SHORT-TERM AIR QUALITY PREDICTION USING A CASE BASED CLASSIFIER

Elias Kalapanidas and Nikolaos Avouris*
University of Patras
Electrical and Computer Engineering Department
GR-265 00 Rio Patras, Greece
email: {ekalap , N.Avouris } @ ee.upatras.gr

* Corresponding Author

Abstract

In the frame of air quality monitoring of urban areas, the task of short-term prediction of key-pollutants concentrations is a daily activity of major importance. Automation of this process is desirable, but development of reliable predictive models with good performance, to support this task in operational basis presents many difficulties. In this paper we present and discuss the NEMO prototype that has been built in order to support short-term prediction of NO_2 maximum concentration levels in Athens, Greece. NEMO is based on a case-based-reasoning approach combining heuristic and statistical techniques. The process of development of the system, its architecture and its performance, are described in this paper. NEMO performance is compared with that of a back propagating neural network, and a decision tree (CART). The overall performance of NEMO makes it a good candidate to support air pollution experts in operational conditions.

Keywords: Short-term NO_2 concentration prediction, Case-Based Reasoning (CBR), Urban air quality, Athens, Air Monitoring Operational data modelling, Air Quality Management Operational Centre

1. INTRODUCTION

Investigation of the suitability of various alternative Artificial Intelligence techniques in the context of short-term air pollution prediction has been the subject of research for some time now. In a study reported in (Kalapanidas 1999), a quantitative comparison of Case-based reasoning, neural networks and decision trees has been performed, using operational data from the Athens Air Quality Operation Centre (AQOC) in Greece. The results of our investigation have been obtained during the design of the new AQOC for the Athens Metropolitan Area, but they can also be useful for other researchers and practitioners involved with development of such Centres. Athens, the Greek capital, has experienced many air pollution episodes during the last years. Athens Metropolitan area is situated in a basin less than 100 Km wide, in which more than 4 million people live and where more than 50% of Greek industrial and commercial activity takes place. Levels of photochemical pollution, especially NO_2 and O_3, have significantly increased during the last years, so prediction of the maximum concentration of them in a daily basis is a prime concern for the environmental experts. In the Athens Operations Centre, daily 12 hour and 24-hour forecast of the main photochemical pollutants, NO_2 and O_3 need to be issued to the public and the authorities concerned. Until now, the predictive process has been based on human expertise, as in most cities of the world. The decisions had been based on pollutant concentration measurements, meteorological measurements and forecasts, as well as other related information, such as traffic in the area, and social factors modelling sources behaviour.

A need has emerged recently for automation of the process. This is envisaged to take place in the frame of an integrated system, under design, that could maintain all necessary data, estimate the episode dangers for certain pollutants, predict short-term and long-term trends, and propose counter- measures when required.

In the frame of a typical Operational Centre, a *fast short-term prediction* is needed in a day-by-day basis, which should determine the likelihood of an air pollution episode. In case of positive answer, a more accurate prediction procedure has to follow, based on complex mathematical models. Using, as input, the prognostic meteorological fields on a continuous basis, the fast prediction model should provide estimates on the level of the air pollutants for the following 12, 24 or 48 hours.

The problem of NOx and other major pollutant concentration prediction, using heuristics and other AI techniques, has been the aim of many researchers during the last years. This is intensified since urban air pollution have become an issue in many metropolitan areas of the world like Los Angeles, Mexico City, Athens etc. (Breiling 1995). The tools used so far in operational basis, are based on various mathematical and computational modelling techniques (Bartzis 1995), which however usually require complex input data and considerable computational resources, so they are not suitable for the fast prediction task. On the other hand statistical forecasting techniques (e.g. regression models or neural networks) as well as expert systems (Simon, 1995), (Avouris 1995), have been proposed as fast prediction tools. In particular, several research papers have been published discussing the role that neural networks and AI techniques can play in predicting photochemical pollution: Lee (Lee 95) tried predicting atmospheric ozone levels using neural networks. This has also been the aim of Perantonis et al. (Perantonis 94) in experimental basis. Both papers reported good results. A similar approach was that of Ruiz-Suarez applied on Mexico city atmospheric data (Ruiz-Suarez et al. 94, 95), and by Yi (Yi 96) over an industrialised urban area, who also contacted research on ozone short-term prediction. Mlakar and Boznar (Mlakar 94, Boznar 95) have presented a study on the usage of neural networks for short-term air pollution prediction. Related work examples exist, dealing with the daily maximum temperature forecasting using machine learning approaches (Abdel 96). However these approaches are hard-wired to the air pollution conditions of model building time, since they do not contain automatic mechanisms for adjusting the developed computational models to the long-term changes which characterise air pollution of a typical urban area, so it is expected that their performance will deteriorate in the long run .

The approach reported in this paper is that of a Case-Based Reasoning system, which uses examples of previous similar incidents in order to classify the expected levels of maximum concentration of the current day. A characteristic of such a system is that it is constantly fed with new cases, so it is expected to adapt its behaviour to the long-term changes of air pollution.

For development of the system a long data modelling process was necessary, followed by testing using data sets from the Athens operational Centre. The originality of the experiments described, is mainly attributed to the fact that the data sets contain commonly available data, coming from monitoring networks and therefore the conducted experiments aim at testing use of the techniques in operational conditions.

In the next chapter we discuss data modelling and transformation issues, which are a key preliminary phase for building any Case-Based Reasoning System.

