Forecasting air pollutant indicator levels with geographic models 3days in advance using neural networks

Article *in* Expert Systems with Applications · December 2010

DOI: 10.1016/j.eswa.2010.05.093 · Source: DBLP

CITATIONS
43

READS
114

2 authors:

Atakan Kurt
Istanbul University
22 PUBLICATIONS **152** CITATIONS
SEE PROFILE

Ayse Betul Oktay
Istanbul Medeniyet Universitesi
19 PUBLICATIONS **95** CITATIONS
SEE PROFILE

Expert Systems with Applications 37 (2010) 7986–7992

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

Forecasting air pollutant indicator levels with geographic models 3 days in advance using neural networks

Atakan Kurt [a,*], Ayşe Betül Oktay [b]

[a] Fatih University, Department of Computer Engineering, 34500 Istanbul, Turkey
[b] Gebze Institute of Technology, Department of Computer Engineering, 41400 Kocaeli, Turkey

ARTICLE INFO

Keywords:
SO_2
CO
PM_{10}
Neural networks
Air pollution
Air quality
Forecasting
Prediction
Data mining
Geographic forecasting
Spatial classification
Environmental expert systems

ABSTRACT

An early warning system for air quality control requires an accurate and dependable forecasting of pollutants in the air. In this study methods based on geographic forecasting models using neural networks (GFM_NN) are presented. The air pollutant data from 10 different air quality monitoring stations in Istanbul was used in forecasting sulfur dioxide (SO_2), carbon monoxide (CO) and particulate matter (PM_{10}) levels 3 days in advance for the Besiktas district. Daily meteorological forecasts as well as the air pollutant indicator values were used as input to feed-forward back-propagation neural networks. The experimental verification of the models was conducted in one-year period between August 2005 and August 2006. The observed and forecasted bands were used to compute the forecasting error. The simplest geographic model proposed uses the observed air pollution indicator values from a selected neighboring district. Where as the second model uses two neighboring districts instead of one. A third model considers the distance between the triangulating districts and the district whose air pollutant level is being forecasted. Each model is tested with at least two different sets of sites. The findings are quite satisfactory. When the right neighboring districts are chosen, the geographic models always yield lower error than non-geographic models. The distance-based geographic model produces considerably lower error than the non-geographic plain model. We argue that models proposed here can be used in urban air pollution forecasting.

1. Introduction

1.1. General overview

Air pollution is a fundamental problem in many parts of the world. It is usually caused by energy production from power plants, industrial processes, residential heating, fuel burning vehicles, natural disasters, etc. The human health concerns are one of the important short-term consequences of air pollution, especially in metropolitan areas. The global warming and the greenhouse effect are among the long term consequences on the global climate.

Air pollution related problems have resulted in an increased public awareness of the air quality in both developing and developed countries. The pressure by the environmentalist, non-governmental organizations and the public has forced governments to take regulatory actions in many countries. National and international environmental agencies such as European Environment Agency (EEA) and Environmental Protection Agency (EPA) have been established to work out the necessary guidelines and the legal ground work on air pollution. The admissible levels of air pollutants are usually determined by these organizations. The air quality guidelines set forth by them describe in detail the standards in the production, reduction, control, measurements, and dissemination of data on air pollution.

There are many air pollutants adversely affecting human health in the polluted air such as carbon monoxide (CO), particulate matter (PM_{10}), ozone (O_3), nitrogen dioxide (NO_2), etc. Sulfur dioxide (SO_2) is an important pollutant among these (Karaca, Alagha, & Ertürk, 2005; Karaca, Nikov, & Alagha, 2006). The high concentrations of these pollutants can be life threatening, causing breathing difficulty, headache, and dizziness. They may even result in heart attacks (Künzli et al., 2000). Thus, authorities advise the monitoring and forecasting criteria pollutants in the air.

Monitoring of criteria pollutants is an expensive work. It requires skilled personnel. On the other hand, forecasting is simple, low cost and efficient.

There are only a few online air pollution or air quality forecasting systems in use today. The UK national air quality information archive and NOAA-EPA Air Quality Forecasting System (USA)[1] are two examples. The Turkish Department of Environment monitors

* Corresponding author. Tel.: +90 212 866 3300x5513; fax: +90 212 866 3412.
E-mail address: akurt@fatih.edu.tr (A. Kurt).

[1] For detailed information, one may visit to related websites on http://www.airquality.co.uk/ and http://www.arl.noaa.gov/ready/aq.html

0957-4174/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2010.05.093

air quality for a number of cities in Turkey.[2] Over 12 million people lives in Istanbul where the air quality is monitored at 10 monitoring stations established and operated by Istanbul Metropolitan Municipality (IBB). The measurements of 10 important air pollutants are announced daily at the IBB official web page.[3] There is no air pollution forecasting system for criteria pollutants in Turkey except the one established by the authors in operation since August 2005.[4] The system uses the past air pollution data and meteorological parameters to forecast air pollutant levels using neural networks.

