Available online at www.sciencedirect.com
ScienceDirect
Environment International 34 (2008) 592–598
ENVIRONMENT INTERNATIONAL
www.elsevier.com/locate/envint

An online air pollution forecasting system using neural networks

Atakan Kurt [a], Betul Gulbagci [a], Ferhat Karaca [b,*], Omar Alagha [b]

[a] Department of Computer Engineering, Fatih University, Buyukcekmece, 34500, Istanbul, Turkey
[b] Department of Environmental Engineering, Fatih University, Buyukcekmece, 34500, Istanbul, Turkey

Available online 30 January 2008

Abstract

In this work, an online air pollution forecasting system for Greater Istanbul Area is developed. The system predicts three air pollution indicator (SO_2, PM_{10} and CO) levels for the next three days (+1, +2, and +3 days) using neural networks. AirPolTool, a user-friendly website (http://airpol.fatih.edu.tr), publishes +1, +2, and +3 days predictions of air pollutants updated twice a day. Experiments presented in this paper show that quite accurate predictions of air pollutant indicator levels are possible with a simple neural network. It is shown that further optimizations of the model can be achieved using different input parameters and different experimental setups. Firstly, +1, +2, and +3 days' pollution levels are predicted independently using same training data, then +2 and +3 days are predicted cumulatively using previously days predicted values. Better prediction results are obtained in the cumulative method. Secondly, the size of training data base used in the model is optimized. The best modeling performance with minimum error rate is achieved using 3–15 past days in the training data set. Finally, the effect of the day of week as an input parameter is investigated. Better forecasts with higher accuracy are observed using the day of week as an input parameter.

Keywords: Air pollution; Prediction; Meteorology; Neural networks; Turkey

1. Introduction

Air pollution is one of the important environmental problems in metropolitan cities (Akkoyunlu and Ertürk, 2003; Karaca et al., 2004, 2005b). There are many air pollution indicators affecting human health (Elbir et al., 2000; Tayanç, 2000). Some of the important indicators are particulate matter (PM_{10}), carbon monoxide (CO) and sulfur dioxide (SO_2). EU and many national environmental agencies have set standards and air quality guidelines for allowable levels of these pollutants in the air. When the concentration levels of these indicators exceed the air quality guidelines, short term and chronic human health problems may occur (Künzli et al., 2000).

Urban air quality management and information systems are required to predict the next day's air pollution levels and for providing proper actions and controlling strategies (Monteiro et al., 2005). Air quality warning systems are therefore needed in order to obtain accurate advance notice that the ambient air concentration levels might exceed the air quality guidelines or the limit values. Warnings can be utilized to alert health care as well as traffic and environmental management so that the adverse effects can be minimized. Such warning systems must be sufficiently reliable and understandable by the majority of people.

Currently the municipality of Istanbul has been monitoring air pollutants at ten permanent stations covering Istanbul from Yenibosna to Kartal. A web-based information system (www.ibb.gov.tr/index.htm) publishes information about the current and earlier periods of air pollution conditions in Istanbul. However, currently there is not an air quality forecast system for the city. This is a shortcoming for not only Turkey but many other countries. A few limited examples of air quality prediction systems include; The UK national air quality information archive: http://www.airquality.co.uk/, A cross-agency U.S. Government Web site: http://airnow.gov/, Australian Air Quality Forecasting System: http://www.epa.vic.gov.au/Air/AAQFS/default.asp.

In recent years, neural network models have been developed and successfully applied to atmospheric pollution modeling in general (Gardner and Dorling, 1998) and air quality problems in particular (Boznar et al., 1993; Gardner and Dorling, 1999a,b; Chaloulakou et al., 2003; Karaca et al., 2005a, 2006a). Unlike other modeling techniques, artificial neural networks (ANN)

* Corresponding author. Tel.: +90 505 616 4632; fax: +90 212 866 3412.
E-mail address: fkaraca@fatih.edu.tr (F. Karaca).