2. OPERATIONAL AIR POLLUTION DATA MODELLING

At the Athens air quality-monitoring Centre, two main predictive tasks are undertaken daily. These relate with prediction of daily peak concentrations of NO_2 and O_3 for the same day (12 hour forecast), and the next day (24 hour forecast). Until now, human experts have exclusively performed these predictive tasks. The predictive decisions were based upon the most recent meteorological data, as well as upon other relevant information, such as traffic in the area, and the behaviour of emission sources (the later constituting to a so-called social factor).

As for NO_2, Operational Centre authorities have defined four levels of pollution, based on maximum mean hourly measurement of at least one monitoring station in the area, shown in Table 1.

Table 1: Pollution levels based on NO_2 concentrations for the Athens area

In the frame of the reported research, a prototype has been developed to predict these indices, based exclusively on the currently daily available data in the Athens Operational Centre. The Air Quality Monitoring Authority for the area, has been our main source of data.

This public organisation maintains an air-quality monitoring network consisting of 12 monitoring stations in the Athens basin, equally distributed throughout the area. Measurements recorded by this network are the hourly mean values of NO, NO_2, CO_2, SO_2, O_3 concentrations, wind speed, wind direction, temperature, and humidity. Additional data have come from the National Meteorological Service (NMS), which provide area meteorological forecast, plus information related with conditions in the upper atmosphere, temperature inversion information etc.

We extracted data concerning future wind speed and direction, temperature, and temperature inversion below 150 m over the surface from NMS prediction bulletins. To this data, the recorded precipitation and solar radiation levels, gathered by the National Observatory of Athens have been added. These correspond to data that at run time are inserted as observations of the experts related to general meteorological conditions in the area.

These data were inserted into a *pollution data warehouse*, covering a two-year time period. Several problems though, regarding the validity and the integrity of these data, still existed. In order to make a useful data set from this database, pollution experts were contacted in order to build adequate abstractions of the raw monitoring data. Using this approach, incomplete or noisy data are handled as well. The objective of this stage was to produce a final working data set, which had to drive the prediction algorithms through the training and testing phases, giving the best possible results. In the final data model only the features that were more relevant to NO_2 pollution levels were included, as described in the following.

The features finally included in the data model were the most recent NO and NO_2 mean hourly concentration measurements, prior to prediction time (morning measurements between 7 am and 10 am), and a set of meteorological attributes. The latter were the *wind factor* (a function of the wind speed and wind direction), the *temperature inversion factor*, the *precipitation factor*, and *the solar radiation factor*. These factors were computed according to the heuristic functions that were built, expressing a measure of how “favourably” each feature contributed to the NO_2 episode evolution. The heuristics used for defining these factors were based on experts’ advise and they were stored in the Knowledge Base of our system. Each factor definition was also backed by statistical analysis, the results of which confirmed or adjusted accordingly the field experts’ advice. The defined factors can often have complementary nature. This is because measurements of the same feature coming from various sources are often used in operational conditions. So for instance meteorological data can come from Forecast Bulletins, experts observations, the pollution monitoring network etc. For this reason, in the following discussion reference is made to the data origin, where this case holds.

The definitions of the main abstractions contained in the pollution data model are described next.

2.1 The wind factor

There are two data sources of wind-related data; the field measurements made by the Pollution Monitoring Network and the predictions made by the National Meteorological Service. The wind factor coming from the actual measurements has been a synthesis of wind speed and wind direction. Two different heuristic functions have been defined, shown in figure 1, relating the wind speed and direction to the wind factor (wf). As shown in figure 1, the first curve concerns winds of Eastern, Northern and North-eastern directions, and the other Southern, Western, and South-western directions. The reason for this discrimination is the fact that the latter category of winds is more favourable towards a NO_2 episode. That is due to the lower temperatures of those winds, that can lead to temperature inversion and correspondingly, to higher pollution.

Figure 1:Heuristic definition of wind factor

The wind factor coming from the NMS meteorological forecast is also classified according to wind direction. The wind forecast provided by NMS contains qualitative values concerning direction and speed of winds, as they are evolving in the time window of the corresponding weather bulletin. The corresponding wind factor is deduced according to the following table 2.

Table 2: Meteorological Wind Forecast and corresponding Wind Factor

2.2 The Temperature Inversion factor

The *temperature inversion* plays a favourable role in the evolution of air pollution episodes. This phenomenon causes high concentrations of photochemical pollutants, especially when no wind blows at the same time, and it is more favourable to cause an episode, as the inversion temperature height is lower. The *temperature inversion factor* is a function of the temperature difference and the inversion height, following the definition, contained in Table 3.

Table 3: Heuristic evaluation of Temperature Inversion Factor

2. 3 Rain Factor

Rain and other phenomena are described according to a heuristic *Rain factor*. There are two sources of data, on which this factor can be based. One from the NMS forecast bulletins, and another one from the National Observatory measurements. Both produce a binary-valued rain factor.

As for the later, two precipitation measurements were available concerning the duration and the height of the precipitation. The rule according to which the rain factor has been computed, is the following:

If (rain Height) > 0.5 then (rain factor) = 0 else (rain factor) = 1,

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

The factor produced out of the NMS forecast data, has been computed according to the following table:

Table 4: Heuristic Evaluation of the Rain Factor

The above factors were combined with other pollution measurements into a Pollution Operational Data Model, the schema of which is shown here.

1) Date
2) Code of measurement station
3) NO hourly concentration at 8 a.m. (measurement)
4) NO hourly concentration at 9 a.m. (measurement)
5) NO hourly concentration at 10 a.m. (measurement)
6) NO maximum hourly concentration after 10 a.m.
7) Hour that the NO maximum hourly concentration occurred
8) NO_2 hourly concentration at 8 a.m. (measurement)
9) NO_2 hourly concentration at 9 a.m. (measurement)
10) NO_2 hourly concentration at 10 a.m. (measurement)
11) NO_2 maximum hourly concentration after 10 a.m.