Neural networks capable of modeling non-linear relationships between input and output variables are often used in forecasting variables in complex systems (Gardner & Dorling, 1998). They are supervised learning techniques involving a training step to create a mathematical model and a prediction step to compute the output for a given set of input values using the model created in the training step. Our work in this study is primarily concerned with forecasting SO_2, CO, PM_{10} levels using *spatial parameters via geographic modeling* for the Besiktas district in Istanbul. The reason for using geographic models in forecasting air pollution is that the atmosphere itself is a geographic entity. In fact the atmosphere is a dynamic intricate system which is quite difficult to model. The air and the pollutants in it move in different ways and directions mainly through natural causes and atmospheric phenomena. These phenomena include temperature, wind, pressure, humidity, etc. The air in close geographic areas intermixes and interacts through these phenomena. We believe that a forecasting system taking the spatial parameters in consideration can more accurately forecast pollutant levels.

This paper is organized as follows: the related work is given below. Section 2 introduces the data set, the environment in which the experiments were conducted and the general architecture of neural networks used in the experiments. The models, the experiments on the models and the results are presented in Section 3. Results are discussed in Section 4. The conclusions and the future work are given in Section 5.

1.2. Related work

In the last decade many applications of neural networks in atmospheric sciences have appeared in the literature. Gardner and Dorling have presented a study on the prediction of nitrogen oxides concentrations (Gardner & Dorling, 1999) and ozone (Gardner & Dorling, 2001). Perez (2001) presented a study on the 8 h early prediction of hourly mean SO_2 values at a location near Santiago, Chili using hourly average concentrations computed at 6 h intervals as well as past temperature, relative humidity and wind speed values. An average 30% error was obtained in the best case. Chelani, Rao, Phadke, and Hasan (2002) predicted SO_2 values at three sites in Delhi using neural networks and compared the results with those of multivariate regression models. Wind speed, wind direction index, relative humidity, and temperature parameters were used as input for a recurrent neural network. The errors obtained by neural networks were lower than multivariate models. Air pollution index reporting system presented by (Jiang et al., 2004) uses date, maximum and average temperatures, pressure, humidity, wind, cloud coverage, daily precipitation as neural network input variables to forecast daily average TSP, SO_2, PM_{10}, and NO2 values. Correlation between the observed and forecasted SO_2 values was 0.7 in this system. Nunnari et al. (2004) compared a number of methods including neural networks, time-series, fuzzy-logic, etc. to predict daily mean, daily maximum and hourly mean SO_2 values using hourly SO_2 values. The results show that there is no best model generating the optimum performance and that a combination of models can be used. The SO_2 and TSP concentrations for the city of Zonguldak, Turkey in winter were predicted 1 day in advance using a neuro-fuzzy technique by Yildirim and Bayramoglu (2006). The prediction of SO_2 with neural networks was also studied by Castro, Prada Sanchez, Gonzalez Manteiga, and Febrero Bande (2003) and Brunellia, Piazzaa, Pignatoa, Sorbellob, and Vitabilec (2008). Ibarra-Berastegi et al. (2008) focused on the hourly prediction of SO_2 and four other pollutants up to 8 h ahead in six locations in Bilbao, Spain with neural networks modeling. Many types of neural network architectures were tried and the best one was chosen depending on the pollutant type, geographic locations, and the number of hours ahead for forecasting.

Our non-geographic model is based on a previous study (Kurt, Gulbagci, Karaca, & Alagha, 2008; Nikov, Karaca, Alagha, Kurt, & Hakkoymaz, 2005). Similar studies are primarily concerned with obtaining optimum neural network architectures to predict pollutant levels or values. Therefore, experiments with a variety of alternative architectures were conducted. Related work also varies in the input meteorological variables, the time frequency of the measurements of input variables, and the kind of predictions computed. Our approach is unique in the sense that it employs geographic models. In order to compare the results with the non-geographic model, the same data set and the same network architecture and the same location were used in all experiments. The aim of this work is to introduce new forecasting models based on geographic parameters.

2. Experimental setup

We present the environment in which data was collected and the experiments were conducted in this section. The error measure is also given here.

2.1. Airpol: an air pollution forecasting system

A web-based forecasting system called Airpol (Kurt et al., 2008; Nikov et al., 2005) is used for collecting data and conducting experiments in this study. A brief description of the system architecture is given in Fig. 1. Airpol is an air pollution forecasting system for 10 districts in Istanbul (Fig. 2) which gathers, forecasts, and publishes 3 day forecast of three air pollutants (SO_2, CO, PM_{10}) using a non-spatial models with neural networks. Airpol is a real time forecasting system in operation since August 2005. It is written in PHP, MySQL, and MATLAB. The system consists of three modules: data collection, forecasting, and web publishing. The data is collected from two different sites by the collection module and placed in a central database:

– *3-day meteorological forecast and observed (current) meteorological data* retrieved from BBC web site which includes: day temperature, night temperature, humidity, pressure, wind speed, wind direction, and daily-condition.
– *Observed (current) air pollution data* retrieved from the IBB web site which provides the daily measurements of 10 air pollution indicators (SO_2, PM_{10}, CO, NO, NO_x, NO_2, NMHC, O_3, CH_4 and THC). This data originates from the air pollution stations placed throughout the city in 10 different districts.