0160-4120/$ - see front matter
doi:10.1016/j.envint.2007.12.020

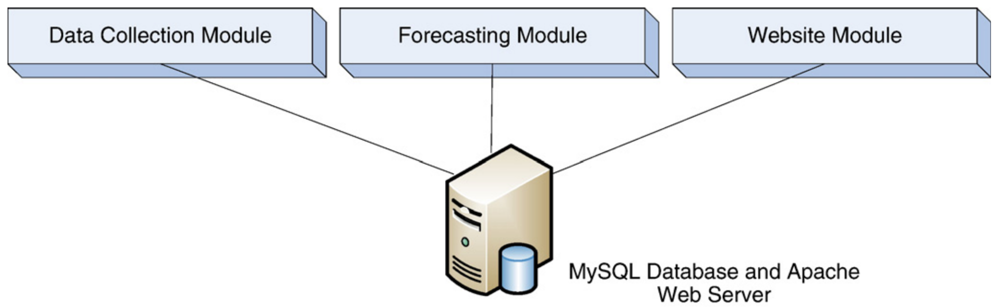

Fig. 1. Air pollution forecasting system modules.

make no prior assumptions concerning the data distribution. ANN is capable of modeling highly non-linear relationships and can be trained to accurately generalize when presented with a new data set.

The Environmental Information System for Planners (EISP) is a proof of concept web-based system designed to support decision making in many countries including the UK planning framework by making information on environmental issues more widely accessible (Culshaw et al., 2005). Recently, the Turkish Government is promoting e-government programs, which are intended to increase use of information technology and web-based services in government administration.

In this paper an online air pollution forecasting system using neural networks is presented. The system is originally based on the study of air pollution prediction model NN-AirPol (Karaca et al., 2006a). An initial online system AirPolTool (Nikov et al., 2005) was developed. Since then, the system has been improved and optimized by examining a number of alternative models. Prediction accuracy is greatly improved. The system can now predict air pollution for all ten districts of Istanbul. The web site has been overhauled. Web services are implemented to enable service oriented architecture software development.

This paper is organized as follows: Section 2 introduces the Air Pollution Forecasting System which consists of three modules: (i) data collection, (ii) forecasting, and (iii) web site modules. In Section 3, a set of experiments on the prediction of air pollutant levels using different models are presented and their results are discussed. The conclusions are presented in Section 4.

2. The air pollution forecasting system

The air pollution forecasting system is a web-based system which gathers meteorological data and air pollution data from the related web sites, forecasts air pollutant levels for the near future, publishes the results on a web site. The system mainly consists of three modules shown in Fig. 1: Data Collection Module, Forecasting Module, and Web Site Module. In this section these modules are briefly introduced.

The software used in the system consists of a database server MySQL version 5 (www.mysql.com), a web server Apache version 2 (www.apache.org), a scripting language PHP (www.php.net) and MATLAB (Palm, 2003) as shown in Fig. 2. MATLAB NN-toolbox is used to generate ANN models for the system. The Apache web server is used for publishing the results as a web page. It invokes application programs that perform various tasks such as conversion, visualization, forecasting, etc. The PHP interpreter interprets the scripts for gathering meteorological data and generating charts on the web site. The MySQL database system stores meteorological data and air pollution data. The database consists of two tables: *Annual* and *Estimation*. The *Annual* table stores daily weather forecast data and measured levels of three pollutants (PM_{10}, SO_2 and CO). The *Estimates* table holds +1, +2, and +3 days estimated values of pollutant levels.

