acting like a trap for the pollutants and blocking them into the lower part of the atmosphere. This phenomenon causes high concentrations of the photochemical pollutants, mainly when no winds blow at the same time, and it is favorable to cause an episode (Avouris et al., 1997).

The wind speed database during the period of analysis has to be processed. Anyway, the strongest winds blew from the Northeast directions, while the prevailing wind direction was from the Southwest and Southeast sectors, according to the rivers direction. In fact, due to the geography of the Galati basin, the winds coming from south directions are more favorable towards a $NO_2$ episode. That is due to the lower temperature of those winds that can lead to

temperature inversion. The wind predictions provided by APM and INMH are categorical values defining the approximate direction and speed of winds, as they are evolving in the period of the corresponding weather bulletin. The linguistic variables considered for meteorological inputs are characterized by the term sets:
`RainToday={YES, NO},`

```
InversionToday={L0, L1, L2, L3},
SolarRadiationAt13={LOW, MED, HIGH},
WindDirection={NE, S, NV},
WindSpeed={LOW, MED, HIGH}.
```

The fuzzy shapes of the normalized values of the wind direction factor, considered as a meteorological linguistic variable, are presented in Fig. 2a.

Table 1. The normalization values for meteorological factors.

| RainToday | InversionToday | SolRadAt13 | WindDirect | WindSpeed |
|---|---|---|---|---|
| 1 | 3 | 4 | 360 | 12 |

Table 2. The values of the rain factor.

| | Precipitation prediction | | | | | | |
|---|---|---|---|---|---|---|---|
| RainToday Factor | Nil | Possible | Light | Snow | Rain | Periodical | Storm |
| 0 | Ï | Ï | Ï | Ï | | | |
| 1 | | | | | Ï | Ï | Ï |

### 4.2. The pollutant factors

Nitrogen oxides in the human respiratory system cause increases in lung infections and asthma, and long-term exposure can also weaken the effectiveness of the lung's defenses against bacterial infection (WHO 1987). Nitrogen oxides and ozone are precursors to ozone and acidic precipitation (Finlayson-Pitts et al., 1986) and take part in the photochemical transformations in the urban atmosphere, therefore it is necessary to examine their dynamics together. Nitrogen oxides $NO_x$ are released into the atmosphere mainly in the form of nitric oxide, as a product of the reaction of NO with $O_2$ during high temperature combustion of the fossil fuel.

The input and the output data set values are normalized and fuzzified with respect of the maximum observed values of pollutant concentrations: 700 µg/m$^3$ ($NO_2$), 1300 µg/m$^3$ (NO), and 130 µg/m$^3$ ($O_3$). The fuzzy shapes of the normalized values of the pollutant considered as linguistic variables are presented in Fig. 2b, while their term sets are:
```
NO2={L1, L2, L3, L4},
NO={LOW, MED, HIGH},
O3={LOW, MED, HIGH}.
```

We defined four levels of concentration for $NO_2$: level 1 (0-200 µg/m$^3$), level 2 (200-350 µg/m$^3$), level 3 (350-500 µg/m$^3$), and level 4 (500-700 µg/m$^3$).
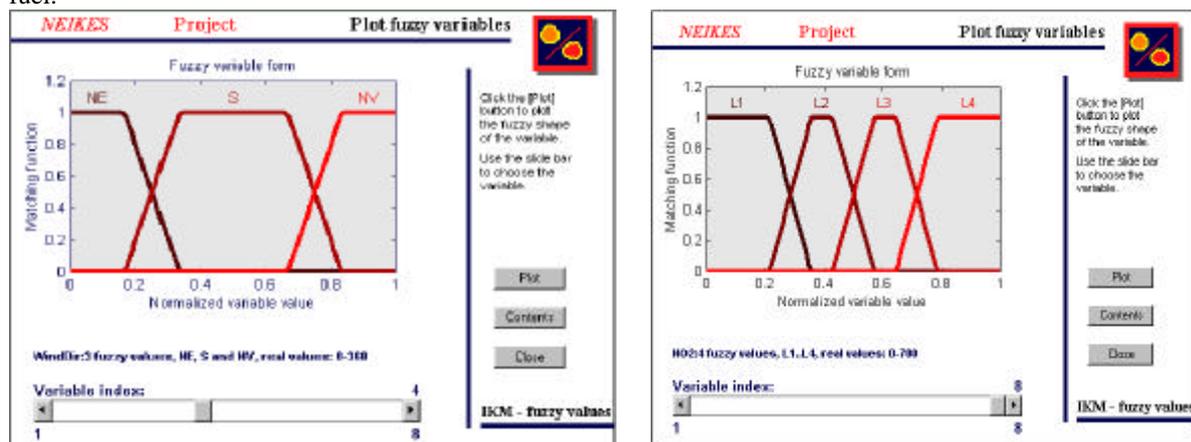


Fig.2. Values of the linguistic variables: (a) WindDirection; (b) $NO_2$.