2.2. Data set

The data set used in the experiments is being collected since August 2005. Even though there are 10 districts where the

[2] Republic of Turkey Ministry of Environmental and Forestry, Air Quality Monitoring Web Site, http://www.havaizleme.gov.tr

[3] Istanbul Metropolitan Municipality, Air Quality Information Web Site; http://application2.ibb.gov.tr/NewIBBUserControls/HavaKalitesi.aspx which can be accessed through the main site http://www.ibb.gov.tr

[4] The Airpol Project, http://airpol.fatih.edu.tr

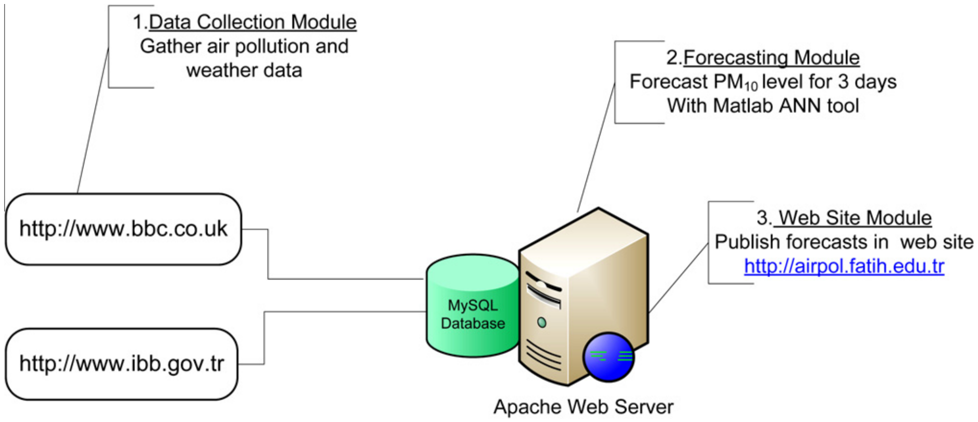

Fig. 1. Airpol system architecture.

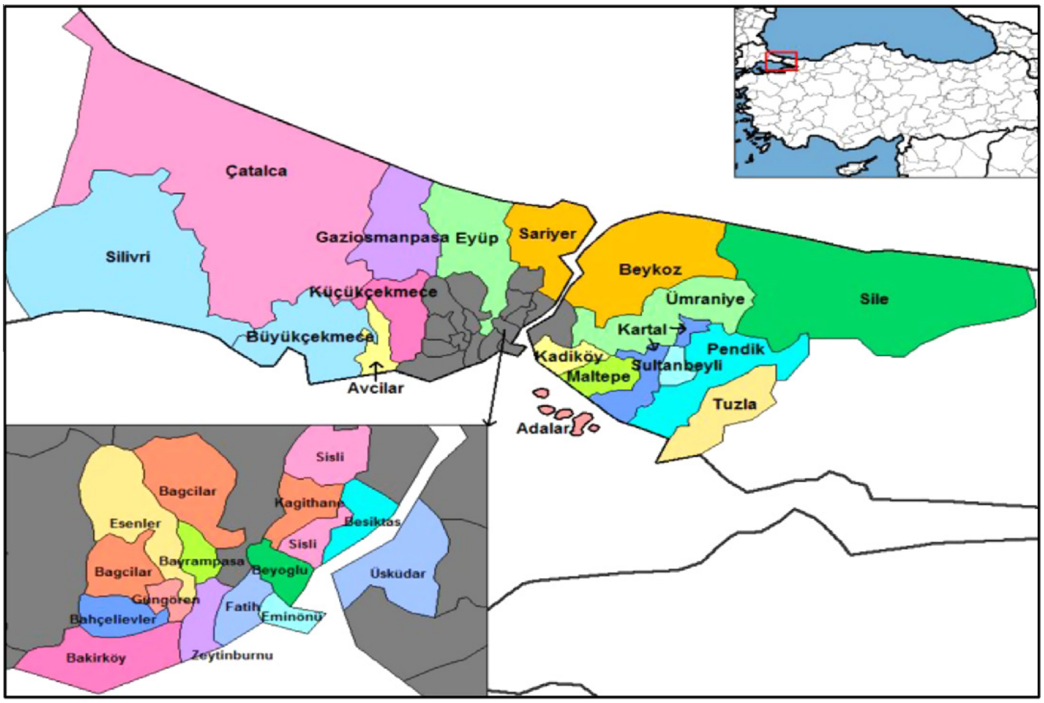

Fig. 2. Greater Istanbul Area and the districts with monitoring stations (bottom-left).

measurements are taken in Istanbul, we chose Besiktas district to conduct the experiments, because this district has the least amount of missing air pollution data among all districts. The data for some holidays and weekends is missing in most districts. We used the mean of the two known nearest values to substitute the missing values. Since the data becomes roughly periodic after one-year period, only one year's (August 2005 through July 2006) data is used in the experiments. The observed meteorological and air pollutant level values are stored in the database daily. The 3-day meteorological forecasts (seven variables) from BBC web site and the forecasted air pollutant values by the system are also stored in the database.

A statistical summary of the variables, including their ranges, mean and standard deviations are given in Table 1. The input variables are general condition, day (high) temperature, night (low) temperature, relative humidity, wind speed, wind direction, pressure, day-of-week, date, SO_2, CO and PM_{10}. The *general condition* attribute has 29 different air conditions such as *sunny*, *stormy*, etc. The day-of-week attribute is an important input parameter in the prediction of air pollution as shown in (Gülbağci, 2006). Numbers 1 through 7 are assigned to Sunday through Saturday in the day-of-week attribute. The input values belong to certain

Table 1
Air pollutant predictor variables.

