# COMPARISON OF ARTIFICIAL INTELLIGENCE PREDICTION TECHNIQUES IN NO AND $NO_2$ CONCENTRATIONS' FORECAST

## I. JUHOS[1], R. BÉCZI[2] and L. MAKRA[2]

[1]*Department of Informatics, University of Szeged, P.O.Box 652, 6701 Szeged, Hungary*
*E-mail: juhos@inf.u-szeged.hu*
[2]*Department of Climatology and Landscape Ecology, University of Szeged,P.O.Box 653, 6701 Szeged, Hungary*

**Összefoglalás** – A műszaki eszközök és berendezések, valamint az épületek védelme iránti állandó igény, továbbá az az elvárás, hogy a gazdasági és gazdálkodási folyamatok költségeit csökkentsük, egyre pontosabb előrejelzési technikákat igényelnek. Az előrejelzés nehéz probléma, mellyel csaknem minden emberi tevékenység szembesül. Kifejlesztettek ugyan számos speciális idősor előrejelzési módszert, ezek mindegyikének azonban vannak bizonyos korlátai. Legtöbbjük sokkal inkább teljes adatsorok modellezésére korlátozódik, semmint az előrejelzési sajátosságok kiemelésére és a terület szakemberei számára általában nehezen értelmezhetők. A mesterséges intelligencia a döntési fákkal történő szimbolikus tanulást ajánlja, mely lehetőséget biztosít a múltbeli adatokban rejlő kapcsolatok felderítésére, számunkra is olvasható formában. Mindezek mellett képes megbecsülni azt, hogy a jövőbeli adatok milyen intervallumba esnek. Precízebb előrejelzést az elmúlt években erre a problémára legtöbbet használt mesterséges neurális hálózatok segítségével nyerhetünk, amelyek pontos függvényillesztést végeznek az adatokon. Ennek ára azonban az összefüggések emberi szem elől való elrejtése. Ha az említett módszerek kombinációját használjuk, akkor pontosabb döntésekre juthatunk a jövőbeni adatokra vonatkozóan, továbbá az okokat is feltárhatjuk. Mindkét esetben a tanulás hatékonysága a tanulóalgoritmusok paramétereinek jó megválasztásától függ. Emiatt a paraméterek beállítására szimulált hűtéssel felügyelt tanulást alkalmaztunk. A dolgozat célja, hogy összehasonlítsuk a fent említett technikákat az NO és $NO_2$ koncentrációk néhány órás előrejelzésében egy forgalmas szegedi közlekedési csomópontban. Ehhez az aktuális értékeik alapján adott hibával előrejelzett meteorológiai paramétereket hívtunk segítségül.

**Summary** – To construct new technical devices, to permanently protect buildings and to reduce the expenses of various economic and business processes more and more accurate prediction techniques are needed. Almost all human activities encounter the hard problem of forecasting. Although several time series prediction methods have been developed, each of them has certain limitations. Most of them are designed rather for modeling complete time series than pointing out different prediction characteristics; furthermore, they can only be interpreted with difficulties. Artificial intelligence offers symbolic learning with decision trees, by means of which we can explore connections in past data and produce them in a readable format. Decision trees can estimate intervals of future data. Recently, artificial neural networks were used to handle this problem. This method offered more precise forecast and more accurate fit of the function to the starting data. However, when applying this method, relationships in the data set examined were hidden. If we combine the methods mentioned above, we can get more precise decisions for the future data and we can also reveal the reasons. In either case, the efficiency of learning depends on a good choice of the learning algorithms' parameters. For this reason, parameters are selected by simulated annealing. The aim of this paper is to compare the above mentioned prediction techniques in several hours forecast of NO and $NO_2$ concentrations at a busy cross-road in Szeged (Hungary). For this object, meteorological parameters predicted with given error on their actual values were used.