## 5. NEUROSYMBOLIC KNOWLEDGE-BASED MODEL FOR AIR QUALITY PREDICTION

The idea of using a neurosymbolic approach as a tool for air pollution prediction came both, from neural networks' abilities to learn and generalize from sets of historical patterns, and the complexity of the problem, which is more suitable to be modeled using

fuzzy techniques (Fuller 1999). More, given the fact that human experts are already using some empirical rules with imprecise and linguistic forms, a separate module is including such rules as specific structured neural networks. This kind of neural networks, with fuzzy inputs and output, are equivalent to the rule set, assuring a homogenous structure model as described in (Neagu et. al, 2001). Depending on the methods

used to combine the outputs of all processing modules in order to predict the overall output of the system (NO2after10), three different structures of a neurosymbolic system-based approach were proposed (Neagu et. al, 2001) and tested on air quality databases.

### 5.1. The Implicit Knowledge Structure

The implicit knowledge structure includes a MAPI-based HNN and an MLP-based network, used as a second IKM component integrated into neuro-fuzzy architecture (Neagu et al., 2001). The first structure is a three-layered network with the eight above described fuzzy inputs and one fuzzy output, the predicted value of $NO_2$ pollutant. The number of hidden neurons parameterizes the neuro-fuzzy network structure. The neuro-fuzzy models will be included into the structure after the database building process will give a satisfactory large collection of data. Currently, this process is in an updating step. This approach would take advantage of predictive capabilities of ANN and gives reasons to explain the output and the patterns discovered by the IKM part of the proposed system, as well as highlights and adjusts some explicit rules given by human experts.

### 5.2. The Explicit Knowledge Structure

The rules, acquired from human experts, are represented as HNN (as described in Neagu et al., 2001). These rules were expressed in a fuzzy form and cover the output domain for premises based on the inputs. The EKM is build as a compact structure, following to be used in FEM or UGN integration procedures (Neagu et al., 2001). The structure of EKM is based on the stand-alone EKM$_i$, $i$=1,2,...,7, built (Fuller 1999) on Fire Each Rule method (Fig.3).

```
Rule 1:
IF (WindSpeed_10 is HIGH)
THEN (NO2After10 is L1)
Rule 2:
IF (RainToday is YES)
AND(SolRadAt13 is HIGH)
AND(WinDirect_10 is S)
AND(WindSpeed_10 is LOW)
AND(NO2_10 is L3)
THEN (NO2After10 is L4)
Rule 3:
IF (WindSpeed_10 is LOW)
AND(NO2_10 is L3)
THEN (NO2After10 is L3)
Rule 4:
IF (WinDirect_10 is S)
AND(WinSpeed is LOW)
AND(O3_10 is HIG)
AND(NO2_10 is L4)
THEN (NO2After10 is L4)
Rule 5:
IF (RainToday is YES)
AND(InversionToday is L2)
AND(WinSpeed_10 is LOW)
```

```
AND(O3_10 is LOW)
THEN (NO2After10 is L2)
Rule 6:
IF (RainToday is YES)
AND(WindDirect_10 is S)
AND(WinSpeed_10 is LOW)
AND(NO2_10 is L2)
THEN (NO2After10 is L3).
Rule 7:
IF (WinSpeed_10 is LOW)
AND(O3_10 is LOW)
AND(NO_10 is HIGH)
AND(NO2_10 is L2)
THEN (NO2After10 is L4)
```

The EKM structures come with the advantage of easy acquisition and implementation (as long as the explicit knowledge is expressed in terms of fuzzy logic), simplicity of the neuro-fuzzy structures used to implement it. There is no additional computation effort other than the one step feedforward inference.
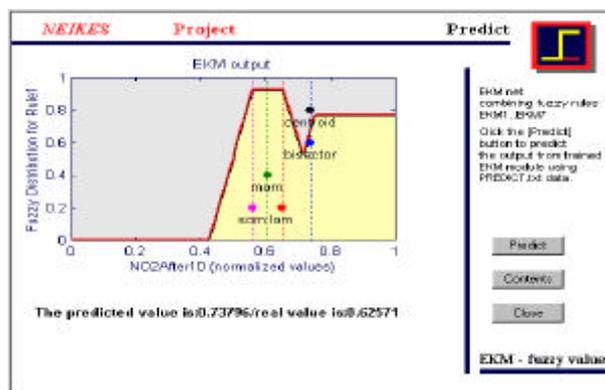


Fig.3. The output of the Explicit Knowledge Module for the given input data set.

## 6. CONCLUSIONS AND FUTURE WORK

In the present work it has been shown that some problems difficult to solve by traditional approaches can be addressed with ANNs: in particular neuro-fuzzy structures can be used both to model and to forecast air pollutant concentrations in urban areas as functions of some chemical variables and of local meteorological parameters. Results obtained confirm the utility of similar neural architectures for predicting air quality pollutant concentrations in industrial and in urban areas. Neuro-fuzzy structures are suitable for investigations working on large data sets and for problems in which the inputs and corresponding output values are known but the relationships between the inputs and the outputs are difficult to understand with usual analysis techniques.