2.1. Data Collection Module

In order to forecast air pollution indicator levels, three day meteorological forecasts and daily measurements of air pollutant levels are collected from related web sites and stored in a database every day by a PHP script. The collected data are:

(i) Daily meteorological forecast data: Meteorological data is available from http://bbc.co.uk which provides day temperature, night temperature, humidity, pressure, wind speed, wind direction, daily condition values. The gathered data is properly parsed and placed in the database and updated twice in a 24-hour period.
(ii) Air pollution measurement data: The following air pollution indicators are measured and published by Istanbul Metropolitan Municipality on their official web

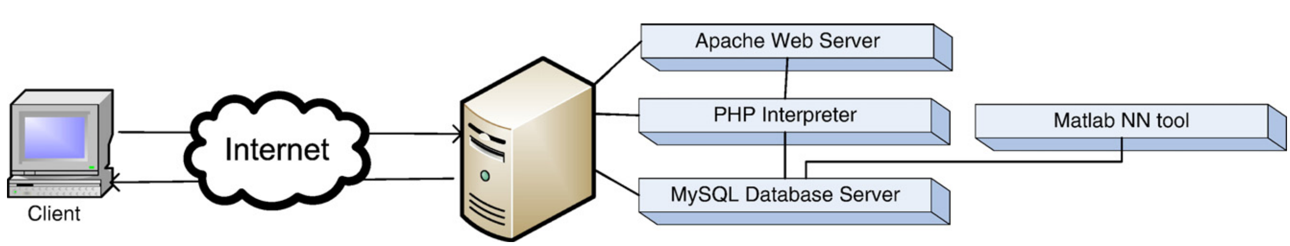

Fig. 2. The system architecture.

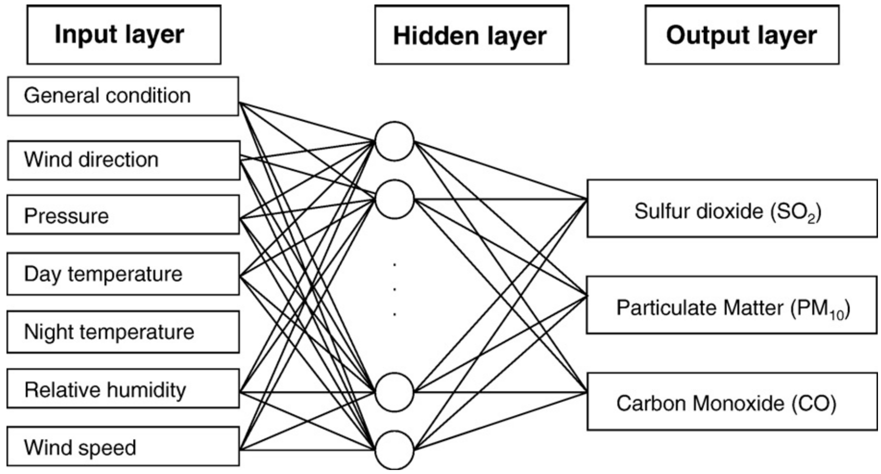

Fig. 3. The neural network model.

site (http://www.ibb.gov.tr): SO_2, PM_{10}, CO, NO, NO_x, NO_2, NMHC, O_3, CH_4 and THC. These ten air pollution indicators for ten districts in Istanbul are monitored and published daily. This data is retrieved and placed in the database.

2.2. Forecasting Module

Air pollution forecasting system is a real time system and it has been working since 01.08.2005. Meteorological air conditions and air pollutant levels are used for building a forecasting model with ANN. The model is then used for predicting near future (+1, +2, and +3 days) pollutant levels. A feed forward back propagation neural network implemented in MATLAB is employed in the system. The ANN model consists of seven input nodes in input layer and ten nodes in hidden layer as shown in Fig. 3. Hyperbolic tangent sigmoid function is used as transfer function. A training function, Levenberg–Marquardt optimization, updates weight and bias values.

2.3. Website Module

The predicted air pollutant concentrations are published on the air pollution forecasting web site http://airpol.fatih.edu.tr