Variable	Unit	Range	Mean	Standard deviation
General condition	–	[1–29]	–	–
Day temperature	°C	[1–37]	18	8
Night temperature	°C	[−1 to 29]	11	7
Humidity	%	[26–100]	66	15
Wind speed	km/h	[1–47]	12	8
Wind direction	°	[0–360]	51	100
Pressure	mm Hg	[997–1033]	1017	6
Day of week	int	[1–7]	–	–
Date	int	[0–14]	–	–
SO_2	$\mu g/m^3$	[69–2861]	948	448
CO	$\mu g/m^3$	[311–2846]	960	451
PM_{10}	$\mu g/m^3$	[9–206]	53	32

date. Integer date values are used to indicate to how many days before the data belongs (0: today, 1: the day before today, 2: 2 days before today). Values are automatically normalized by MATLAB neural network toolbox prior to processing.

2.3. Error measure

The error is usually reported as *band error* in air quality web sites. Band error represents the difference between the observed and forecasted bands in which the observed and forecasted values fall. SO_2, PM_{10}, CO value range is divided into five intervals or bands. The bands for SO_2 are: [0–12], [13–24], [25–36], [37–48], [49–60]. The bands for PM_{10} are: [0–60], [61–120], [121–180], [181–240], [241–300]. The bands for CO are: [0–600], [601–1200], [1201–1800], [1801–2400], [2401–3000].

The five intervals system is chosen because it is the preferred method in air quality reporting. For example, the actual and forecasted pair (12, 13) for SO_2 is reported as +1 band in the band error method, since 12 falls into the first interval, and 13 falls into the second interval.

2.4. Neural network architecture

In this section, we present the architecture of neural networks employed in forecasting SO_2, PM_{10} and CO levels 3 days into future (denoted by +1 day, +2 day, and +3 day) in the experiments below. The neural network represents a forecasting model for the output variables SO_2 based on the input variables shown in Fig. 3. The air pollutants PM_{10} and CO are forecasted using the same neural network model with the appropriate input pollutant variables.

The neural network consists of 10-node input layer, 10-node hidden layer, and a single node output layer. Feed-forward backpropagation neural networks are employed in the system. Hyperbolic tangent sigmoid function is used as the transfer function. The Levenberg–Marquardt optimization is applied for training weights and bias values. The neural networks are implemented with MATLAB neural network toolbox.

Since forecasting is performed for 3 days into future, there are three distinct neural networks with the same architecture represented with triangle symbols in Fig. 4. In this figure, solid-edge and dotted-edge rectangles represent days with observed and forecasted values. These three neural networks are connected together in a cascaded fashion. The cascading is preferred because the output air pollution indicator values depend not only on past meteorological variables, but also on the previously forecasted air pollution indicator values. Therefore, when forecasting air pollutant levels for +2 and +3 day, the system should use meteorological forecast data and the previous day's predicted air pollution values. It has been shown that cascaded model is able to forecast +2 and +3 days better than the straight model in which previously forecasted air pollutant values are not used (Gülbağci, 2006).

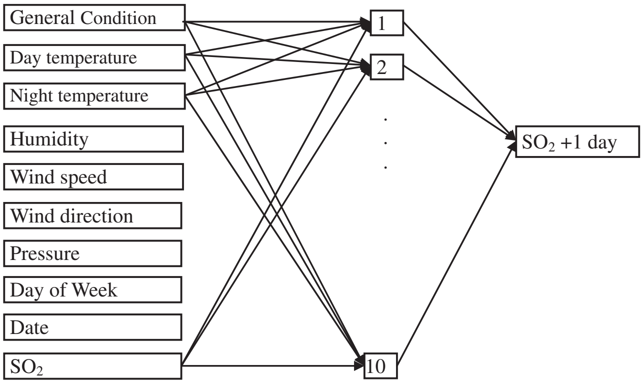

Fig. 3. +1 day neural network forecasting SO_2.

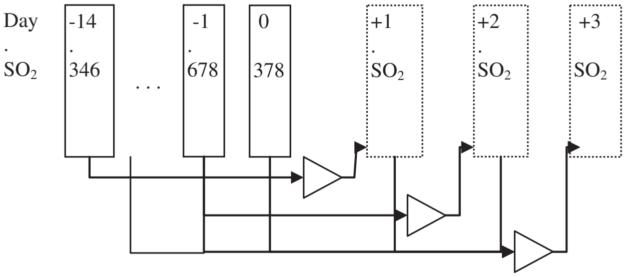

Fig. 4. Multi-day neural network architecture.

3. Results

The purpose of the experiments in this study is to examine various geographic forecasting models for SO_2, PM_{10} and CO up to 3 days ahead of time. Most of the similar studies in this area are concerned with forecasting only the next day's (tomorrow) air pollution, since the error rates for 2 or 3 days into the future are naturally are higher and therefore more difficult. We use a non-geographic plain model in the first experiment. This model is a reference model for the geographic models used in experiments 2–4. The cascaded neural network architecture described in the previous section is used in all experiments. Only the most recent 3 days' (−2 day: 2 days before, −1 day: the day before, 0 day: current day) observed meteorological and air pollutant values are used and all experiments are performed for 365 days.

3.1. Plain model

The objective of this model is to forecast SO_2, PM_{10} and CO levels using past meteorological and SO_2, PM_{10} and CO values. We do not consider any geographic variables in this model. The architecture of this model is given in Figs. 3 and 4. A year around (365 days) prediction is performed. The observed and predicted air pollution indicator band levels are compared. The days with errors are reported. Experimental results are summarized in Table 2.