These conditions are commonly found in many air quality applications, owing to the relationships between physical and chemical processes in the environmental systems. ANNs do not eliminate the preparatory work on the data sets: indeed the data, before being processed, need a very careful statistical

analysis to discover links between variables in order to avoid redundant patterns and improve accuracy. The approach described could be a component for an air monitoring network implementation in the urban/industrial area (such is Galati) designed to evaluate regulatory programs in improving ambient air quality. For example, ANNs could be utilized for controlling urban air quality pollution by regulation of traffic flows during severe weather conditions. ANNs were confirmed as a computational approach improving classical models in which the solution is learned from a set of examples. It is important to note that ANNs are pattern recognition diagnostic techniques, with some predictive skill, allowing assimilation of high quantities of data.

Finally, the following considerations can be drawn: a) ANNs can provide considerable support in conducting quality controls on experimental data; b) results can be better when data from geographically distributed areas are available; c) in order to replace missing data or to assess outliers, it is advisable to conduct series training and forecasting on the same sets; d) the weights and the structures provide useful insight into the most important forecasting variables and their relevant links; e) ANNs are particularly suitable whenever a large quantity of data is available (since networks train by example, the more data they have experienced, the better the network will perform) and whenever there is no simple solution by traditional technology.

The study is strongly dependent by the quality, and amount of databases which has to be raised up as a proposal to local environmental protection agencies and other administrative official data sources.

## 7. REFERENCES

Abdel, R.E., Elhadidy, M.A. (1996) Modeling and forecasting the daily maximum temperature using abductive machine learning. *Oceanographic Literature Review*, **43(1)**, pp.25-37.

Avouris, N.M, Page, B., (1995), *Environmental Informatics*, Kluwer Academis, Dordrecht.

Avouris, N.M., Kalapanidas, E. (1997), Expert Systems and AI Techniques in Air Pollution Prediction, in *Mathematical Modelling of Atmospheric Pollution*, (N. Moussiopoulos, ed.), Thessaloniki,.

Boznar, M., (1997), Pattern Selection Strategies for a Neural Network-Based Short Term Air Pollution Prediction Model, in Procs. of the IEEE International Conf. on Intelligent Information Systems, IIS'97, pp.340-344.

CNS (Comisia Nationala pentru Statistica, 2000), Rapoartele anuale ale Comisiei Nationale pentru Statistica, http://www.cns.ro.

EEA (European Environment Agency, 2000), Emissions of atmospheric pollutants in Europe, 1980-1996, Topic report No 9/2000.

Finlayson-Pitts, B.J., Pitts, J.N., *Atmospheric Chemistry: Fundamentals & Experimental Techniques*, Wiley: NY, 1986.

R. Fuller, *Introduction to Neuro-Fuzzy Systems*, Advances in Soft Computing Series, Springer-Verlag, 1999.

M. Funabashi, A. Maeda, "Fuzzy and Neural Hybrid Expert Systems: Synergetic AI", IEEE Expert, pp. 32-40, 1995.

Houghton, J. (1994), *Global warming: the complete briefing*, Lion Publishing, Oxford.

IPCC (1996), Climate Change 1995, In: The Science of Climate Change, (J.T.Houghton, L.G. Meira Filho, B.A. Callander, N. Harris, A. Kattenberg, K. Maskell), Contribution to the 2nd Assessment Report of the Intergovernmental Panel on Climate Change, CambridgeUniversity Press 1996.

IPCC (Intergovernmental Panel on Climate Change, 1998), The Regional Impacts of Climate Change, An Assessment of Vulnerability, (R.T. Watson, M.C. Zinyowera, R.H. Moss, D.J. Dokken, eds.). Published for the Intergovernmental Panel on Climate Change, Cambridge University Press.

Lee, S.S., (1995), Predicting atmospheric ozone using neural networks as compared to some statistical methods, in Procs. of IEEE Technical Applications Conference and Workshops, NORTHCON'95, pp. 309-316.

Lekkas, G.P., Avouris, N.M., Viras, L.G., (1994), Case-Based Reasoning in Environmental Monitoring Applications, *Applied Artificial Intelligence*, 8(3), pp. 359-376,

MAPPM (Ministerul Apelor si Protectiei Mediului din Romania, 2000), *Date privind starea mediului 2000*: www.mappm.ro/WebTest/rom/oriz2/StareaMediului.

Met.Office (1997), Climate change and its impacts: a global perspective, December 1997.

Neagu C.D., Avouris, N., Kalapanidas, E., Palade, V., (2001), Neurosymbolic Integration in a Knowledge-based System for Air Quality Prediction, *Applied Intelligence* (accepted), Kluwer.

Murthy, S.K., Kasif, S., Salzberg, S., (1994), A System for Induction of Oblique Decision Trees, *Journal of Artificial Intelligence Research*, **2**, pp.1-32.

Sucar, L.E., Perez-Brito, J., Ruiz-Suarez, J.C., Morales, E. (1997), Learning Structure from Data and Its Application to Ozone Prediction, *Applied Intelligence*, **7**, pp. 327-338.

WHO (World Health Organization) 1987. Air quality guidelines for Europe, WHO: Copenhagen.

Zickus, M., (1999), Influence Of Meteorological Parameters On The Urban Air Pollution And Its Forecast, PhD Thesis, University of Vilnius.