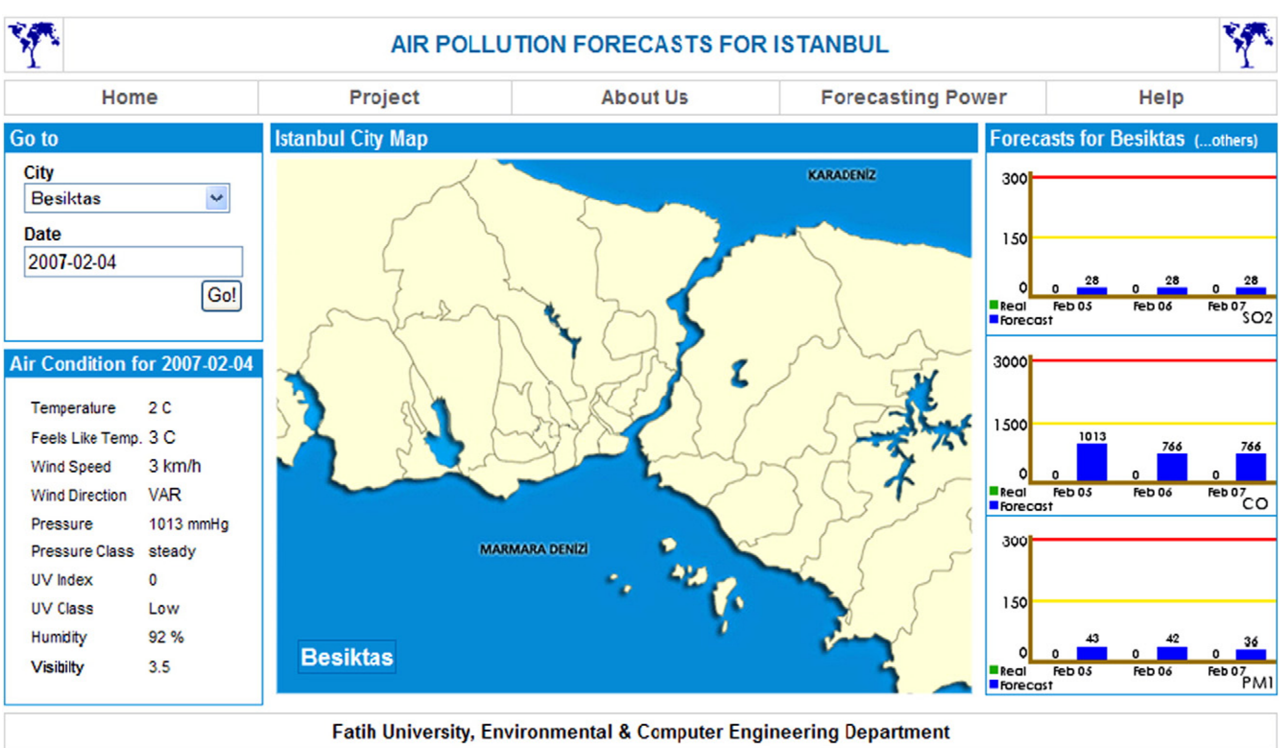

Fig. 4. The web page (http://airpol.fatih.edu.tr).

(Fig. 4). The web site provides forecasts for ten different districts in Greater Istanbul area. Currently, three of the criteria pollutants (SO_2, PM_{10} and CO) are published.

The forecast results fall into one of three different categories: Good/healthy, warning, and hazardous/dangerous. Alternatively, they could be roughly considered as low, medium, and high pollution levels separated by the yellow and red lines on the charts (appears on the right side of the web page).

The past air pollution data and air pollution forecasts are also available on the web site through a search/query facility. Users can specify a date and a district to retrieve past data (the top left side of the web page). The forecasts are shown on the left as charts.

3. Experiments

The purpose of the experiments in this study is to improve the prediction power of the air pollution parameters up to three days ahead of time by examining various neural network models. Most of the similar studies in this subject are concerned with forecasting only the next day's (tomorrow) air pollution, since forecasting two days or three days into the future are naturally more difficult (Gülbağcı, 2006).

In this section three experiments are set up. In the former part of the first experiment, next three days (+1, +2, and +3 days) are predicted independently using same training data and in the latter part +2 and +3 days are predicted cumulatively using previous days' predicted values. The same neural network architecture is used in both parts. The objective of the second experiment is to study the optimum number of past days used in the training of neural networks to achieve better predictions. Finally, in the third experiment, the effect of the day of week as an input parameter is investigated.