The errors for +1, +2, and +3 days are 37%, 40% and 40%, respectively, for SO_2 prediction which are considerably high for air pollution forecasting. The best forecasting is achieved for +1 day. The distribution of band errors (±1, ±2, ±3 bands) is listed for each day. The (∼70%) majority of errors are in ±1 band range. Relatively less error is reported for ±2 (∼20%) and ±3 (∼10%) bands in all cases.

Table 2
Forecasting errors for Besiktas with the plain model.

Forecasting day	Error	SO_2	PM_{10}	CO
+1 day	±1 band	100 (27%)	81 (22%)	95 (26%)
	±2 bands	26 (7%)	2 (0.5%)	13 (3%)
	±3 bands	8 (2%)	1 (0.02%)	0 (0%)
	Total	134 (37%)	84 (23%)	108 (29%)
+2 day	±1 band	105 (29%)	100 (27%)	104 (29%)
	±2 bands	30 (8%)	4 (1%)	16 (4%)
	±3 bands	12 (3%)	1 (0.02%)	1 (0.02%)
	Total	147 (40%)	105 (29%)	121 (33%)
+3 day	±1 band	103 (24%)	107 (29%)	101 (27%)
	±2 bands	35 (10%)	3 (0.08%)	19 (5%)
	±3 bands	8 (2%)	4 (0.01%)	1 (0.02%)
	Total	146 (40%)	114 (31%)	121 (33%)

The errors for +1, +2, and +3 days are 23%, 29% and 31%, respectively, for predicting PM_{10}; and 29%, 33% and 33% for predicting CO. The +1 day's forecasts are better than the other forecasts and ±1 band range includes the most of the errors.

3.2. Geographic models

We believe that using geographic factors in air pollution forecast can increase the forecasting accuracy, because the districts in consideration for the forecast are in close proximity (Fig. 2). There are no rural areas between urban cities or districts in Istanbul. A district starts on a street where another one ends. We expect that there should be atmospheric interactions between the districts nearby. The geographic terrain, the geometric characteristic of the man-made structures in the city, along with a number of other geographic and environment attributes determines how the pollutants in different areas come into interaction. These complex interactions play a significant role in the forecasting of air pollution. The industrialization, the amount of traffic, the concentrations of population, and the heating methods are among the parameters considered in air pollution analysis and prediction. The location and distances between districts are also important. As the distance between districts decreases, the resemblance should be higher. In this section, we present three geographic models in SO_2, PM_{10} and CO forecasting with the increasing order of complexity: single-site neighborhood model, two-site neighborhood model, and distance-based model.

3.3. Single-site neighborhood model

This model is developed on the idea of using one or more neighboring districts' air pollution indicator value as an extra input for the current district. The selection of districts should be primarily based on the distance between the districts and the correlation of variables. When selecting the districts, we have two geographic criteria to consider. The first one is whether to select a near or far district (all districts are in a relatively small region). The second one is the number of districts to use in the model. Assuming that as the distance between districts decreases, the air pollution values resemble more, we decided to use near districts. On the second criteria, we created models with one and two districts. The experiment with one neighboring city uses one extra input air pollutant variable whereas the experiment with two cities uses two extra air pollutant input variables.

Separate experiments are conducted for the following neighboring districts and the results are given in Table 3: Esenler, Sariyer, Uskudar, Yenibosna. The distance between Besiktas and the neighboring districts are given in Table 5. The best results for each forecasted day is indicated with an asterisk (∗). Esenler produces the best results for +1 and +2 days, while Sariyer delivers the best results for +3 day for SO_2. In all three cases, the neighborhood-based model yields lower error than the plain model for SO_2. The PM_{10} forecasting errors are greater than the plain model. The CO forecasts with single-site neighborhood model give more accurate results than the plain model.

Table 3
Forecasting errors with the single-site neighborhood model.

