AIR POLLUTION FORECASTING BY USING DATA MINING

by

A. Betül GÜLBAĞCI

August 2006

AIR POLLUTION FORECASTING BY USING DATA MINING

by

A. Betül GÜLBAĞCI

A thesis submitted to

the Graduate Institute of Sciences and Engineering

of

Fatih University

in partial fulfillment of the requirements for the degree of

Master of Science

in

Computer Engineering

August 2006
Istanbul, Turkey

I certify that this thesis satisfies all the requirements as a thesis for the degree of Master of Science.

____
Prof. Dr. Kemal FİDANBOYLU
Head of Department

This is to certify that I have read this thesis and that in my opinion it is fully adequate, in scope and quality, as a thesis for the degree of Master of Science.

____
Assist. Prof. Dr. Atakan KURT
Supervisor

Examining Committee Members

Assist. Prof. Dr. Atakan KURT ______________________

Assist. Prof. Dr. Veli HAKKOYMAZ ______________________

Assist. Prof. Dr. Omar ALAGHA ______________________

It is approved that this thesis has been written in compliance with the formatting rules laid down by the Graduate Institute of Sciences and Engineering.

____
Assist. Prof. Dr. Nurullah ARSLAN
Director

Date
August 2006

AIR POLLUTION FORECASTING BY USING DATA MINING

A. Betül GÜLBAĞCI

M. S. Thesis - Computer Engineering
August 2006

Supervisor: Assist. Prof. Atakan KURT

ABSTRACT

Air pollution is one of the most important environmental problems in metropolitan cities like Istanbul; it affects human health and may even cause deaths. It is necessary to develop warning systems for air pollution by forecasting air pollution indicators. Artificial Neural Networks (ANN) which is a data mining technique has been widely used in extracting knowledge and data from very large databases. In this study, a system for forecasting air pollution indicators of SO_2, PM_{10} and CO with ANN is developed for Istanbul. The relationship between local meteorological data -like temperature, humidity, pressure, wind direction, wind speed - and air pollution indicators is used to construct the ANN. Air pollution forecasts for three days into future are done with this model and furthermore these forecasts are published in the air pollution forecasting website. In order to get lower error rates in forecasts, the ANN model is enhanced and different experiments are performed on the dataset and the results are presented.

Keywords: Air pollution forecasting, data mining, artificial neural networks

VERİ MADENCİLİĞİ ile HAVA KİRLİLİĞİ TAHMİNİ

A. Betül GÜLBAĞCI

Yüksek Lisan Tezi – Bilgisayar Mühendisliği
Ağustos 2006

Tez Yöneticisi: Yrd. Doç. Dr. Atakan KURT

ÖZ

Hava kirliliği, İstanbul gibi çok büyük şehirlerdeki en önemli çevre problemlerinden biridir; insan sağlığını etkiler ve hatta ölümlere bile neden olabilir. Hava kirliliği göstergelerini tahmin ederek, hava kirliliği uyarı sistemleri geliştirmek gerekmektedir. Bir veri madenciliği tekniği olan yapay sinir ağları, çok büyük veritabanlarından değer taşıyan bilgiyi çıkarmak için çok kullanılmaktadır. Bu çalışmada, yapay sinir ağları ile SO_2, PM_{10} ve CO hava kirliliği göstergelerini tahmin eden bir sistem İstanbul için geliştirilmiştir. Sıcaklık, nem, basınç, rüzgâr yönü, rüzgâr hızı gibi meteorolojik veriler ile hava kirliliği göstergeleri arasındaki ilişki yapay sinir ağı oluşturmak için kullanılmıştır. Gelecek üç gün için hava tahmini bu modelle yapılmış ve bunun da ötesinde tahminler hava kirliliği tahmini sitesinde yayınlanmıştır. Hava kirliliği tahminlerinde daha düşük hata oranları elde etmek için yapay sinir ağı modeli iyileştirilerek geliştirilmiş, çeşitli deneyler veri kümesi üzerinde yapılmış ve sonuçları sunulmuştur.

Anahtar Kelimeler: Hava kirliliği tahmini, veri madenciliği, yapay sinir ağları

ACKNOWLEDGEMENT

I express sincere appreciation to Assist. Prof. Atakan KURT for his guidance and insight throughout the research.

I also express gratitude to Assist. Prof. Ferhat Karaca, Prof. Alexander Nikov, Assist. Prof. Omar Alagha and Hüseyin Hakkoymaz who have previous works and creative ideas on this thesis subject.

Thanks go to the committee member Assist. Prof. Veli Hakkoymaz for his valuable suggestions and comments.

The assistance of Ahmet İzgi who has worked on the web page design is gratefully acknowledged.

I express my thanks and appreciation to my family for their understanding, motivation and patience.