12) Hour that the NO_2 maximum hourly concentration occurred

13) The wind before 10 a.m. (factor deduced from measurement)

14) NMS wind forecast after 10 a.m. (factor deduced from forecast)

15) The wind after 10 a.m. (factor deduced from measurement)

16) Precipitation level from the Observatory(factor deduced from measurement)

17) NMS rain forecast (factor deduced from forecast)

18) Temperature inversion (factor deduced from measurement)

19) Solar radiation at 10 a.m. (factor deduced from measurement)

20) NMS forecast of solar radiation after 10 a.m. (factor deduced from forecast)

21) Solar radiation at 1 p.m.

22) Maximum temperature of the day in Celsius degrees

23) NMS forecast for next day’s winds

24) NMS forecast for next day’s precipitation

25) NMS forecast for next day’s inversion

26) NMS forecast for next day’s solar radiation

27) NMS forecast for next day’s maximum temperature

After this data modelling and transformation phase, a Database was built, implementing the above schema, which was populated with the available data set. In order to minimise the search space, the most relevant attributes were selected, according to the physics of the photochemical phenomenon, as the environmental experts described it. All NOx values between 7 a.m. and 10 a.m. were included, as well as some major meteorological attributes like the wind factor, the inversion factor, the rain factor, and the solar radiation factor. All these attributes were normalised before input, by dividing their values with each field’s maximum value.

The input file finally consists of ten input attributes and one output (NO2DAY), which (the later) is the predicted maximum NO_2 value for the remaining of the day, after 10 a.m.

3. DEVELOPMENT OF NEMO PROTOTYPE

A Case-Based Reasoning (CBR) algorithm uses past knowledge to solve new problems; the basic idea behind CBR is the use of a systematically recorded repository of past cases of the problem to be solved. When the system is presented with a new problem, it retrieves the most similar past cases from the database which subsequently are adapted to the present conditions, in order to provide the new solution (Aamodt 94), (Kolodner 93), (Watson 1996). NEMO development has followed this principle and was based on our past experience of a previous more research-oriented prototype AIRQUOP (Lekkas et al. 94), developed for the same application domain. Figure 2 shows a block diagram of NEMO. All metrics (modification heuristics, similarity metrics) have been replaced by the described above " knowledge base". Due to the nature of the problem, the “test solution” phase is replaced with measurement of success of the prediction when compared with real measurements. At the top level, NEMO consists of three main modules. Two of them deal with retrieval and filtering of cases similar to the new case, while the third one adapts the solutions proposed by the remaining similar cases to form the proposed solution for the new case. Since the attributes of each case are numerical, the indexes are of statistical nature rather than based on a symbolic index, like a vocabulary of words as in other CBR systems. Therefore the retrieval strategy is based on aggregate matching among the cases rather than dimensional matching. According to the later, the attributes of each case are matched one by one, while the former implies the computation of a numerical evaluation function that combines the degree of match along each dimension with a value representing the importance of the dimension (Kolodner 93). A typical flat memory was used to store the system cases, while the search method implemented was a serial search on the case indexes. Due to the lack of internal case explanation, the human intervention is asked during the presentation of the similar cases to discard irrelevant cases that the system has retrieved. While this possibility is foreseen in the algorithm, it has not been used in the experiments described here, in order to concentrate our tests in measuring the algorithm performance without any user intervention.

In order to narrow the search space, some heuristic knowledge has been included in form of rules. Examples of such rules are: “ when strong winds are forecasted for the day under prediction, it is known that the main mass of the photochemical products such as the NO_2 and O_3 would be transported out of the area". Also, "when there is a coincidence of high solar radiation and high ambient temperature, there is a significant possibility of NO_2 episode development".

NEMO in its present version can predict the maximum NO_2 concentration of a previously specified measuring station area. When searching for past cases, the case base consists of the filtered part of the whole database that refers to the specified station.

Figure 2: CBR System (NEMO) block diagram

For each pair of a new case and a past case, a vector made up from the attribute differences is produced. From the differences that are related to the meteorological attributes, the *Weather index* is computed, while from the pollutant emissions related differences, the respective *Pollution index* is computed, as seen at figure 3.

Figure 3: NEMO: Temporal relation of the prediction factors involved

Decision about whether a past case is a relevant case, is taken if the sum of these two major indexes is greater than a pre-defined threshold value. Nevertheless the user, who is a pollution expert, has the power to alter the list of similar cases proposed by the system, removing a case from the list of similar cases found, or adding a new one that is considered to be relevant in predicting the maximum pollutant concentration. Figure 4 presents a part of the system's user interface available during the inspection of a new case.