3.1. Data set

The data set used in the experiments is being collected from August 2005 to July 2006. Even though there are ten districts in Istanbul, only one region (Besiktas) is used to conduct the experiments, since there is nearly complete air pollution data for that region. Most other districts have missing data due to instrumental faults especially in the holidays and weekends. The missing data accounts for 22% of SO_2, 8% of CO and 10% of PM_{10} data sets for Besiktas. As data treatment procedure, the missing data is replaced by the average of nearest two days. Naturally, missing data introduces extra error to the system. The data sets used in the experiments are available in http://airpol.fatih.edu.tr/project.php. The attributes (inputs and outputs) used in ANN, ranges, means and standard deviations are given in Table 1. For example; air condition has 29 different air conditions such as sunny, stormy, etc and numerical values are assigned to those nominal values according to their order.

3.2. The prediction accuracy (absolute and relative error)

The error rate is reported as *absolute error* (AE) and *relative error* (RE) in the figures. AE represents the difference between the actual amount and the predicted amount, while RE represents the difference between the actual and the predicted *bands*. A band is defined (or measured) as a range of values. AE and RE are calculated as;

Table 1
AirPolTool Model Parameters

Attribute	Parameter	Unit	Range [min; max]	Mean	SD
Input	Condition	No unit	[1; 29]	4	5
Input	Day temperature	°C	[1; 37]	18	8
Input	Night temperature	°C	[−1; 29]	11	7
Input	Humidity	%	[26; 100]	66	15
Input	Wind speed	km/h	[1; 47]	12	8
Input	Wind direction	°	[0; 360]	51	100
Input	Pressure	mm Hg	[997; 1033]	1017	6
Output	SO_2	$\mu g/m^3$	[0; 63]	18	12
Output	CO	$\mu g/m^3$	[311; 2846]	960	451
Output	PM_{10}	$\mu g/m^3$	[9; 206]	53	32

$$AE = |X_a - X_p/X_a|*100 \tag{1}$$

$$RE = |B_a - B_p| \tag{2}$$

Here;

X_a= Actual or measured value
X_p= Predicted value
B_a= Actual or measure band
B_p= Predicted band

In this study, the pollutant ranges are divided into 5 bands or intervals. For the end users, the warning signal in the form of color or explanation of the band or the interval do carry a meaning not the actual pollutant concentration. Because of this, using absolute error could be misleading. For example, predicting 10 instead of 5 produces 100% absolute error. However in relative error that is considered 0% if both 10 and 5 values are in the same band. Hence both absolute and relative errors are included in the experiments. The ranges are defined according to minimum and maximum values of the data sets (Table 1). The relative errors—normalized from 365 days—for all experiments are shown in Table 6 using 3 days of training data.

3.3. Experiment 1: independent vs. cumulative forecasting

This experiment is designed to assess the effect of independent and cumulative predictions on the accuracy of the model. The predictions for +1, +2, and +3 days are performed independently in the Part 1 and cumulatively in the Part 2. In the Part 1, a single neural network model is used for predicting +1, +2, and +3 days pollution levels separately. Three independent neural networks are used for SO_2, PM_{10} and CO. In the Part 2, the cumulative model construction which is performed for each indicator as follows:

(i) Create a model and predict the pollution level for the next day (+1 day) using the given training data.

Table 2
Comparison of independent and cumulative predictions (values are given as AE)

Predicted day	SO_2		PM_{10}		CO	
	Part 1 (%)	Part 2 (%)	Part 1 (%)	Part 2 (%)	Part 1 (%)	Part 2 (%)
+1 day	43	43	34	33	32	32
+2 day	52	48	41	41	37	32
+3 day	52	51	47	46	38	36

(ii) Append the prediction of +1 day to the training data, create a new model using this new training data, and predict pollution level for +2 day.
(iii) Append the prediction of +2 days to the training data, create a new model using this new training data, and predict pollution level for +3 day.