TABLE OF CONTENTS

ABSTRACT IV

ÖZ V

ACKNOWLEDGEMENT VI

TABLE OF CONTENTS VII

LIST OF TABLES IX

LIST OF FIGURES XI

CHAPTER 1 INTRODUCTION 1

CHAPTER 2 BACKGROUND 4

2.1 Air Quality Forecasting Systems 4

2.2 Neural Networks Used In Ecological Modeling 7

2.3 NN-Airpol Tool 9

CHAPTER 3 THE AIR POLLUTION FORECASTING SYSTEM 11

3.1 Data Collection Module 12

3.1.1 Air Pollution Indicators 12

3.1.2 Meteorological Air Condition 15

3.2 Database Module 18

3.3 Forecasting Module 19

3.4 Web Site Module 21

CHAPTER 4 EXPERIMENTS, RESULTS AND EVALUATION 24

4.1 Environment 24

4.2 Air Pollution Dataset 25

4.3 Experiments 28
4.3.1 Finding Optimum Training Dataset Size 28
4.3.2 Arithmetic Mean 31
4.3.3 Three Days Data in One Record 32
4.3.4 Using the Day of Week Parameter in Neural Network 34
4.3.5 Location based Data Mining 36
4.3.5.1 Forecasting One District with the Nearest District's Values 38
4.3.5.2 Forecasting a District with Nearest Two Districts' Values 39
4.3.5.3 Forecasting Air Pollution with Three Districts' Values Propagated with Distance ……………………………………………………………………...40
4.3.5.4 Forecasting Air Pollution with Three Districts Values Using Angles 41
4.4 Evaluation 43
CHAPTER 5 CONCLUSIONS 47
REFERENCES 49

LIST OF TABLES

Table 3.1 Coordinates of air pollution measurement stations in Istanbul 13
Table 3.2 Meteorological air condition data 17
Table 4.1 Test data format 25
Table 4.2 Training data format 26
Table 4.3 General statistical values of parameters in data set 27
Table 4.4 Average error percentages of arithmetic mean experiment 31
Table 4.5 Average error percentages of using the last true value as prediction for tomorrow experiment 32
Table 4.6 Three days data in one row experiment error percentages 33
Table 4.7 Average of each day's pollutant 34
Table 4.8 The day of week 1-7 is used as input parameter in neural network recent 3 days in training data 35
Table 4.9 The day of week 1-7 is not used as input parameter in neural network recent 3 days in training data 35
Table 4.10 Correlations of 10 districts in Istanbul 37
Table 4.11 Forecasting a district with the nearest district's values 38
Table 4.12 Forecasting by not using nearest district's values 39
Table 4.13 Forecasting a district with nearest two districts' values 39
Table 4.14 Distances to Besiktas 40
Table 4.15 Error percentages of forecasting air pollution with three districts values propagated with distance 41
Table 4.16 Distances, sinus, cosines values of three districts...42
Table 4.17 Error percentages of forecasting air pollution with three districts values using angles 42

Table 4.18 Experiment names and numbers 43

LIST OF FIGURES

Figure 2.1 UK air quality forecast 5

Figure 2.2 US air quality web site 6

Figure 2.3 Australian air quality forecasting system 7

Figure 2.4 NN-Airpol tool website 10

Figure 3.1 System Architecture 11

Figure 3.2 Air pollution levels in Istanbul from www.ibb.gov.tr 15

Figure 3.3 Air pollution forecasting neural network 20

Figure 3.4 Air pollution forecasting system web site 22

Figure 4.1 Pseudo code of finding optimum training dataset size 29

Figure 4.2 SO_2 error percentages for past days 3 to 15 30

Figure 4.3 CO error percentages for past days 3 to 15 30

Figure 4.4 PM_{10} error percentages for past days 3 to 15 31

Figure 4.5 Pseudo code of three days data in one row experiment 33

Figure 4.6 Ten stations in Istanbul 36

Figure 4.7 Four districts on the x-y plane 42

Figure 4.8 Average error percentages of SO_2 44

Figure 4.9 Average error percentages of PM_{10} 44

Figure 4.10 Average error percentages of CO 45

LIST OF SYMBOLS AND ABBREVIATIONS

SYMBOL/ABBREVIATION

$\mu g/m_3$	Micrograms per cubic meter
ANN	Artificial neural networks
CH_4	Methane
CO	Carbon monoxide
GNU	General public license
NMHC	Non-methane hydrocarbon
NO	Nitric oxide
NO_2	Nitrogen dioxide
NOx	Nitrogen oxides
O_3	Ozone
PHP	Personal home page
PM_{10}	Particulate matter smaller than 10 μm
SO_2	Sulfur dioxide
THC	Total hydrocarbon
UTM	Universal transverse mercator
IBB	Istanbul metropolitan municipality
UV	Ultra violet
DEWP	Dew point
Avg	Average

CHAPTER 1

INTRODUCTION

Air pollution is one of the most important environmental problems especially in urban cities like Istanbul (Akkoyunlu and Ertürk, 2003). Air is essential for human beings to continue living. The air humans breath is made up of mixture of gases and small particles. Pollutants in the air are chemicals or substances that are harmful to humans, other species and ecosystems. These pollutants can come from human sources or from natural sources such as volcanoes or dust storms. They are responsible for large numbers of deaths and cases of respiratory disease.