Forecasting	SO_2 error	District			
		Esenler	Sarıyer	Üsküdar	Yenibosna
+1 day	±1 band	98 (26%)	104 (28%)	103 (28%)	116 (31%)
	±2 bands	27 (7%)	25 (7%)	20 (6%)	24 (7%)
	±3 bands	4 (1%)	7 (2%)	8 (2%)	7 (2%)
	Total	*129 (35%)	136 (37%)	131 (36%)	147 (40%)
+2 day	±1 band	102 (28%)	102 (28%)	101 (28%)	106 (29%)
	±2 bands	27 (7%)	33 (9%)	29 (8%)	29 (8%)
	±3 bands	6 (2%)	6 (2%)	5 (1%)	6 (2%)
	Total	*135 (37%)	141 (39%)	135 (37%)	141 (39%)
+3 day	±1 band	112 (31%)	96 (26%)	110 (30%)	111 (30%)
	±2 bands	27 (7%)	32 (9%)	22 (6%)	25 (7%)
	±3 bands	3 (0.8%)	6 (2%)	5 (1%)	3 (0.8%)
	Total	142 (39%)	*134 (37%)	137 (38%)	139 (38%)
	PM_{10} error	District			
		Esenler	Sarıyer	Üsküdar	Yenibosna
+1 day	±1 band	87 (24%)	85 (23%)	87 (24%)	92 (25%)
	±2 bands	4 (1%)	4 (1%)	2 (0.5%)	2 (0.5%)
	±3 bands	1 (0.02%)	1 (0.02%)	1 (0.02%)	1 (0.02%)
	Total	92 (25%)	*90 (25%)	90 (25%)	95 (26%)
+2 day	±1 band	90 (25%)	89 (24%)	95 (26%)	93 (25%)
	±2 bands	3 (0.8%)	2 (0.5%)	4 (1%)	3 (0.8%)
	±3 bands	1 (0.02%)	1 (0.02%)	1 (0.02%)	1 (0.02%)
	Total	94 (26%)	*92 (25%)	100 (27%)	97 (27%)
+3 day	±1 band	97 (27%)	98 (27%)	96 (27%)	103 (28%)
	±2 bands	2 (0.5%)	3 (0.8%)	3 (0.8%)	2 (0.5%)
	±3 bands	1 (0.02%)	1 (0.02%)	1 (0.02%)	1 (0.02%)
	Total	*100 (27%)	102 (28%)	100 (27%)	106 (29%)
	CO error	District			
		Esenler	Sarıyer	Üsküdar	Yenibosna
+1 day	±1 band	105 (29%)	100 (27%)	101 (27%)	110 (30%)
	±2 bands	10 (3%)	10 (3%)	10 (3%)	12 (3%)
	±3 bands	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Total	115 (31%)	110 (30%)	111 (30%)	122 (34%)
+2 day	±1 band	98 (27%)	101 (27%)	102 (28%)	106 (29%)
	±2 bands	15 (4%)	12 (3%)	8 (2%)	12 (3%)
	±3 bands	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Total	113 (31%)	113 (31%)	*110 (30%)	118 (32%)
+3 day	±1 band	111 (30%)	98	99 (27%)	107 (29%)
	±2 bands	9 (2%)	14 (4%)	11 (3%)	14 (4%)
	±3 bands	0 (0%)	0 (0%)	0 (0%)	0 (0%)
	Total	120 (33%)	112 (31%)	*110 (30%)	121 (33%)

3.4. Two-site neighborhood model

This neighborhood model considers two neighboring districts instead of one. The rationale for this model is that using more predictor variables should achieve higher accuracy. The results for this experiment are given in Table 4. The experiments were conducted on separate district pairs: Sariyer–Yenibosna and Uskudar–Fatih. The results are a bit surprising, because the error is higher than the plain model for +1 day in both district pairs. The Uskudar–Fatih pair produces the best results for +1 days, while the Sariyer–Yenibosna pair delivers the best results for +2 and +3 days in this experiment for SO_2. For +2 day, Sariyer–Yenibosna pair yields lower error than both the plain model and the geographic model with a single-site above. It should also be pointed out that Yenibosna and Sariyer when paired produces lower error (32% in Table 4) than the case they are used alone in the single-site model. PM_{10} and CO error percentages are nearly same as the single-site neighborhood model.

3.5. Distance-based model

In the neighborhood models above, air pollutant values in one or two near districts are considered. However, the actual distances between sites are not given any consideration. In this model we consider a model based on distance in which the distance values are used to compute a weighted average of air pollutant from the air pollutant values in three neighboring districts. The model is based on the idea that the effects of air pollutant levels of the neighboring district are inversely proportional to the distance between the two districts. Thus, we can compute the combined

Table 4
Forecasting errors with the two-site neighborhood model.

Forecasting	Error	SO_2		PM_{10}		CO	
		Üsküdar–Fatih	Sarıyer–Yenibosna	Üsküdar–Fatih	Sarıyer–Yenibosna	Üsküdar–Fatih	Sarıyer–Yenibosna
+1 day	±1 band	102 (28%)	112 (30%)	103 (28%)	87 (24%)	103 (28%)	106 (29%)
	±2 bands	30 (8%)	21 (6%)	1 (0.02%)	3 (0.8%)	11 (3%)	12 (3%)
	±3 bands	5 (1%)	6 (2%)	1 (0.02%)	1 (0.02%)	1 (0.02%)	0 (0%)
	Total	*137 (38%)	139 (38%)	105 (29%)	91 (25%)	115 (32%)	118 (32%)
+2 day	±1 band	96 (26%)	99 (27%)	107 (29%)	95 (26%)	108 (30%)	111 (27%)
	±2 bands	31 (8%)	29 (8%)	3 (0.8%)	3 (0.8%)	10 (2%)	13 (3%)
	±3 bands	8 (2%)	4 (1%)	1 (0.02%)	1 (0.02%)	0 (0%)	0 (0%)
	Total	135 (37%)	*132 (36%)	111 (30%)	99 (27%)	118 (32%)	124 (34%)
+3 day	±1 band	117 (32%)	105 (29%)	110 (30%)	98 (27%)	104 (28%)	109 (30%)
	±2 bands	29 (8%)	35 (10%)	3 (0.8%)	1 (0.02%)	13 (3%)	12 (3%)
	±3 bands	3 (0.8%)	3 (0.8%)	1 (0.02%)	1 (0.02%)	0 (0%)	0 (0%)
	Total	149 (40%)	*143 (39%)	114 (31%)	100 (27%)	117 (32%)	121 (33%)

Table 5
Distances between Besiktas and triangulating districts.