In Table 2, the values in Part 1 and Part 2 columns show the obtained absolute errors in percentages. Overall, the cumulative method produces better results as shown both in Table 2 using absolute error and in Table 6 using relative error. The highest gain is on +2 day in CO and on +2 day in SO_2.

3.4. Experiment 2: optimization of the training data set size with cumulative method

The objective of this experiment is to find the optimum training data set size. In this experiment, the performance of different numbers of days—from 3 days to 15 days—are compared. Considering more than 15 days is not of much practical value, since—as revealed in these trials—it doesn't provide any advantage and even deteriorates the performance. The absolute errors of predictions are shown in Table 3. The most recent data is used in training. The pollution levels of next day (+1 day) can be predicted with higher accuracy than the pollution level of day after tomorrow (+2 day).

The error is generally higher for SO_2 than the other two pollutants. SO_2 predictions for +2 and +3 days have high absolute errors but relative errors are still acceptable as shown in Table 6. As more training data is used, the error goes higher. Interestingly, using only the most recent three days, today and the two days before today produces the best SO_2 predictions. Using more than 7 days completely confuses the model and yields unexpected values.

Carbon monoxide levels can be predicted with higher accuracy than SO_2 levels. The lowest error is achieved by using previous three days data in training. If more training data is used, the error gets higher until nine days for +1 and +2 days. The results showed that using more than nine days of data in training is not practical and yields no useful prediction. The most interesting observation is that prediction of CO levels for tomorrow (+1 day) is more difficult than for day after tomorrow (+2 day). Using only previous three days in training produces the best results in both cases (SO_2 and CO).

The error rate for +1, +2, +3 days predictions of PM_{10} levels are less than 45%. The best prediction is obtained for +1 day. Naturally higher error rates for +2 and +3 day are realized. The error goes up slightly as more training data is used in the prediction of +1 day PM_{10} levels. Less than 35% error rate is possible using 3–6 days of data. Inversely, the error rates for +2 and +3 day go down as more training data is used. As mentioned by Karaca et al. (2005b), PM_{10} has a seasonal behavior. The increasing performance can be explained by this seasonality.

Overall, it was observed that this cumulative method, which uses optimized training data set, produces higher accuracy than the independent method, which uses all of the past data as training data set (cf. Experiment 1, Part 2).

Table 3
3–15 days absolute errors

No. of days	+1 day			+2 days			+3 days		
	SO_2 (%)	CO (%)	PM_{10} (%)	SO_2 (%)	CO (%)	PM_{10} (%)	SO_2 (%)	CO (%)	PM_{10} (%)
3	43	32	33	48	32	41	51	36	46
4	47	34	37	51	33	41	53	35	46
5	48	35	35	50	33	43	54	35	44
6	45	36	35	52	34	43	54	35	45
7	50	35	39	53	34	41	56	36	42
8	47	36	38	52	34	39	54	37	43
9	50	37	37	51	35	39	56	36	43
10	49	34	36	55	35	40	54	35	44
11	50	35	38	52	34	40	53	35	42
12	49	40	37	53	33	39	53	36	44
13	50	35	36	54	35	39	52	37	43
14	51	32	36	54	33	39	51	34	40
15	51	34	37	53	34	39	52	37	42

3.5. Experiment 3: effect of day of week as input parameter

Previous studies (Karaca et al., 2005b, 2006b) reported that air pollution levels changes depending on conditions such as traffic density, closed factories on weekends, etc., during any day of week. Daily average air pollution concentrations are examined in Besiktas from January 2005 to December 2005 as shown in Table 4. The objective of this experiment is to evaluate the effect of day of week as an input parameter in air pollution forecasting.