Figure 4: NEMO: User Interface: Presentation of a parameter on the area map

The similar case adaptation is based on parameter adjustment, where changes in parameters in an old solution are made in response to differences between problem specifications in an old and a new case. This approach is similar to the one used in PERSUATOR (Sycara 87). The adaptation in NEMO is handled by a statistically driven formula, which exploits the fact that the daily NO_2 peak is usually an evolution in time of the morning (until 10 am) NO_2 maximum;

Let the j^{th} case in the case base to be written as X_j where X_j is the vector $(x_1, x_2, ..., x_l, ..., x_m, d_p, d_w, y)_j$.

x_{1j}, x_{2j}, ..., x_{mj} are the m input attributes for the jth case, x_{lj} is the attribute that indicates the maximum morning NO_2 concentration, and y_j is the real value that is to be predicted. If in the process the similar case selector and the filter have retrieved k similar cases for the nth case under prediction, then the value y_n is computed as follows (formula 2.1):

$$y_n = x_{\ln} + \frac{\sum_{i=1}^{k} \left(y_i - x_{li}\right) \cdot \left(\frac{d_p \cdot w_p + d_w \cdot w_w}{w_p + w_w}\right)}{\sum_{i=1}^{k} \left(\frac{d_p \cdot w_p + d_w \cdot w_w}{w_p + w_w}\right)}$$

where d_p is the pollution index between the nth day and its ith day of the similar days in the case base, d_{wj} is the weather (meteorological) index between those two days, W_p and W_w represent the predefined weights for the pollution index and the weather index respectively. The factor

$$\left(\frac{d_p \cdot w_p + d_w \cdot w_w}{w_p + w_w}\right)$$

is used as a closeness measure between the nth and the ith days and behaves as a weighted average, favouring the close days to the more distant ones. Formula (2.1) is a typical similarity function of a CBR system, derived after long experimentation with the test data and taking into consideration the physical meaning of the attributes involved.

The result of experimentation reported here is an evolution of the simpler technique described in (Lekkas 94), resulting in better performance of the NEMO system than the original AIRQUAP CBR system.

3. Fine-tuning the Case-Based Classifier

The NEMO system was trained over the above-described training set. Since the CBR methodology does not include any learning cycle, the training set was used only for parameter adjustment and general data calibration. This involved consecutive runs of the system, during which the system parameters where fine-tuned, until acceptable predictions were obtained. Then the final system was tested using the testing set. Examples of these parameters are C_{max}, C_{min} i.e. the maximum and minimum number of past cases retrieved, the attribute weights W_i, the high and low attribute thresholds, the similarity criteria, the adaptation methods used, etc.

The methodology followed involved, tuning a parameter at a time. In the following table 5, the values of several variables are presented, along with their description, as they have been set at the end of the fine-tuning process.

This has been a long tedious process that determines optimal settings for the NEMO system for the given training data set.

Table 5 Parameters settings for NEMO

4. CLASSIFIER PERFORMANCE

First a description of the testing set is provided. The testing set contains 240 cases (30% of the available data set). The distribution of the NO_2 concentration levels is this data set is shown in Table 6.

Table 6. Testing set classification

It should be stressed at this point that the success criterion for the system should contain a bias towards accurate prediction of the, more rare, pollution episodes (level 4), in order for the system to have any operational value. In other words the predictive accuracy at level 3 and level 4 is more desirable than that of levels 1 and 2. In the following table 7, the NEMO system performance is shown.

Table 7 NEMO System performance for the 240 cases of the Testing Set.

According to the results shown in Table 7, the system was able to make a prediction for 236 cases (98,3%) and made no prediction in 4 cases (1,7%). It predicted correctly the pollution level in 169 cases (70%). However relaxing the success criterion by including the adjacent classes as correct predictions, it missed only 3 cases (level 2 values were predicted as level 4), or 1,2%. Finally according to the level 4 criterion, i.e. correct predictions of level-4 values, in Table 7, only 4 cases (50%) were accurately predicted. However we should take into account the fact that the Level-4 cases which were not predicted, in fact were not missed, but were classified as "no prediction" cases. This is explained by the fact that the system had not enough similar episode cases during the limited period of the two years, to make an accurate prediction.

The performance of the NEMO system was also compared to that of an Artificial Neural Network and a Decision Tree which were built in order to solve the same problem, using the same data set, as described in more detail in (Kalapanidas, 2000). Table 8 contains the overall comparative performance results of the three systems according to the above-discussed criteria.

Table 8 Comparative System performance of NEMO, Decision Tree and Artificial Neural network.

From Table 8, it can be deduced that the NEMO system produces similar results with the other two systems, while the Decision Tree (DT) seems to have a slight overall advantage. However it should be observed that the number of cases is too small, especially for level 4, for this conclusion to have any general validity. However the CB Classifier NEMO presents the important advantage over the other two competing systems, that it is designed to maintain this performance over time and perhaps improve, as the number of stored cases increases. In contrary, the other two techniques present no adaptive behaviour over the time, so they are expected to deteriorate their performance as the air pollution conditions change. It should be added that a difficulty in predicting the second level NO_2 peaks by all 3 models was observed. However this was due to the fact that the algorithms could not distinguish clearly a first level case from a second level one. That implies that the causing circumstances and the related attributes were very similar for these two classes.

5. DYNAMIC PERFORMANCE

During the development of NEMO, the assumption was made that as more cases are processed by the system, the response should improve. This is because as the Case Base is filled with more past cases, the likelihood that the system will find cases with similar characteristics and produce more accurate prediction increases. This assumption leads to the expectation of a measurable improvement of performance. In order to test this assumption the time series of the data set was divided into several consecutive periodes. If the hypothesis were true, then the mean of squares of the difference between the prediction and the actual NO2 maximum concentration of every period should follow a decreasing slope as the period number increases.

If the pair (x $_{i,j}$,y $_{i,j}$) is the jth case of the ith zone, where $x_{i,j}$ is the prediction that NEMO gave and $y_{i,j}$ is the actual outcome, and each time-period contains m_j and m_{j+1} cases respectively, then

$$\frac{\sum_{j=1}^{m_i}\left(x_{i,j}-y_{i,j}\right)}{m} > \frac{\sum_{j=1}^{m_{i+1}}\left(x_{i+1,j}-y_{i+1,j}\right)}{m_{i+1}}$$ should in general be true.