*Key words:* NO and $NO_2$ concentrations, neural network, decision tree, forecast, prediction

## INTRODUCTION

Considering the special meteorological and geographical conditions of Szeged, dispersion of air pollutants – especially during permanent anticyclone weather conditions in
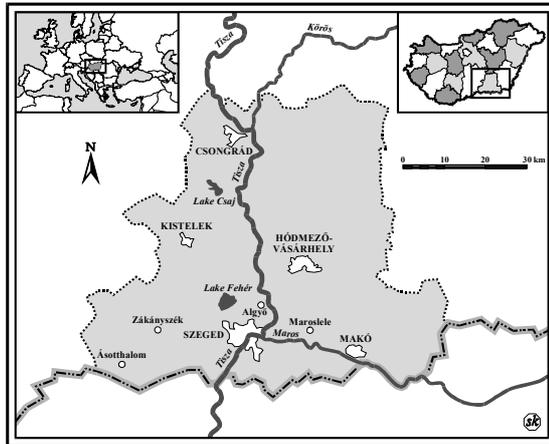
*Fig. 1a* Geographical position of Szeged, Hungary in Europe (top left); Csongrád county in Hungary (top right); and Szeged in Csongrád county (centre)
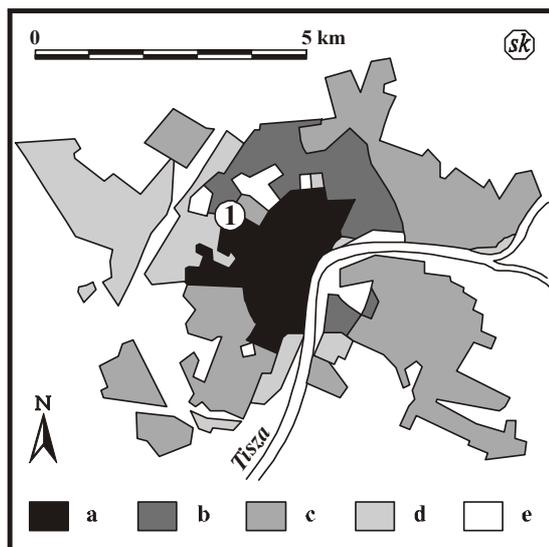


*Fig. 1b* The built-up types of Szeged [a: centre (2-4-storey buildings); b: housing estates with prefabricated concrete slabs (5-10-storey buildings); c: detached houses (1-2-storey buildings); d: industrial areas; e: green areas, (1): monitoring station]

the summer and winter seasons – is very slow. A reliable forecast of concentrations of the air pollutants is very important. It makes possible for the authorities to prepare people for heavy air pollution episodes. Several authors have used methods using classical statistics as well as methods of neural networks in order to make short term forecast of various gases and particulate matter. *Gardner and Dorling* (1998) give an excellent account on the applications of neural network methods for forecasting in atmospheric sciences. It is worth to mention here paper of *Jorquera et al.* (1998), which compares a linear model – a neural network model – and a fuzzy model as a prediction of daily maximum ozone concentrations. *Perez et al.* (2000) presents an application of neural networks for a few hours prediction of $PM_{2.5}$ in the atmosphere of Santiago city (Chile), where the input data are hourly $PM_{2.5}$ concentrations in the previous day. *Ziomas et al.* (1995) analyses the possibility of forecasting maximum ozone concentrations in Athens city. He applied discriminant analysis to forecast possible increase and decline of $NO_2$ levels. He considered the following parameters: previous daily maximum ozone concentration; forecasted temperature; wind velocity and direction; an index of the given day's short term emission change; an index of the effect of the precipitation on the given day. In average, 80 % of the forecasts were successful. If we use multivariate linear regression analysis, these forecasts give quantitative estimates for the maximal hourly averages of $NO_2$ concentrations on the next day.

In this paper, a test is presented and different forecasting methods are compared; furthermore, they are used to forecast hourly averages of NO and $NO_2$ concentrations.

DATA BASIS

Data basis of the paper comes from the monitoring station located in Szeged city in a crossroad (Kossuth Lajos Avenue and Damjanich Street). The data basis consists of a five-year data set from the period between January 1, 1997 – December 31, 2001. The elements considered are, on the one hand, the average mass concentrations ($\mu g\ m^{-3}$) of $SO_2$, $O_3$ and $PM_{10}$ and, on the other hand, the average values of the main climatic elements (temperature, humidity, air pressure, global radiation, wind direction and wind speed). The verified database of the study consists of the 30-minute averages of the above-mentioned pollutants and climatic elements. The monitoring station is operated by the Environmental Protection Inspectorate of Lower-Tisza Region, Szeged, under auspices of the Ministry of Environment, Hungary.