The values in Table 4 show that, air pollution is highest on Thursday and lowest on Sunday. Numbers 1 to 7 is assigned to each day representing Sunday to Thursday and the day of week is used as an input parameter in ANN. The other input parameters in ANN are the same as in Experiment 2. The same

Table 4
Daily average air pollution concentrations in Besiktas from January 2005 to December 2005

Day of week	SO_2 ($\mu g\ m^{-3}$)	PM_{10} ($\mu g\ m^{-3}$)	CO ($\mu g\ m^{-3}$)
Monday	16	53	904
Tuesday	17	50	941
Wednesday	17	52	910
Thursday	18	52	1027
Friday	17	52	1004
Saturday	16	51	1024
Sunday	14	47	818

Table 5
Comparison of obtained performances in Experiment 2 and Experiment 3

Predicted day	SO_2		PM_{10}		CO	
	Experiment 3 (%)	Experiment 2 (%)	Experiment 3 (%)	Experiment 2 (%)	Experiment 3 (%)	Experiment 2 (%)
+1 day	43	43	35	33	28	32
+2 day	45	48	39	41	32	32
+3 day	48	51	40	46	36	36

data set is used in both experiments and the error percentages are shown in Table 5.

The error percentages for SO_2 are nearly same as the results of Experiment 2. Error rates for +1 and +2 days are 2% lower than the previous experiment. However the error rates for PM_{10} and CO are much lower while than the results obtained in Experiment 2. Especially for +3 day, the day of week parameter causes 5% lower error percentage. The error rate for CO prediction of +1 day pollution level decreases from 32% to 28% compared to the results obtained in Experiment 2. As a concrete result, including the day of week as an input parameter makes better forecasts with higher accuracy.

3.6. Discussions on error rates

The error rates achieved in this study is affected by a number of factors as explained below;

(i) Meteorological data is measured in one station in Istanbul until recently, which means that for ten different locations the same meteorological data is used in the experiments. This has an adverse effect on forecast. As a result, lower than possible prediction accuracy rates are achieved. All available meteorological data sources will be used, as they become available in the future.

(ii) Another important point is the use of absolute or relative errors in calculation. Absolute error produces higher values (indicates lower performance) but it is a very important indicator to compare performances while studying on model optimization. The relative error rates can be lower, but more meaningful for the end user, as the results are usually expressed in intervals of values. Table 6 presents the relative error rates, the number of day with incorrect band predictions out of one year (356 predictions) in the experiment.

(iii) Another aggravating factor is the missing values and the method for handling them as indicated before. The missing value (replaced with the mean of the closest existing values) ratio is 10–20% for the above experiments.

4. Conclusion

A web-based system for air pollution prediction for ten districts in Istanbul is developed. Neural networks based models are used to predict the pollutant levels for the next three days. Satisfactory prediction results are obtained.

The system provides to municipality managers and government authorities some useful and necessary information to enable them to take relevant actions to reduce emissions of air pollutants to non-harmful levels.

The outcome of this study can be summarized in a few points;

(i) an air quality prediction system in Istanbul;
(ii) an interactive map and the visualization of predicted air pollutants;

Table 6
Relative errors

	Experiment 1			Experiment 2			Experiment 3		
	+1 Day	+2 Day	+3 Day	+1 Day	+2 Day	+3 Day	+1 Day	+2 Day	+3 Day
SO_2									
No. of days 1 band off	41	51	51	37	44	50	47	48	52
No. of days 2 bands off	14	19	16	14	16	17	11	20	17
No. of days 3 bands off	5	5	6	5	7	6	4	6	6
Relative error (%)	16	20	20	**15**	**18**	**20**	17	20	21
PM_{10}									
No. of days 1 band off	20	26	33	16	19	29	17	14	21
No. of days 2 bands off	0	4	3	0	3	5	0	2	3
No. of days 3 bands off	1	0	0	1	1	0	1	1	0
Relative error (%)	6	8	10	5	6	9	5	5	7
CO									
No. of days 1 band off	31	46	37	30	34	40	16	21	25
No. of days 2 bands off	2	4	9	2	4	8	1	2	5
No. of days 3 bands off	0	0	0	0	1	0	0	0	0
Relative error (%)	9	14	13	9	11	13	5	6	8

(iii) an effective support for proper actions by managers and relevant authorities
(iv) warnings for sensitive groups such as elderly people, children, asthmatics who should stay at home on the days with dangerous levels of air pollution.