During 3 runs of NEMO, each with the same but randomly shuffled data set, 3 different series of ($x_{i,j}$,$y_{i,j}$) pairs have been produced. We partitioned them into 18 time-periods, containing 44 cases each. For every period, the mean of the difference $(x_{i,j} - y_{i,j})^2$ was computed, as well as the mean per level of $y_{i,j}$, and the error at a 95% of confidence interval for the mean. These indices are shown in a graphical way in the diagrams of figure 5.

Figure 5. Time evolution of NEMO performance

From figure 5 one can deduce a decreasing trend in the time evolution of average error of selected time periods, following a sinusoidal shape downward slope. From this, one can conclude that the assumption holds true, especially in the training set comprising the first 14 periods. However, a larger data set is needed for a more decisive prove. Although there are certain periods, that seem to interrupt the pattern, which can be attributed to abrupt changes of the patterns of meteorology and air pollution indices, as for instance is the case with seasons changes, the charts show a tendency towards decrease of the mean error. So in general it can be observed that even during the limited time represented by the testing data there seem to be indications of system performance improvement.

All cases in the data set used during the reported experiments, were derived from one of the principal Air Pollution Monitoring Stations of central Athens (Patission st.). This has been the Station where most NO_2 episodes were monitored during the past years, and where some of the higher NO_2 concentrations have been measured. So the developed system NEMO related its prediction to a specific monitoring station for which it develops gradually a Case Base. A future application of the predictive NEMO algorithm for a network of monitoring stations means deployment of a set of NEMO modules, each one of which is tied to a specific Monitoring Station. A Unit, that should subsequently perform correlation of the results according to the distribution of the hourly concentrations over that area, should control these modules. This way an overall prediction of the NO_2 levels in the area will be produced.

We assume that a larger data set would present a better distribution of cases, so that more representative episode cases would exist, closer to the episode occurrence.

6. CONCLUSIONS

By interpreting the results of the performed experiments, a conclusion can reached that the NEMO prototype presents characteristics that make it a good candidate for deployment in an Air Quality Monitoring Centre under operational conditions. The conducted experiments showed that in complex and multi-dimensional domains of real world problems, as it is the case with air quality prediction, tree-induction classifiers perform well. While NEMO CBR algorithm performed well during the low NO_2 emissions prediction, it was unable to produce a prediction in some of the high emissions cases. However even in this case the prototype did not make false classification.

The fact that the algorithm has not been tested on a historical database containing a great number of air pollution episodes, which is not available, limits its role to that of a decision support system. The overall performance though of the algorithm can be described as significant, considering that the human experts at the Greek AQOC do not exceed these accuracy levels. This fact makes the system proposed here, useful tool, to be used in operational conditions.

The adaptive nature of CBR algorithms, an intrinsic characteristic of the method, which was proven even in the case of the used data set, is also an important feature of the proposed hybrid architecture. The described system can be easily implemented on a low-cost computing platform. It delivers prediction at run time, unlike other modelling techniques, it does not require a dense monitoring network, and it can deal with noisy data or uncertainty, since the defined abstractions in the data model hide the noise of the raw data. On the other hand, NEMO can not contribute to the understanding of the phenomenon, it produces only a qualitative indication, unable to produce a solid explanation of its response. The system, contrary to other machine learning techniques like ANN and DT, can be adapted to time-evolving problems, not making necessary frequent intervention for adjusting the system.

Acknowledgements

The data used in this study have been supplied by the Athens Air Quality Monitoring Centre (PERPA), the National Observatory of Athens (NOA) and the National Meteorology Service of Greece. Special thanks are due to Dr L. Viras, air pollution expert of PERPA, for his continuing advice and support during this ongoing research effort. Financial support has been provided by the Greek Ministry for the Environment and Public Works (EPPER4 Project - Specifications of the Athens Air Quality Monitoring Centre) and General Secretary of Research and Technology (Project PENED).

References

Aamodt, A; Plaza, E. (1994). Case-based reasoning: foundational issues, methodological variations, and system approaches AI communications, Vol.7 no1, March 1994.

Abdel-Aal R. E.; Elhadidy M. A. (1996). Modelling and forecasting the daily maximum temperature using abductive machine learning. Oceanographic Literature Review, 43 (1).

Avouris N.M., (1995) Co-operating Knowledge-Based Systems for Environmental Decision Support, Knowledge-Based Systems, vol. 8,n.1, pp.39-54, February 1995.

Avouris N.M., Kalapanidas E., (1997) Expert Systems and Artificial Intelligence Techniques in air pollution prediction, in N. Moussiopoulos (ed) *Mathematical Modelling of Atmospheric Pollution*, pp. 21-36, Thessaloniki.

Bartzis J. G. (1995), Environmental Monitoring and Simulation, Environmental Informatics - Methodology and Applications of Environmental Information Processing, N. M. Avouris & B. Page (eds.), pp.237-255, Kluwer Pubs., 1995

Boznar, M.; Mlakar, P. (1995). Neural networks - a new mathematical tool for air pollution modelling. In Air Pollution Theory and Simulation International Conference on Air Pollution - Proceedings v1 1995. Computational Mechanics Inc., Billerica, MA, USA. p. 259-266.

Breiling M, Alcamo J., (1992), Emergency Air Protection: A survey of Smog Alarm Systems, IIASA technical report of Leader project, Laxenburg, Austria 1992.