Szeged (20°06'E; 46°15'N) is situated near the confluence of the Tisza and Maros Rivers. It is the largest town in the south-eastern part of Hungary, at about 20 km from Hungarian-Serbian border. The city and its surroundings are flat and low and the elevation is 79 m above sea level, the lowest in the Carpathian basin. The number of inhabitants of the city is up to 160,000 and the territory of its built-up area is about 46 km$^2$ (*Fig. 1a-b*).

The basis of the city structure is a boulvard-avenue street system crossed by the Tisza River (*Fig. 1b*). In this way, the structure of the city is simple; however, following to this system, motor vehicle traffic as well as air pollution are concentrated in the city.

The industrial area is found main-ly in the north-west part of Szeged. Thus, the prevailing westerly and northerly winds transport air pollutants, coming from this area, towards the city centre.

METHODS

*Inductive learning of atmospheric parameters*

Inductive learning of a concept means recognizing a hypothesis regarding this concept after presenting the training instances to the learner. The simplest learning case is that, where one part of the training instances is true (positive) and another part is false (negative). A set of the instances can be regarded as a function. The domain of this function consists of the instances, while the values are either true or false according to the instance. During the training process, the instances are generally at disposal in the following format:

$$x_1,\ x_2,\ \ldots\ ,\ x_n, \qquad y$$

*instance*       *class*

where $x_i$ is the *i*-th attribute of the instance and *y* is the class of the instance (true or false). A training instance and its class is a training example (in certain case an instance phrase is equal to a mentioned example). In order to find the inductive hypothesis, a function of $y = h(x_1,\ x_2,\ \ldots\ ,\ x_n)$ based on the training instances have to be approximated. The number of the classes can be extended to more than two; thus, the problem can be generalized to the classification into more than two discrete classes or to the learning of functions with not discrete range. Goodness of the *h* hypothesis can be determined by applying it on the not presented instances (which were not in the training set).

The estimate of the atmospheric parameter corresponds to an inductive learning model. Past data are the instances, and the forecast of the data will be determined by an inductive hypothesis. The accuracy of the learning depends on the number and on the accuracy of the training data (data can be come from real process by measurement), while the quality of the learning (the finding of the inductive hypothesis) depends on the chosen learning algorithm.

The accuracy of the meteorological data is determined by measuring instruments. We chose a database of approximately five months for the learning period. The data taken into account are hourly average values of concentrations between 28th February 2003, 00:00 CET to 9th July 2003, 24:00 CET. In the paper artificial neural networks are applied as learning algorithm or machine learner, which are frequently used in this field. To understand the relationships among the data, decision trees were applied.

*Decision trees*

The decision tree is a tree-graph, vertices of which denote branch points. The branch points are determined by the attributes of the learning instances. The leaves of the tree mean a classification of the learning instances. The applied algorithm of the C4.5 decision tree (*Quinlan,* 1993) builds up the tree based on the so-called information gain. The gain determines the extent to which an attribute describes the proper classification of a given set of instances; i.e., which decreases mostly the entropy of the classification. If we choose the attribute with the largest gain, the algorithm puts it to the root of the tree. The sub-trees branching from the root are determined following the same procedure from the partition classes given by the chosen attribute. The learning, which is exactly the construction of the tree, stops if we reach the last attribute. This algorithm is working on discrete attribute values but easily extends to real attributes. After the learning, the decision tree is able to classify previously unknown instances. Certain rules between the attributes can be derived from the tree and we can also infer their effect on the decision making. If the tree grows too large, we may loose generality. To avoid this, we can do pruning of the branches of the decision tree. Before building the tree, categories have to be made if the learning class is not discrete. In our case, this particularly means the partitioning of the estimated range into intervals.

*Artificial neural network*

In this paper estimations are provided for future concentrations by backpropagation multi-layer perceptron (BMLP) 24 hours beforehand (*Kaastra and Boyd,* 1996). In the hidden layer, sigmoid activation function is used:

$$g(x) = \frac{1}{1 + e^{-x}} \qquad (1)$$

The main task is to approximate the functions. The three-layer BMLP is capable of approximating arbitrary finite sets of real numbers (*Hornik et al.,* 1989). When the attributes of the learning instances take the form $x_{i1}, \dots, x_{in}$ then the class of the i-th instance produced by the perceptron is

$$y_i = g(w_1 x_{i1} + \dots + w_n x_{in} + b) \qquad (2)$$

48

where $w_j$ ($j = 1, ... , n$) are the weights of the input layer; $b$ is the threshold of the activation function; $g$ is the activation function.