There are many air pollution indicators that affect human health. Some of the most dangerous indicators are briefly explained below.

Particulate matter (PM)*:* Particulate matter is composed of very small solid or liquid particles suspended in the air. They vary in the chemical makeup and the size of particles. $PM_{2.5}$ refers to particles less than or equal to 2.5 μm in aerodynamic diameter. PM_{10} are those that are between 2.5 and 10 μm in diameter. Both of these particles can accumulate in the respiratory system and are associated with numerous health effects on humans.

PM is the most noticeable pollutant because it dramatically reduces visibility in urban areas. Diesel smoke is a good example of PM. Several epidemiological studies have indicated a strong link between increased PM_{10} concentrations and increased mortality and morbidity from all causes (Tittanen et al., 1999; Schwartz, 1996; McDonnell et al., 2000; Dockery et al., 1993)

Carbon monoxide (CO)*:* Carbon Monoxide is a major urban air pollutant. CO is a colorless, odorless gas. It forms during the incomplete combustion of fuels that contain carbon. Vehicle exhaust makes up more than 60% of all CO emissions, and is one of the most dominant pollutants in cities. CO can also come from forest fires, and its concentrations are the highest during cold weather.

Sulfur dioxide (SO_2)*:* Sulfur dioxide pollution is produced when sulfur containing fuels are burned. High concentrations of SO_2 can aggravate respiratory problems, such as asthma, bronchitis, and emphysema. In high quantities, SO_2 can harm plants and cause rain to become acidic (Saral and Erturk, 2000).

Nitrogen oxides: Nitrogen oxides are formed during high temperature combustion processes from the oxidation of nitrogen in the air or fuel. The principal source of nitrogen oxides - nitric oxide (NO) and nitrogen dioxide (NO_2), collectively known as NOx - are emissions from vehicles and from power plants and other fossil fuel-burning industries. NO_2 levels vary with traffic density.

Ozone (O_3)*:* Ozone is a gas that can form and react under the action of light. It forms a layer that shields the earth from ultraviolet rays high up in the atmosphere. However, at ground level, ozone is considered a major air pollutant. It mainly affects the lungs, but it can also affect the eyes.

Hydrocarbons: Hydrocarbon emissions data from mobile sources is measured as total hydrocarbon (THC). Methane (CH_4) is an organic gas that is orders of magnitude less reactive than other hydrocarbons. Non-Methane Hydrocarbon (NMHC) is the sum of all hydrocarbon air pollutants except methane; significant precursors to ozone formation. CH_4 and NMHC come together and form THC.

Spatial and temporal variations in emissions of air pollutants and the accompanying variability in meteorological conditions can lead to occurrences of pollutant levels, which can cause adverse short-term and chronic human health impacts (Künzli, 2005; Elbir ,2000). Urban air quality management and information systems are required to predict following day's air pollution levels and provide necessary actions and controlling strategies (Monterio et al, 2005). When ambient air concentration levels exceed air quality guideline, some air quality and warning systems must be developed.

Air-quality models play a significant role in all aspects of air pollution control and planning, where prediction is a major component (Longhurst et al., 1996). The artificial neural network (ANN) which is a data mining technique makes no prior assumptions concerning the data distribution. ANNs are capable of modeling highly non-linear relationships and can be trained to accurately generalize when presented with a new data set (Gardner and Dorling,1998). ANNs are parallel computational models, comprised of densely interconnected adaptive processing units. The important characteristic of neural networks is their adaptive nature, where 'learning by example replaces programming' (Bose and Liang, 1998). This feature makes the ANN techniques very appealing in application domains for solving highly non-linear phenomena (Zurada, 1997).

In this study, the objective is developing an air pollution forecasting system by neural networks and present results in the project web site daily. This study makes predicts for ten districts in Istanbul and publishes air pollutant indicators' (SO_2, CO, PM_{10}) levels for three days into future in the web site. So that people are informed about three days forecasts by the web site and they are able to take precautions for their health.

This thesis is organized as follows: Chapter 2 gives the background about air pollution modeling, some air quality warning systems and the NN-Airpol tool. Chapter 3 discusses the meteorological and air pollution data collection and storage, data mining models and the project website. Chapter 4 discusses the experiments conducted to develop best data mining model in air pollution forecasts. Chapter 5 discusses future work and concludes the study.