Districts	Distance	Normalized distance
(F)atih	6381	0.32
(U)sküdar	4494	0.45
(S)ariyer	8963	0.53
(K)artal	24,664	0.19
(Y)enibosna	16,580	0.28

effects of neighboring three districts' air pollutant values on the triangulated district in the center. A new weighted average air pollutant variable is created using the corresponding variables and the distances of three triangulating nearby districts using the following formula for SO_2 below:

$$\text{New variable } SO_2 = SO_2@\text{Fatih} * 0.32 + SO_2@\text{Uskudar} * 0.45 + SO_2@\text{Sariyer} * 0.23 \quad (1)$$

The coefficients represent the normalized distances given in Table 5 for the first triple below. The neural networks in this model have the air pollutant variable above as an extra variable representing the triangulating districts' effect on the prediction. We conducted two experiments by choosing two distinct triples as the triangulating districts in this model. Triangle one consists of Fatih, Uskudar and Sariyer, and triangle two consists of Kartal, Yenibosna, and Sariyer districts. The second triple consists of farther districts than the districts in first triple from the district in the center. The results are shown in Table 6 which are quite satisfactory for triangle one. Even though triangle two yields lower error for both +1 and +2 days than the plain model, it has at least 10% higher error than triangle one for all 3 days. Triangle one produces lower error than the neighborhood-based models. We see that the distance-based model with the first triple is better than the neighborhood-based models.

The accuracy of this model is mainly affected by the selection of the districts and the number of districts used. The selection of the districts is mainly based on the correlation between variables and on relative location, and distance. On the distance criterion, we preferred the near districts, so that a better comparison can be achieved with the previous experiment. On the relative location criterion, our guideline is to choose the three districts that create as good triangle as possible. The triangle should satisfy two conditions: (i) the district in the middle should be placed in the center of the triangle as much as possible. (ii) The triangle should be equilateral. In reality many triangles are possible. Each triangle produces one model. The best triangle and its model can be chosen with experimentation.

4. Discussion

A summary of the experimental results is given in Table 7. The districts are indicated by the first letter of the names. The days with errors out of 365 are given as the error measure. The best results for each model are indicated in bold type face. Single-site neighborhood model indicates Esenler as the best choice for +1 and +2 day, Sariyer for +3 day. Two-site neighborhood model with Sariyer–Yenibosna produces better accuracy than single-city model for +2 day. It should be noted that although Sariyer and Yenibosna separately produces higher error in single-site model, surprisingly they yield lower error together in two-site model. Distance-based model using Fatih–Uskudar–Sariyer is the model with lowest error. Fatih–Uskudar–Sariyer triple is closer to Besiktas

Table 6
Forecasting errors for Besiktas with the distance-based model.

Forecasting	Error	SO_2		PM_{10}		CO	
		F–U–S	K–Y–S	F–U–S	K–Y–S	F–U–S	K–Y–S
+1 day	±1 band	74 (20%)	94 (25%)	57 (16%)	83 (22%)	73 (20%)	92 (25%)
	±2 bands	15 (4%)	26 (7%)	2 (0.5%)	4 (1%)	3 (0.8%)	14 (4%)
	±3 bands	6 (2%)	8 (2%)	1 (0.2%)	1 (0.2%)	0 (0%)	0 (0%)
	Total	*95 (26%)	128 (35%)	60 (16%)	88 (24%)	76 (20%)	106 (29%)
+2 day	±1 band	68 (19%)	101 (27%)	64 (18%)	88 (24%)	68 (19%)	105 (29%)
	±2 bands	18 (5%)	34 (9%)	1 (0.2%)	7 (2%)	5 (1%)	18 (5%)
	±3 bands	11 (3%)	13 (3%)	1 (0.2%)	1 (0.2%)	0 (0%)	1 (0.2%)
	Total	*97 (27%)	148 (40%)	66 (18%)	96 (26%)	73 (20%)	124 (34%)
+3 day	±1 band	57 (15%)	97 (26%)	81 (22%)	97 (27%)	66 (18%)	104 (28%)
	±2 bands	22 (6%)	29 (7%)	2 (0.5%)	9 (2%)	10 (2%)	20 (5%)
	±3 bands	12 (3%)	10 (3%)	1 (0.2%)	1 (0.2%)	0 (0%)	1 (0.2%)
	Total	*91 (25%)	136 (37%)	84 (23%)	107 (29%)	76 (21%)	124 (34%)

Table 7
Model accuracy summary (number of days with error out of 365 days).

	Model								
	Plain	Single-site				Two-site		Distance-based	
		E.	S.	U.	Y.	U.-F.	S.-Y.	K.-Y.-S.	F.-U.-S.
Forecasting SO_2									
+1 day	134	129	136	131	147	137	139	128	95
+2 day	147	135	141	135	141	135	132	148	97
+3 day	146	142	134	137	139	149	143	136	91
Forecasting PM_{10}									
+1 day	134	92	90	90	95	105	91	88	60
+2 day	147	94	92	100	97	111	99	96	66
+3 day	146	100	102	100	106	114	100	107	84
Forecasting CO									
+1 day	134	120	112	110	121	115	118	106	76
+2 day	147	113	113	110	118	118	124	124	73
+3 day	146	115	110	111	122	117	121	124	76

than Kartal–Yenibosna–Sariyer triple. On one hand Kartal–Yenibosna–Sariyer triple has low accuracy for distance-based model, on the other hand this triple yields higher accuracy than single-site and two-site neighborhood model for +1 day, higher accuracy than two-site neighborhood model for +3 day. This means that distance alone is not the sole determinant of predictability and that many complex atmospheric and geographic variables play a role in this system. Overall, the distance-based model yields higher accuracy. In all models, the accuracy depends on the choice of district used in the model. Certain sites, site-pairs and site-triples seem to produce better results than the others. Then the best choices can be determined by a series experiments as shown above.