Kalapanidas E., Avouris N. (1999), Applying Machine Learning Techniques in Air Quality Prediction, Proc. ACAI 99, pp. 58-64, Chania, July 1999.

Kolodner, J.L., (1993), Case-based Reasoning, Morgan Kaufmann Publishers, pp.545-555 1993.

Lee, S. S. (1995). Predicting atmospheric ozone using neural networks as compared to some statistical methods. Proceedings of the 1995 IEEE Technical Applications Conference and Workshops, NORTHCON'95. Portland OR, USA

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

Mlakar, P.; Boznar, M. (1994). Short-term air pollution prediction on the basis of artificial neural networks. In International Conference on Air Pollution - Proceedings 1 1994. Computational Mechanics Publ, Southampton, Engl. P. 545-552.

Perantonis S.J., Vassilas N., Amanatidis G.T., Varoufakis S.J. and Bartzis J.G. (1994), Neural Network Techniques for SO_2 Episode Prediction,

Ruiz-Suarez, J. C.; Mayora-Ibara, O. A.; Smith-Perez, R.; Ruiz-Suarez, L. G. (1994). Neural network-based prediction model of ozone for Mexico City. In International Conference on Air Pollution - Proceedings 1 1994. Computational Mechanics Publ., Southampton, Engl. P. 343-400.

Ruiz-Suarez, J. C.; Mayora-Ibara, O. A.; Torres-Jimenez, J.; Ruiz-Suarez, L. G. (1995). Short-term ozone forecasting by artificial neural networks. In Advances in Engineering Software v23 n3 1995, p.143-149.

Simon K.H., Jaeschke A., Manche A., (1995), Environmental Applications of Expert Systems Technology, in N.M. Avouris, B. Page, Environmental Informatics, Kluwer Academic, pp. 93-109, 1995.

Sycara E.P., (1988), Using case-based reasoning for plan adaptation and repair, in Proceedings: Workshop on case-based reasoning (DARPA), Clearwater, Florida. San Mateo, CA. Morgan Kaufmann.

Watson I.D., Case-Based Reasoning Development Tools: A Review, http://www.salford.ac.uk/docs/depts/survey/staff/IWatson/cbrtools.htm, World Wide Web (1996).

Yi J.; Prybutok V. R. (1996). A neural network model for the prediction of daily maximum ozone concentration in an industrialised urban area. Environmental Pollution 92(3) p. 349-357.

level 1 (low)	0 - 200 µg/m^3
level 2 (medium)	200 - 350 µg/m^3
level 3 (high)	350 - 500 µg/m^3
level 4 (alarm)	over 500 µg/m^3

Table 1 : Pollution levels based on NO_2 concentrations for the Athens area

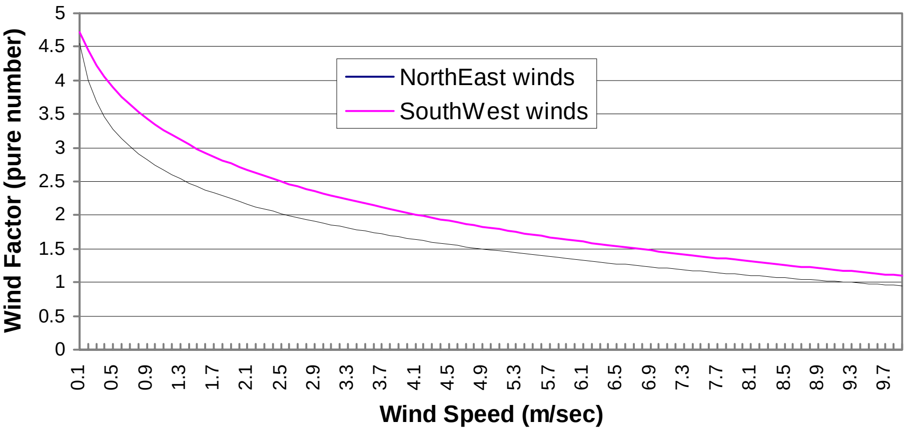

Figure1:Heuristic definition of wind factor

Wind Factor		Wind Description by NMS
SE-S-SW-W	NW-N-NE-E	
4	4	NIL
2.5	2.5	Light
2	1.5	Moderate
1	1	Strong or very strong

Table 2: Meteorological Wind Forecast and corresponding Wind Factor

Temperature Inversion Factor	**DT(temperature difference), H(height from ground)**
0	No inversion
1	DT = 0, H at 50 m
2	DT > 0, H over 50 m
3	DT > 0, H over 50 m, multiple inversion layers up to H 1500 m

Table 3: Heuristic evaluation of Temperature Inversion Factor

	Precipitation forecast						
Rain factor	Nil	possible	light	snow	rain	Periodical rain	storm
1					×	×	×
0	×	×	×	×			