### *Overfitting*

A general drawback in machine learning is overfitting. When the precision of the approximation of the desired function is increased, the generality may be destroyed. Thus, in case of the instances taken outside of the training set the sum of the errors increases. Generally, these phenomena may be observed in the later phases of the training process. Using a larger training set or stopping the training process in due time may provide a solution to this problem; nonetheless, there are no exact definitions of the correct stopping time. In the case of decision trees, we saw a branch-pruning technique, which could lead to a solution, but which also introduces more and more parameters. It is not determined when or which branch pruning is needed. We see that learning process is significantly influenced by certain parameters, which are usually set through several trials.

### *Setting of the parameters*

Learning, as function of fitting, involves optimization. The function to be optimized has, in general, bad behaviours. It has many parameters and several local extrema, thus the searching algorithms are often stuck in local minimum points. The multi-parameters optimization and the approximation is carried out during the training process (in case of the BMLP it is done using backpropagation), its goodness is the function of certain parameters (in case of BMLP, learning rate used in backpropagation, momentum, etc.) regarding the learners. These are usually determined by experiments. If more precision is required, another optimization is needed but this time not among the training instances rather in the parameter space of our training algorithm; which, of course, depends on the training data.

### *The heuristic search*

We use a heuristic search algorithm to handle and control the above-mentioned problems of overfitting and the setting of training parameters and thus to facilitate a better learning. During this algorithm, we split the historical data set in two parts: a larger set of training and a smaller set of test instances; more exactly, the union of these gives back the original training set. Considering an initial parameterization of our learner, which is installed on a training set, a hypothesis is obtained. We check its correctness on the set of test instances and we measure the error (it is called validation). The error determines a so-called fitness function $f$ depending on the training parameters, the smaller the error is the smaller the values of $f$ are. We try to minimize the fitness function using the heuristic search. Reaching the minimum of $f$ it is plausible that our posed hypothesis is appropriate according to the given data. Using the parameters yielded by the heuristic search, we train our learner again on the union of the sets of training and test instances, i.e., on the historical training set. Thus, using more instances in the training process hypothesis becomes better.

Simulated annealing was applied as heuristic search, which is frequently used in the field of artificial intelligence. This is a quick algorithm, which provides precise estimates on the location of optimum (*Nahar et al.,* 1986).

*The simulated annealing*

The simulated annealing (SA) is a heuristic method of locating the extrema of functions. In our case, we are looking for the minimum of the fitness function *f* (or the maximum of *1/f*, if it exists). The algorithm has a physical motivation. During the annealing of liquids, the fine structure of the material follows the principle of optimal disposition. There are numerous implementation of SA. We apply a version, which is controlled by the following five parameters: initial temperature, current temperature, minimal temperature (freezing), annealing velocity and maximum number of steps.

Initially we assume that the distance between the elements of the search space (the space of the learners' parameters) is the same; i.e., our assumption is that we have only neighbouring elements. Later, as the temperature falls, the distance between the elements increases according to a coefficient *μ* depending on the temperature.

We use the Manhattan-distance *d* to measure the distance between two elements $P = (p_1, p_2, \dots , p_n)$ and $P' = (p_1', p_2', \dots , p_n')$:

$$d(P, P') = \sum_{i=1}^{n} \left| p_i - p_i' \right| \tag{3}$$

We say that the parameter P' is in the *μ*- neighbourhood of P, if the following condition holds:

$$d(P, P') < \mu \sum_{i=1}^{n} \left| \max(p_i) - \min(p_i') \right| \tag{4}$$

where $\mu = temp / temp_{max}$; *temp* = actual temperature; $temp_{max}$ = maximum temperature; $max(p_i)$ and $min(p_i)$ denote the maximum and the minimum of the possible values of the i-th parameter $p_i$ respectively. The function that we would like to optimize is the fitness function *f(P)*. Calculating the value of *f* for the initial parameter *P* we decide whether *f(P)* is an optimum. If not, we continue the procedure with choosing an element *P'* from the *μ*-neighbourhood of *P*. In any case, we accept the step from *P* to *P'*, if *f(P')* is better value than *f(P)*. According to SA, we also accept a worse parameter with probability *v*, which makes it possible that the algorithm leaves local optima. If the better values are only accepted, the search would certainly get stuck in local optima. The probability *v* depends also on the temperature:

$$v(temp) = e^{\frac{f(P') - f(P)}{temp}} \tag{5}$$

At each step, the temperature falls according to annealing velocity. Reaching the optimum or the maximum number of steps, the algorithm stops and returns the previously found best parameter.

Since the algorithm requires no assumption on the shape of *f*, it provides a widely applicable tool. Taking small enough decrease in the temperature at each step, we may get sufficiently close to or reach the optima.

The SA is used to set the following parameters of the learners. In case of decision trees: it is necessary to apply pruning, pruning threshold and maximal number of nodes in the tree. In case of BMLP: training time, learning rate and moments.

## RESULTS

The historical data used in the training are the hourly means of NO, $NO_2$, $O_3$ concentrations and the intensity of global radiation in one-hour steps between 28th February 2001 00:00 CET and 9th July 2001 24:00 CET. The forecasted interval is the period between 10th July 2001, 00:00 CET and 13th July 2001, 24:00 CET, also in one-hour steps.

For the precise estimates we used the 24-1-1 type BMLP (24 input, 1 hidden and 1 output neurons). To explore the relationships between the historical data, the C4.5 decision tree was applied.

Firstly, the attributes of one training instance were the values measured in the 24 hours of a given day (attributes of the instance) and one concentration value of NO or $NO_2$ taken from the following day (class of the instance, i.e., the material which is forecasted) (see Example 1). Then, we added some future external factors measured in the same hour as the data to be predicted. Such factors are the concentration of $O_3$, intensity of the global radiation, and $NO_2$ or NO concentrations depending on which material (NO or $NO_2$) is to be forecasted (see Example 2). Thus, we have two types of estimations, i.e., learning and estimations with or without external factors. Naturally, the future external factors, similarly to the data that is the subject of the estimation, are not known in the moment of the forecast, since both relate to the same future time. Therefore, these future external factors have to be estimated (see *Fig. 2-6*). With this technique our original estimate may become more accurate. Since we predict the concentration values from Tuesday to Friday for each hour of a day, the attributes of a training instance, i.e., the data that are used in the estimation, are the 24 values of each factor measured on the previous day (e.g., we estimate the values on Tuesday from the values on Monday).

For example, one instance is like the following without external factor (the emphasized (bold) part of the examples denotes the class of the instances):

Example 1

$$NO_{0h}(today) \, , \, ... \, , \, NO_{24h}(today), \quad \mathbf{NO_{12h}(tomorrow)}$$
$$\text{\textit{instance}} \qquad\qquad\qquad \text{\textit{class}}$$

with external factors:

Example 2

$$NO_{0h}(today) \, , \, ... \, , \, NO_{24h}(today), \, NO_{2\ 12h}(tomorrow)' \, , \, O_{3\ 12h}(tomorrow)' \, , \, glob.$$
$$rad._{12h}(tomorrow)', \, \mathbf{NO_{12h}(tomorrow)}$$

Since we would like to predict the concentrations between Tuesday and Friday for each 24 hours of a day, we have 24x4 independent training periods. In each training process the parameters of the learners were set by the SA, and we split the data set in 80%-20% parts for the validation. In the validation the fitness function was the root mean squared error on the test set (Eq. 6).

$$e_{rmse} = \sqrt{\frac{\sum_{i=1}^{m}(\bar{y} - y)^2}{m}}$$ (6)

$\bar{y}$ = the estimated value; $y$ = actual value; $m$ = number of instances in the test set.

In each period of the training carried out by the BMLP, we obtained estimations on the means of the observed parameter in a given hour by averaging the results of five independent executions of the SA. Because of the heuristic search, the SA does not always give the same results. That is why the average of five independent estimations have been used. In the sequel, we present the results on the estimations of the parameters in Section "*Hourly prediction of the factors ($O_3$, NO, $NO_2$ and global radiation)*", while the forecast of NO and $NO_2$ will be discussed in Section "*Hourly forecast of NO and $NO_2$ concentrations with external factors*".