As a conclusion, the performance of AirPolTool is increased by the experiments carried out during this study. Some points, which increase the performance of air pollution predictions, are addressed. These points can be summarized as;

(i) A cumulative method is proposed to obtain better air pollution predictions for +2 and +3 days predictions.
(ii) The optimization of the number of days, used in training, increases the prediction performance.
(iii) Finally, including the day of week as a parameter in input data set gives more precise results.

References

Akkoyunlu A, Ertürk F. Evaluation of air pollution trends in İstanbul. Int J Environ Pollut 2003;18:388–98.

Boznar M, Lesjack M, Mlakar P. A neural network based method for short-term predictions of ambient SO_2 concentrations in highly polluted industrial areas of complex Terrain. Atmos Environ 1993;270:221–30.

Chaloulakou A, Saisana M, Spyrellis N. Comparative assessment of neural networks and regression models for forecasting summertime ozone in Athens. Sci Total Environ 2003;313:1–13.

Culshaw MG, Nathanail CP, Leeks GJL, Alker S, Bridge D, Duffy T, et al. The role of web-based environmental information in urban planning — the environmental information system for planners. Sci Total Environ 2005;360(1–3):233–45.

Elbir T, Muezzinoglu A, Bayram A. Evaluation of some air pollution indicators in Turkey. Environ Int 2000;26(1–2):5–10.

Gardner MW, Dorling SR. Artificial neural networks (the multilayer perceptron) — a review of applications in the atmospheric sciences. Atmos Environ 1998;32 (14/15): 2627–36.

Gardner MW, Dorling SR. Neural network modelling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. Atmos Environ 1999a;33(5):709–19.

Gardner MW, Dorling SR. Statistical surface ozone models: an improved methodology to account for non-linear behaviour. Atmos Environ 1999b;34(1): 21–34.

Gülbağcı A.B. Air Pollution forecasting by using data mining. Master Thesis, Fatih University, 2006.

Karaca F, Ölmez I, Aras NK. A radiotracer method to study the transport of mercury (II) chloride from water to sediment and air. J Radioanal Chem 2004;259:223–6.

Karaca F, Alagha O, Ertürk F. Application of inductive learning: air pollution forecast in Istanbul, Turkey. Intell Autom Soft Co 2005a;11(4):207–16.

Karaca F, Alagha O, Ertürk F. Statistical characterization of atmospheric PM_{10} and $PM_{2.5}$ concentrations at a non-impacted suburban site of Istanbul, Turkey. Chemosphere 2005b;59(8):1183–90.

Karaca F, Nikov A, Alagha O. NN-AirPol: a neural-network-based method for air pollution evaluation and control. Int J Environ Pollut 2006a;28(3/4):310–25.

Karaca F, Alagha O, Elçi E, Ertürk F, Yılmaz YZ, Özkara Y. PM_{10} and $PM_{2.5}$ chromium concentrations of Buyukcekmece Lake Basin. Ekoloji 2006b;15(61): 16–21.

Künzli N, Kaiser R, Medina S, Studnicka M, Chanel O, Filliger P, et al. Public-health impact of outdoor and traffic-related air pollution. Eur Assess 2000;356(9232):795–801.

Monteiro A, Lopes M, Miranda AI, Borrego C, Vautard R. Air pollution forecast in Portugal: a demand from the new air quality framework directive. Int J Environ Pollut 2005;5:1–9.

Nikov A, Karaca F, Alagha O, Kurt A, Hakkoymaz H. AirPolTool: a web based tool for Istanbul air pollution forecasting and control. In: İncecik S, Muezzinoglu A, Atimtay A, editors. Proceedings of 3rd International Symposium on Air Quality Management at Urban, Regional and Global Scales; 2005. (Istanbul, Turkey).

Palm WJ. Introduction to MATLAB 7 for Engineers. McGraw–Hill; 2003.

Tayanç M. An assessment of spatial and temporal variation of sulfur dioxide levels over Istanbul, Turkey. Environ Pollut 2000;107(1):61–9.