Table 4: Heuristic Evaluation of the Rain Factor

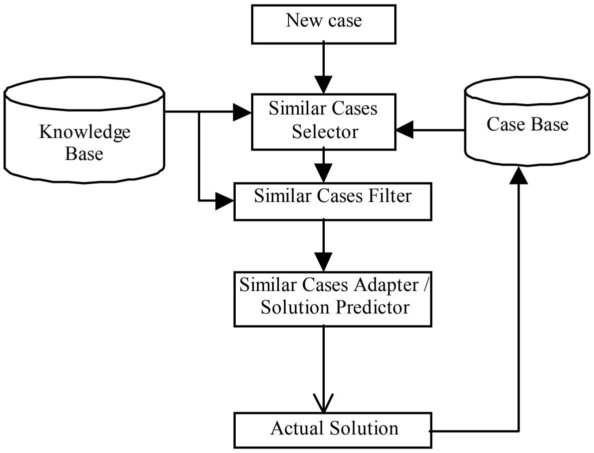

Figure 2: CBR System (NEMO) block diagram

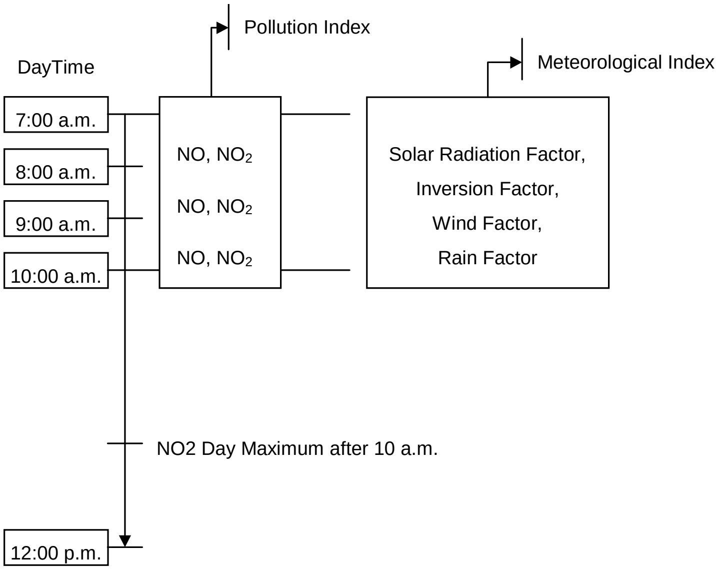

Figure 3: NEMO: Temporal relation of the prediction factors involved

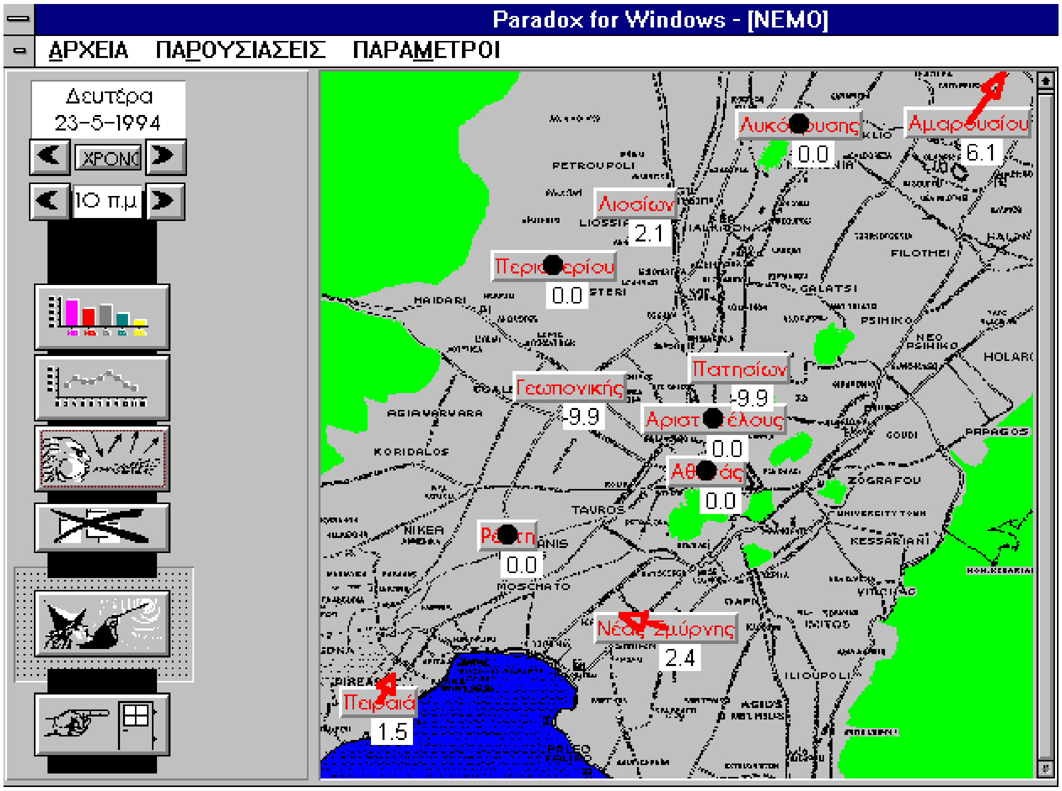

Figure 4: NEMO: User Interface: Presentation of a parameter on the area map

Variable Name	Value	Description
filter A	pollution	Where the first similarity filter is based on
filter B	meteorology	Where the second similarity filter is based on
prevCaseMarginRelaxConst	*2	Relaxation factor for the attribute margins for the previous day
HighWindMarginRelaxConst	*1.5	Relaxation factor for the attribute margins when the new case is a day with high winds
LowNoMarginRelaxConst	*1.5	Relaxation factor for the attribute margins when the new case is a day with low morning NO
HighWindThresholdConst	0.8	Wind threshold, over which the new case is considered as a day with high winds (and so the pollution attribute selection margins should be relaxed)
LowNoThresholdConst	0.1	NO threshold, under which the new case is considered as a day with low morning NO concentrations (and so the meteorological attribute selection margins should be relaxed)
meteoThresholdConst	0.8	Meteorological threshold, over which the old case to be considered as similar meteorologically with the new case, and therefore pass one of the two similarity criteria
pollutionThresholdConst	0.7	Pollution threshold, over which the old case to be considered as similar over a pollutants concentrations average with the new case, and therefore pass one of the two similarity criteria
SufficientSimilarCasesNo	25	Minimum number of the old similar cases, in order for the final prediction formula to extract an acceptable predictive result
totalCasesConst)	50	Maximum number of pastCases retrieved for similarity assessment (related to the casesToSearch variable)
casesToSkip	10	Number of cases from the start of the caseBase to be skipped before starting predicting (made necessary in order for the caseBase to have sufficient pastCases to predict when NEMO starts predicting)
Final predictive algorithm	NO2Morning	This is a statistical formula that extracts the maximum NO2 value of the new case from the corespondent NO2 value of the similar past cases remained, weighted by the similarity distance of each pair (new case, past case)