### *Hourly prediction of the factors ($O_3$, NO, $NO_2$ and global radiation)*

In *Fig. 2-5*, the solid line denotes the actual hourly average concentrations; the dashed line depicts the results obtained by the BMLP using the average of five SAs, while the dotted line draws the picture of the results yielded by the BMLP without parameter setting.
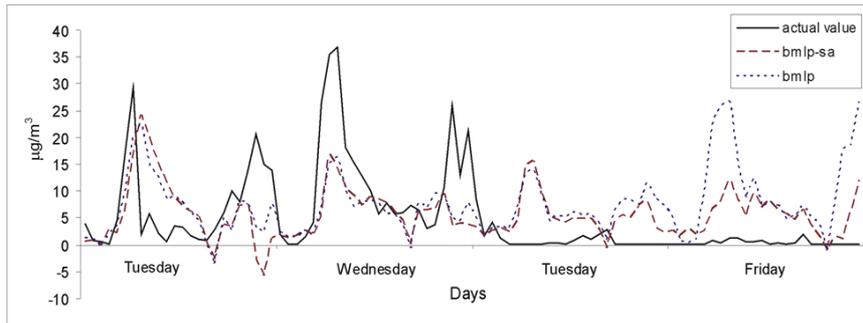


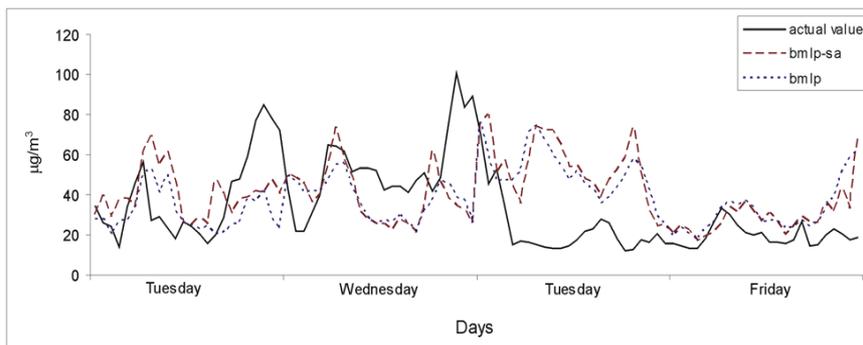*Fig. 2* Estimation of NO concentration without external factors



*Fig. 3* Estimation of $NO_2$ concentration without external factors

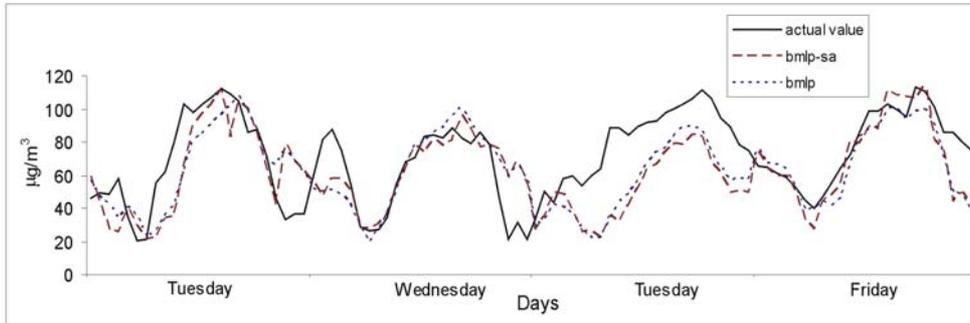As we can see in *Fig. 2,* the BMLP combined with SA is more capable to fit to the flat piece occurring on Friday.