CHAPTER 2

BACKGROUND

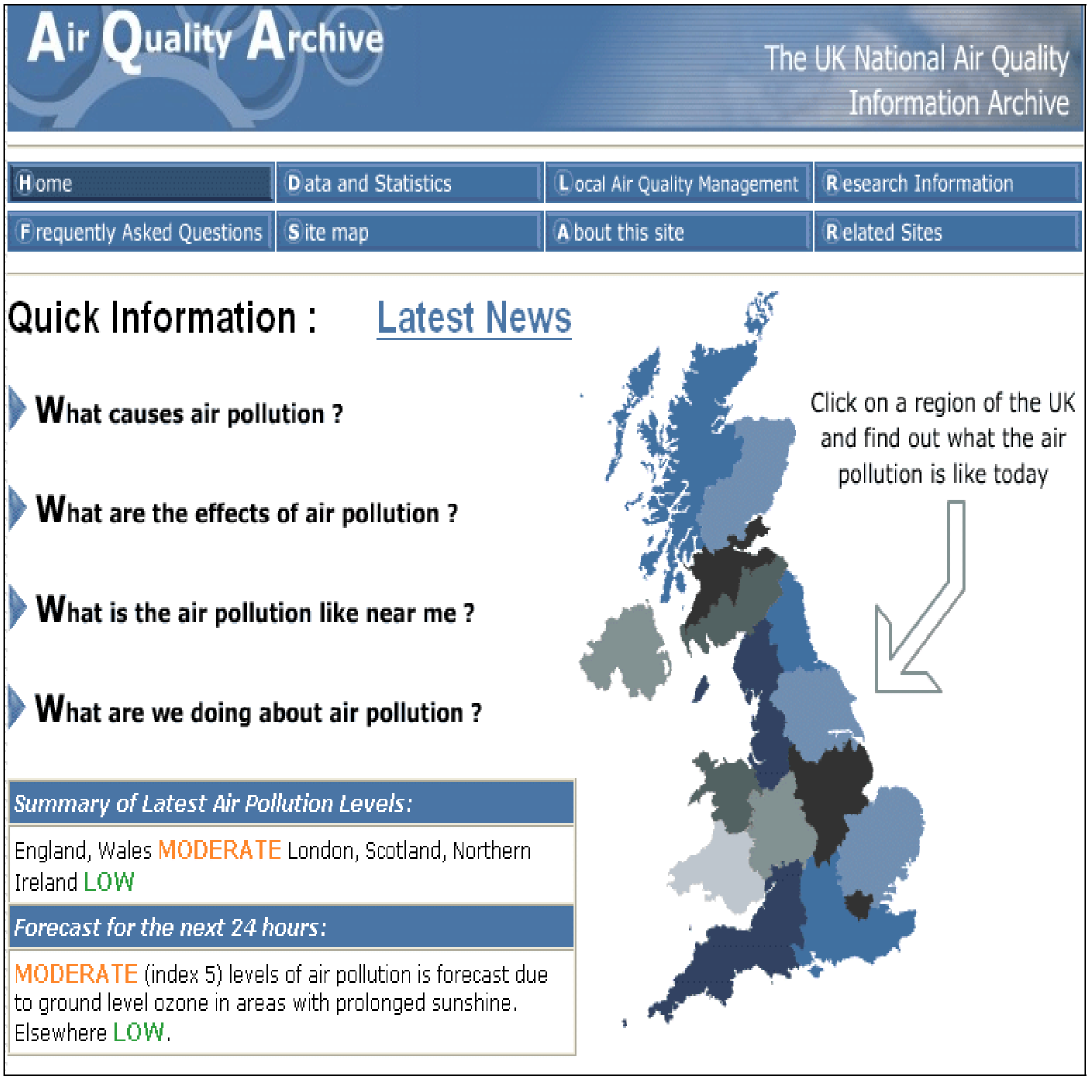
This chapter includes general background information and preceding works about air pollution modeling with ANN. This chapter is divided into several sections. In section 2.1, the air quality warning systems currently on internet are presented. In section 2.2, the use of neural networks in ecological modeling is introduced by specific examples. The NN-Airpol project which is designed previously to predict air pollution and forms the base of this thesis is discussed in section 2.3.

2.1 AIR QUALITY FORECASTING SYSTEMS

Air quality warning systems are being developed to forecast air pollution indicators' levels for days ahead. Unfortunately, there are a few limited examples in giving information about the air pollutant's levels around the world. These air quality forecasting systems are the UK air quality information archive, a cross-agency US government web site and Australian Air Quality forecasting system.

1. The UK air quality information archive: http://www.airquality.co.uk/

In UK air quality information archive web site, daily updated forecasts for up to 24-hours ahead of UK air pollution can be monitored. Forecasts are made by using different sources: Pollutant concentrations, weather forecasts for two days before and real time results from the ozone trajectory model. By combining these data with the model they have developed, air pollution forecasts for 24 hours is made. Figure 2.1 shows the main web page of the web site. People can learn air quality by clicking the area on the UK map.


Air Quality Archive
The UK National Air Quality Information Archive
Home
Data and Statistics
Local Air Quality Management
Research Information
Frequently Asked Questions
Site map
About this site
Related Sites
Quick Information :
Latest News
What causes air pollution ?
What are the effects of air pollution ?
What is the air pollution like near me ?
What are we doing about air pollution ?
Click on a region of the UK and find out what the air pollution is like today

Summary of Latest Air Pollution Levels:

England, Wales MODERATE London, Scotland, Northern Ireland LOW

Forecast for the next 24 hours:

MODERATE (index 5) levels of air pollution is forecast due to ground level ozone in areas with prolonged sunshine. Elsewhere LOW.