Table 5 Parameters settings for NEMO

	Cases	%
Level 1	139	58
Level 2	64	27
Level 3	29	12
Level 4	8	3

Table 6. Testing set classification

	Predicted as level					Prediction
	Unable	1	2	3	4	Accuracy
Level 1		134	5			96.4
Level 2		38	22	2	3	33.8
Level 3			12	9	8	33.34
Level 4	4				4	50

Table 7 NEMO System performance for the 240 cases of the Testing Set.

	System		
Criterion	***NEMO***	***DT***	***ANN***
Overall Success	169 (70%)	183 (76,2%)	166 (69,1%)
Relaxed Overall Success	233 (97,1%)	236 (98,3%)	232 (96,7%)
Level 4 success	4 (50%)*	8 (100%)	4 (50%)

**4 missed cases that NEMO was unable to predict.*

Table 8 Comparative System performance of NEMO, Decision Tree and Artificial Neural network.

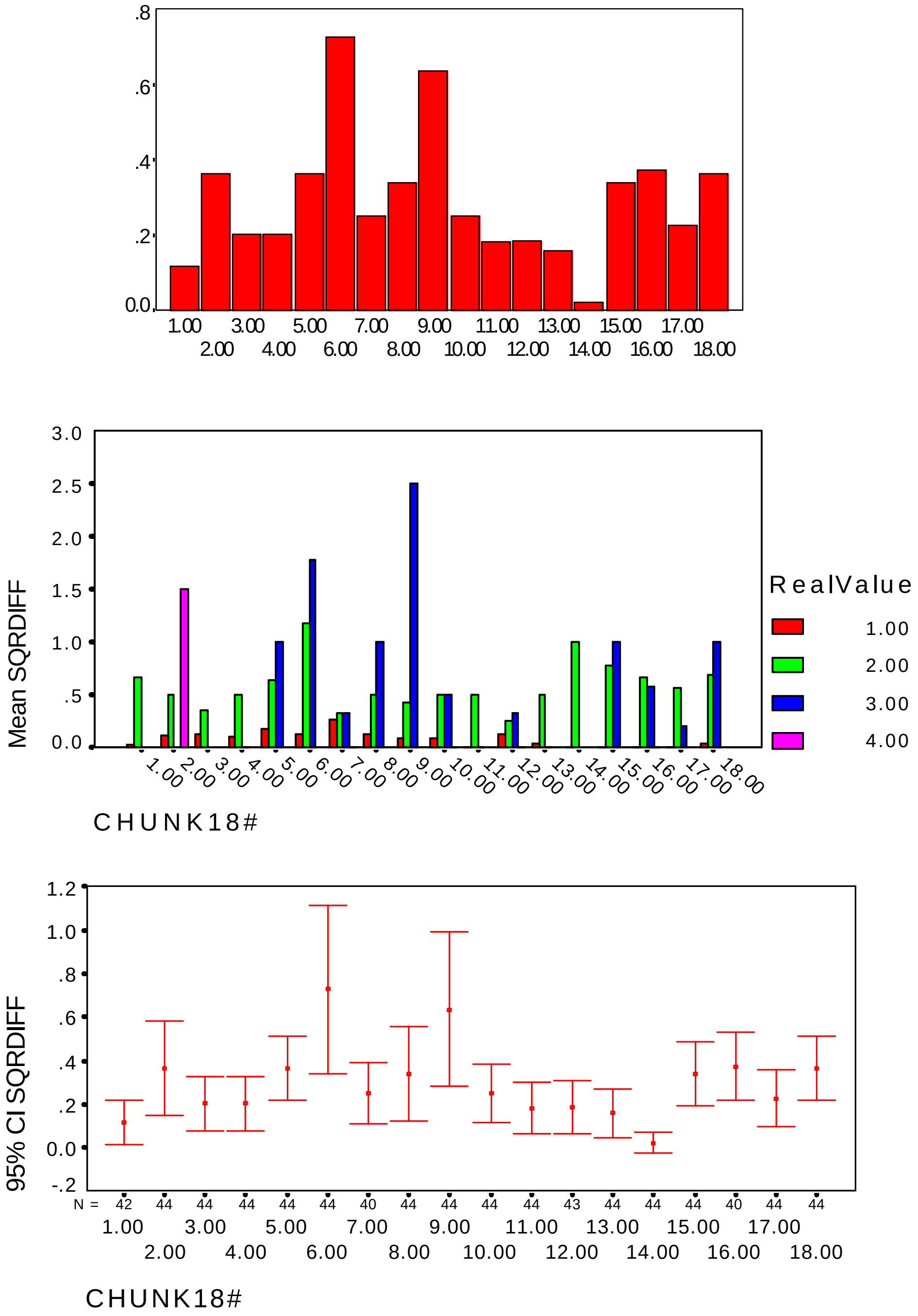

Figure 4 Time evolution of NEMO performance