52

*Fig. 4* Estimation of $O_3$ concentration without external factors

As we can see in *Fig. 4-5*, the $O_3$ concentration and the global radiation are easier to learn; thus, their utilization as external factors is beneficial. This is not the case for NO and $NO_2$. We can see that the BMLP with SA gave worse estimations in a couple of points. One reason of this could be that the training set does not give an appropriate representation of the problem; furthermore, it may happen that not all possible training and test instances occur during the validation. It has also to be mentioned that after performing 200 iterations the execution of the SA was always stopped; hence, the precision of the algorithm is limited in this way.
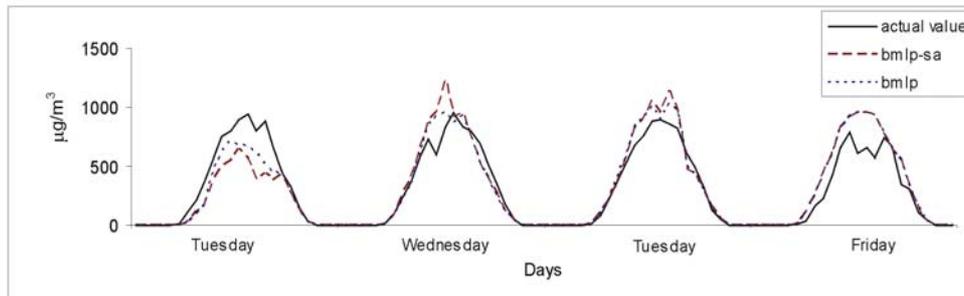


*Fig. 5* Estimation of the global radiation without external factors

*Hourly forecast of NO and $NO_2$ concentrations with external factors*

In *Fig. 6-7*, the solid line denotes the actual hourly average concentrations; the dashed line depicts the results obtained by the BMLP using the average of five SAs with factors, while the dotted line draws the results without factors.

We see in *Fig. 2-5* that setting the parameters with SA resulted in better results in most of the cases. It can also be seen in *Fig. 6-7* that using factors in the estimations of NO and $NO_2$ yield much better approximation compared to that of Section "*Hourly prediction of the factors ($O_3$, NO, $NO_2$ and global radiation)*". We remark that the examples presented here are based on actual values on future factors.

The largest deviation from the actual value of NO concentration occurs on Tuesday between 20:00 and 23:00 (*Fig. 6*). To explore the reasons of this, let us consider the decision trees belonging to Wednesday 20:00 and 23:00 (*Fig. 8-9*). In this case the route of the decisions is denoted along the edges of the graphs. This shows that the most significant factors are the

$NO_2$ and $O_3$ concentrations, while the global radiation does not play an important role; thus, it is not really involved in the calculations. This can also be seen in *Fig. 5*, since global radiation shows periodic values; hence, they do not provide extra information neither in the training nor in the estimations.
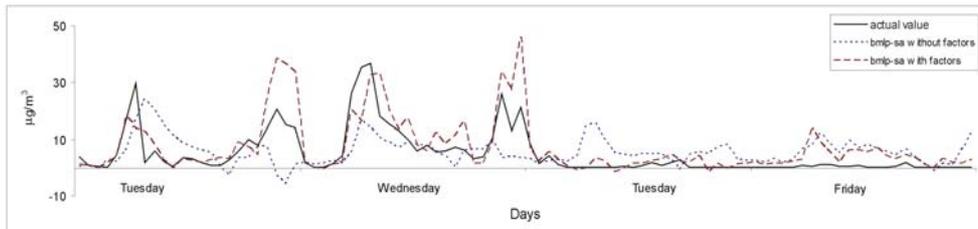


*Fig. 6* Estimation of NO concentration using the external factors; $NO_2$, $O_3$ and global radiation
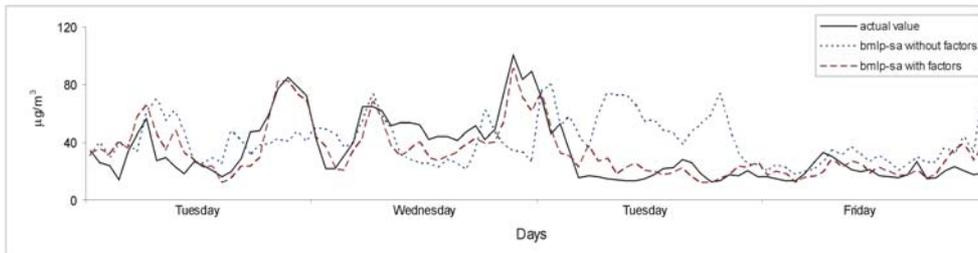