Figure 2.1 UK air quality forecast

2. A cross-agency US government web site: http://airnow.gov/

The US government web site gives air quality information for all cities in US. The levels of O_3 and PM_{10} in each day can be monitored from web site and forecasts for next 24 hours can be found in web site. There are 6 levels of air quality from good to hazardous. On the map next days air pollution levels for many places shown as the can be seen as in the Figure 2.2.

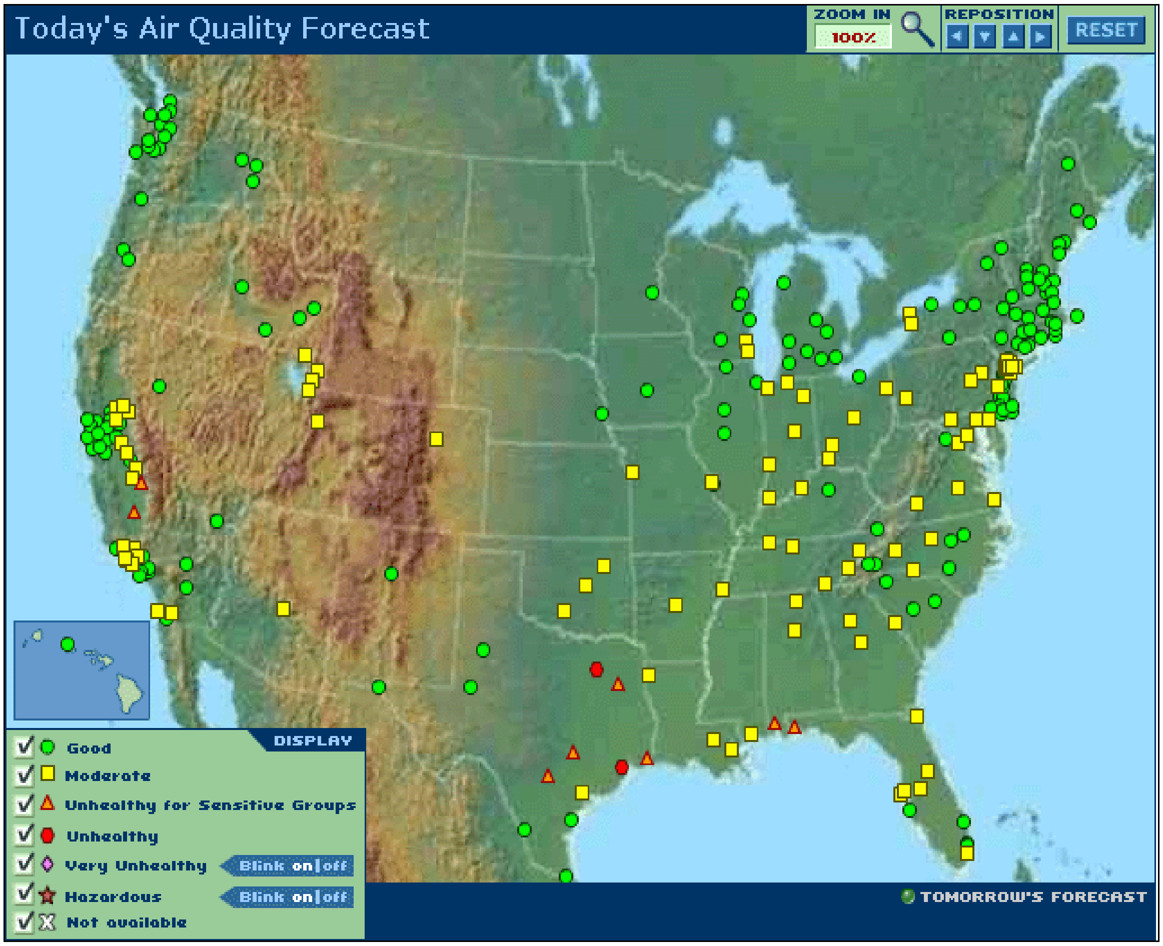


Figure 2.2 US air quality web site

3. Australian Air Quality forecasting system:
http://www.epa.vic.gov.au/air/AAQFS/default.asp/

The Australian Air Quality Forecasting is a pilot system that only forecasts 24 hour air pollution in Australia. It is a state of the art modeling system where meteorological and emissions information is entered into the model which aims to accurately forecast air pollution by using statistical forecasting techniques. This system (Figure 2.3) only gives information about the air quality with five levels.