5. Conclusions

Air pollution forecasting models play an important role in air quality management and control if properly designed and implemented. Therefore techniques for increasing the forecasting accuracy of existing systems are quite valuable. SO_2, CO and PM_{10} are the pollutants which are mainly responsible for urban air pollution and particularly difficult to forecast. A number of geographic models for air pollutant forecasting are developed using neural networks. The forecasting accuracy of these models is studied through a set of experiments for a district in Istanbul and the results are presented. Results show that the geographic models involving spatial parameters outperform the non-geographic model. Especially three-site distance-based model outperforms all other models significantly. Results also indicate that a best nearby district or districts yielding lowest error can be experimentally determined for each forecasted day. These districts' air pollution value can be used by a real time forecasting system. The models here can easily be adapted to other regions in the world. The best neighboring cities can be adaptively obtained. An adaptive system dynamically reselecting those cities seasonally can be built.

We believe the geographic models proposed here can be employed for other pollutants. We also believe that other geographic models yielding higher accuracy can be developed utilizing this approach. For example, a model using the exact relative locations of a set of selected cities to the city in the center could be one of the models to consider.

References

Brunellia, U., Piazzaa, V., Pignatoa, L., Sorbellob, F., & Vitabilec, S. (2008). Three hours ahead prevision of SO_2 pollutant concentration using an Elman neural based forecaster. *Building and Environment, 43*(3), 304–314.

Castro, F. B. M., Prada Sanchez, J. M., Gonzalez Manteiga, W., & Febrero Bande, M. (2003). Prediction of SO_2 levels using neural networks. *Journal of the Air and Waste Management Association, 53*, 532–539.

Chelani, A. B., Rao, C. V. C., Phadke, K. M., & Hasan, M. Z. (2002). Prediction of sulphur dioxide concentrations using artificial neural networks. *Environmental Modelling and Software, 17*(2), 161–168.

Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron) – A review of applications in the atmospheric sciences. *Atmospheric Environment, 32*(14–15), 2627–2636.

Gardner, M. W., & Dorling, S. R. (1999). Neural network modeling and prediction of hourly NO_x and NO_2 concentrations in urban air in London. *Atmospheric Environment, 33*(5), 709–719.

Gardner, M. W., & Dorling, S. R. (2001). Artificial neural network derived trends in surface ozone concentrations. *Journal of the Air and Waste Management Association, 51*, 1202–1210.

Gülbağci, A. B., 2006. *Air pollution forecasting by using data mining*. Master Thesis, Fatih University, Istanbul, Turkey.

Ibarra-Berastegi, G., Elias, A., Barona, A., Sáenz, J., Ezcurra, A., & de Argandoña, J. D. (2008). From diagnosis to prognosis for forecasting air pollution using neural networks: Air pollution monitoring in Bilbao. *Environmental Modelling and Software, 23*(5), 622–637.

Jiang, D., Zhang, Y., Hu, X., Zeng, Y., Tan, J., & Shao, D. (2004). Progress in developing an ANN model for air pollution index forecast. *Atmospheric Environment, 38*(40), 7055–7064.

Karaca, F., Alagha, O., & Ertürk, F. (2005). Application of inductive learning: Air pollution forecast in Istanbul City. *Intelligent Automation and Soft Computing, 11*(4), 207–216.

Karaca, F., Nikov, A., & Alagha, A. (2006). NN-AirPol: A neural-networks-based method for air pollution evaluation and control. *International Journal of Environment and Pollution, 28*(3/4), 310–325.

Künzli, N., Kaiser, R., Medina, S., Studnicka, M., Chanel, O., Filliger, P., et al. (2000). Public-health impact of outdoor and traffic-related air pollution: A European assessment. *The Lancet, 356*(2932), 795–801.

Kurt, A., Gulbagci, A. B., Karaca, F., & Alagha, O. (2008). An online air pollution forecasting system using neural networks. *Environment International, 34*(5), 592–598.

Nikov, A., Karaca, F., Alagha, O., Kurt, A., & Hakkoymaz, H., 2005. AirPolTool: A web based tool for Istanbul air pollution forecasting and control. In S. Topcu, M. F. Yardım, A. Bayram, T. Elbir, & C. Kahya (Eds.), *Proceeding of third international symposium on air quality management at urban, regional and global scales, Istanbul, Turkey*. pp. 247–255.

Nunnari, G., Dorling, S., Schlink, U., Cawley, G., Foxall, R., & Chatterton, T. (2004). Modelling SO_2 concentration at a point with statistical approaches. *Environmental Modelling and Software, 19*(10), 887–905.

Perez, P. (2001). Prediction of sulfur dioxide concentrations at a site near downtown Santiago, Chile. *Atmospheric Environment, 35*(29), 4929–4935.

Yildirim, Y., & Bayramoglu, M. (2006). Adaptive neuro-fuzzy based modeling for prediction of air pollution daily levels in city of Zonguldak. *Chemosphere, 63*(9), 1575–1582.