*Fig. 7* Estimation of $NO_2$ concentration using the external factors NO, $O_3$ and global radiation
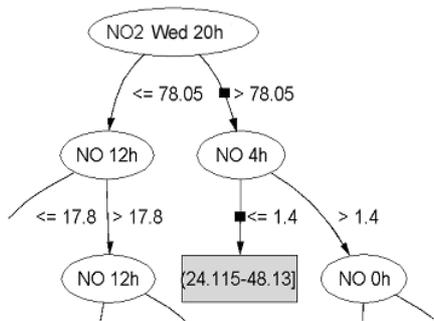


*Fig. 8* Part of the decision tree belonging to the estimation of the NO value on Wednesday 20h using the external factors $NO_2$, global radiation and $O_3$
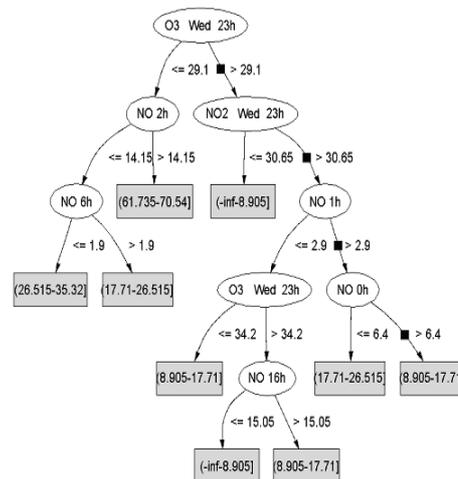
*Fig. 9* Part of the decision tree belonging to the estimation of the NO value on Wednesday 23h using the external factors $NO_2$, global radiation and $O_3$.

54

## DISCUSSION

*Table 1* shows that the BMLP gives precise estimates, while the decision tree is only able to determine intervals of the estimated variable but, at the same time, it is applicable to determine the relationships among the data.

Comparing the two methods, we may see that both yielded more or less the same predicted values. According to this result and observing the historical data used in the training process, we may find out that reasons of inaccurate estimates lie in the possibly inappropriate modeling of the problem and in the fact that some additional factors which are not taken into account at this time may also affect the estimates. Nevertheless, the presented system follows the trend of the NO concentration quite well.

*Table 1* Estimated values of NO concentrations ($\mu g/m^3$) given by BMPL and decision tree by SA parameters adjusting

| Wednesday | 20 h | 21 h | 22 h | 23 h |
|---|---|---|---|---|
| Actual value | 10.70 | 26.10 | 13.05 | 21.35 |
| 5SA-BMLP-$NO_x$ | 8.74 | 33.90 | 26.80 | 46.10 |
| 5SA-dtree-$NO_x$ | (8.45-16.81) | (24.10-48.13) | 24.50-49.10) | (8.90-17.71) |

## REFERENCES

*Gardner, M.V. and Dorling, S.R.,* 1998: Artificial neural networks (the multilayer perceptron) – a review of applications in atmospheric sciences. *Atmos. Environ. 32,* 2627-2636.

*Hornik, K., Stinchcombe, M. and White, H.,* 1989: Multilayer feedforward networks are universal approximators. *Neural Networks 2,* 359-366.

*Jorquera, H., Perez, R., Cipriano, A., Espejo, A., Letelier, M.V. and Acuňa, G.,* 1998: Forecastinh ozone daily maximum az Santiago, Chile. *Atmos. Environ. 32,* 3415-3424.

*Kaastra, I. and Boyd, M.,* 1996: Designing a neural network for forecast financial and economical time series. *Neurocomputing 10,* 215-236.

*Nahar, S., Sahni, S. and Shragowitz, E.,* 1986: Simulated annealing and combinatorical optimization. *The 23rd Design Automation Conference,* 293 - 299.

*Perez, P., Trier, A. and Reyes, J.,* 2000: Prediction of PM 2.5 concentration several hours in advance using neural networks in Santiago, Chile. *Atmos Environ. 34,* 1189-1196.

*Quinlan, J. R.,* 1993: C4.5: *Programs for Machine Learning.* Morgan Kaufmann.

*Ziomas, I.C., Melas, D., Zerefos, C.S., Bais, A.F. and Paliatsos, A.G.,* 1995: Forecasting peak pollutant levels from meteorological variables. *Atmos. Environ. 24,* 3703-3711.