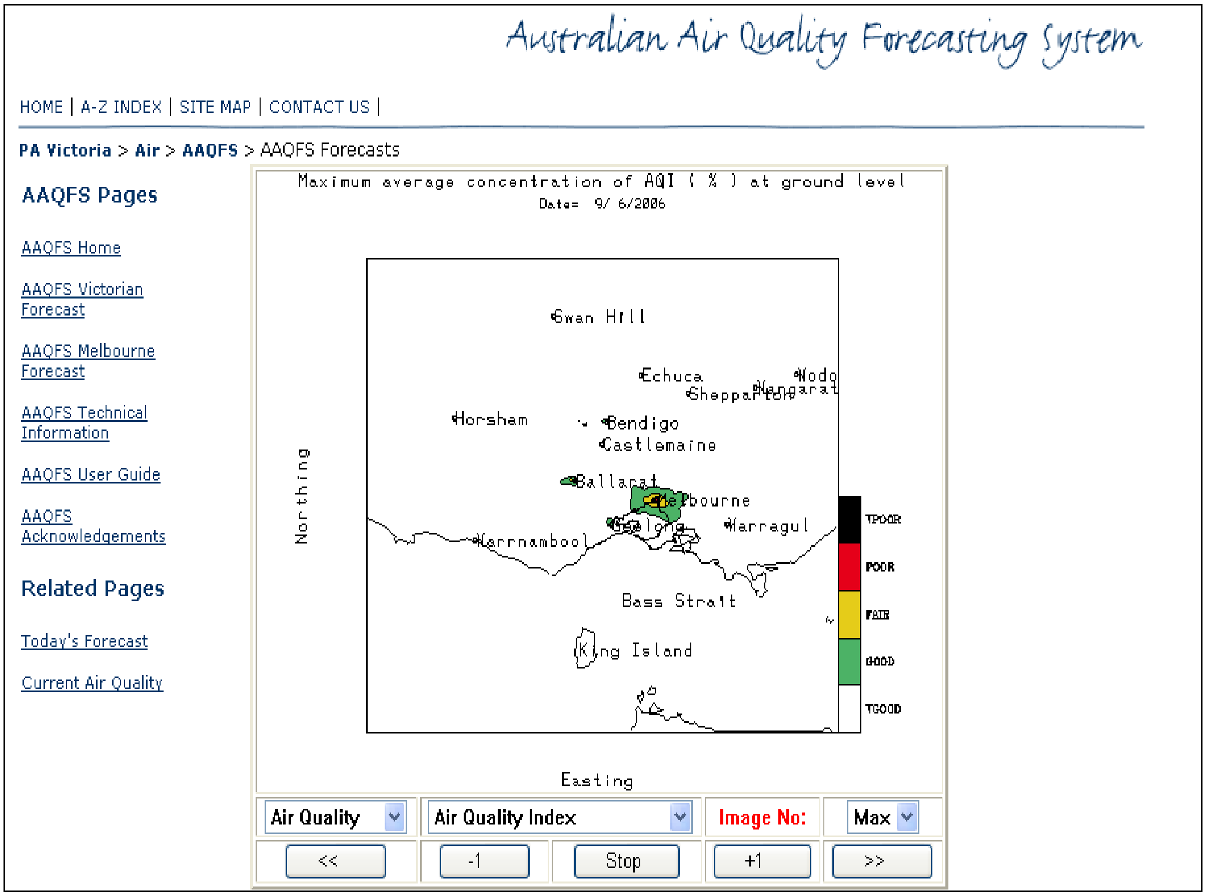


Figure 2.3 Australian air quality forecasting system

2.2 NEURAL NETWORKS USED IN ECOLOGICAL MODELING

Data mining has recently become important alternative tool to conventional methods in modeling complex non-linear relationships in many areas like ecological modeling. In the recent past, a data mining technique ANN have been applied to model large dimensionality of ecological data (Gevrey et al., 2003). Some important applications by data mining in environmental science are:

- Lek et al. (1996) compared multiple regression and ANN models in predicting density, biomass, reproduction potential, growth in brown trout management.
- Paruelo and Tomasel (1997) compared the performance of ANN models with regression models in predicting functional attributes of ecosystem and they indicated better performance of ANN models.
- Nunnari et al. (1998) used ANN technique to model the pollutants produced by alteration of photolytic cycle of NO_2, due to the presence of hydrocarbons released into the atmosphere.
- Dimopoulos et al. (1999) developed a neural network model to estimate the lead concentration in grasses using urban descriptors as model inputs in the Athens city, Greece.
- Manel et al. (1999) compared the performance of multiple discriminant analysis, logistic regression and ANNs in predicting the river bird's presence or absence from 32 variables consisting stream altitude, slope, habitat structure, chemistry and invertebrate abundance. The study indicated out-performance of the ANN model when compared with other traditional ecological modeling methods.
- Karul et al. (2000) used a three-layer leven berg-Marquardt feedforward learning algorithm to model the eutrophication process in water bodies in Turkey.
- Heymans and Baird (2000) used ANN technique to analyze the carbon flow in the northern Benguela upwelling ecosystem of Namiba.
- Antonic et al. (2001) forecasted the forest survival after building the hydroelectric power plant on the Drava river, Croatia using the ANN.
- Olden and Jackson (2002) described randomization approach for statistically assessing the importance of network connection weights and the contribution of input variables in the neural network.
- Park et al. (2003) used 'unsupervised' and 'supervised' network training algorithms to classify the sampling sites and to predict the aquatic insect species richness in running waters in France.

- Ryan et al. (2004) used ANN technique to simulate nitrous oxide (N_2O) emissions from temperate grassland in New Zealand.

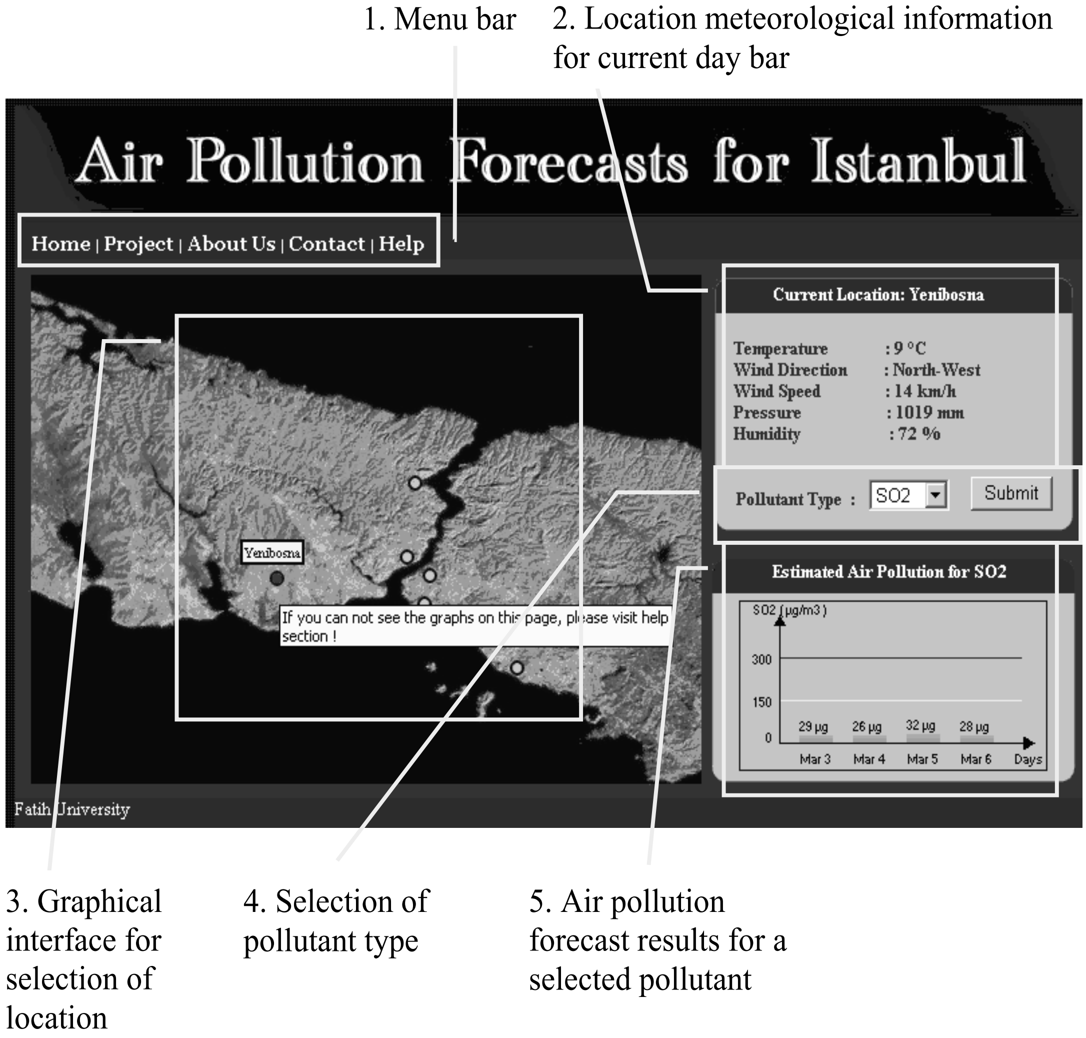
In the recent past, ANN technique has become more and more popular in modeling air-quality data (Nagendra and Khare, 2004). The multilayer neural network technique has been used to forecast the O_3 (Comrie, 1997; Gardner and Dorling, 1996, 2000), the SO_2 (Boznar et al., 1993), the NO_2 (Gardner and Dorling, 1999) and the PM (Perez and Trier, 2001) in the ambient environment.

2.3 NN-AIRPOL TOOL

The NN-AirPol tool is an ANN based method for the evaluation and control of air pollution developed by Karaca et al (2005). It forecasts the air pollution levels for three days into future. The forecasted air pollution levels are displayed in the web site.

NN-Airpol tool predicts the air pollution with best fitting backpropagation algorithm for Yenibosna district in Istanbul metropolitan city. It uses meteorological data and air pollution data. The concentrations of air pollution indicators -SO_2, PM_{10} and CO- are determined with ANN. The prognoses of these concentration indicators present the outputs of the neural network. If the forecasted air pollutant concentrations are higher than the threshold values relevant episode measures and strategic action plans are proposed. If the concentrations are dangerous for human health red ecowarnings are proposed. If the concentrations are higher than national air quality standards for certain areas, yellow ecowarnings are proposed. If the concentrations are lower, green episodes are proposed. A two-layer neural network with tan-sigmoid transfer function, a hidden layer and a linear transfer function at the output layer were used in NN-Airpol tool.

The system was forecasting the next three days air pollution indicators’ levels and showing the forecasts on the website. The web site address of that version of the system is: http://airpol.fatih.edu.tr/old/index.php and screenshot is shown in the Figure 2.4.



1. Menu bar

2. Location meteorological information for current day bar

Air Pollution Forecasts for Istanbul

Current Location: Yenibosna

Temperature : 9 °C
Wind Direction : North-West
Wind Speed : 14 km/h
Pressure : 1019 mm
Humidity : 72 %

Pollutant Type : SO2

If you can not see the graphs on this page, please visit help section !

Estimated Air Pollution for SO2

3. Graphical interface for selection of location

4. Selection of pollutant type

5. Air pollution forecast results for a selected pollutant

Figure 2.4 NN-Airpol tool website

This thesis is based on the preceding works in NN-Airpol tool and the motivation is improving NN-Airpol tool. First of all, the system developed was working for only Yenibosna district. There are other nine air pollution measurement stations in Istanbul metropolitan city. One of the objectives in this thesis is improving the NN-Airpol tool to forecast air pollution in ten districts in Istanbul. Another aim in this thesis is to improve the neural network model in NN-Airpol tool. For this reason, we proposed to do some experiments to do better forecasts with ANN.

CHAPTER 3

THE AIR POLLUTION FORECASTING SYSTEM

In the air pollution forecasting system, meteorological data and air pollution data are gathered and stored in database. The data gathered is used as inputs in ANN and forecasts are made for three days into future. The forecasts are shown in the air pollution forecasting website by graphs.

The air pollution forecasting system consists of four main modules as shown in the Figure 3.1: These modules are data collection module, database module, forecasting module and web site module. Data collection module is responsible for gathering meteorological data and air pollution data from web sites. Database module stores and manages the data collected from web sites. Forecasting module makes forecasts for three days into future. Web site module shows the forecasted values in web.

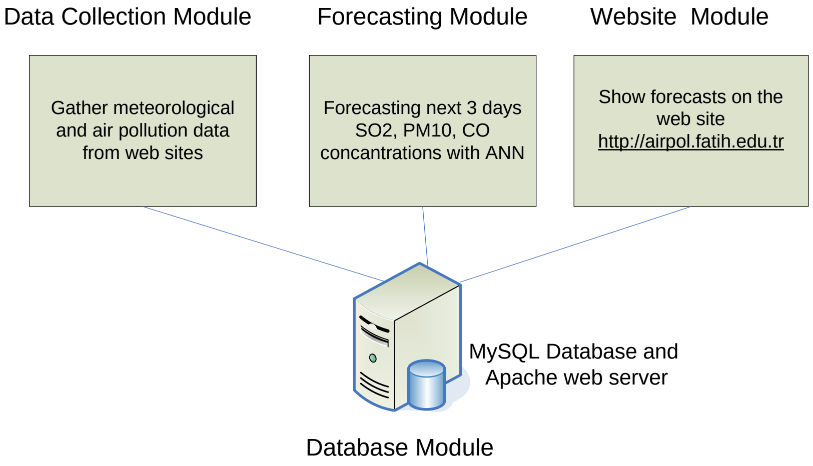

Data Collection Module
Forecasting Module
Website Module
Gather meteorological and air pollution data from web sites
Forecasting next 3 days SO2, PM10, CO concantrations with ANN
Show forecasts on the web site http://airpol.fatih.edu.tr
MySQL Database and Apache web server
Database Module

Figure 3.1 System Architecture

This chapter introduces the air pollution forecasting system architecture developed for Istanbul. This chapter is divided into several sections. In section 3.1., data collecting module which gathers meteorological and air pollution data from web sites is introduced. In section 3.2, database module is presented. In section 3.3, forecasting model is presented and in section 3.4 the web site module is discussed.

3.1 DATA COLLECTION MODULE

Air pollution forecasting system uses the relation between meteorological air conditions and air pollution for forecasting. Meteorological air condition data and air pollution data are gathered to be used in the forecasts.

The best way of collecting data is from the Internet. Data in html or xml are converted to strings and stored in the MySQL database. This module collects daily weather forecasting and air pollution data. Air pollution data and meteorological air condition data are collected from related websites and inserted into MySQL database. There is a scheduled task in the server to get the most recent data. It collects data from related web pages and put into the database on the server.

Air pollution data measured by the Istanbul metropolitan municipality (IBB) and published in its official web site http://www.ibb.gov.tr. Meteorological air condition data is gathered from http://bbc.co.uk since January 2005 and from www.weather.com since March 2006.

3.1.1 Air Pollution Indicators

The levels of air pollution indicators define the air quality so they must be monitored periodically. Especially in cities that have high population like Istanbul, these indicators must be monitored daily. Currently Istanbul municipality has been monitoring ten permanent air pollutants at ten stations in Istanbul. These stations and their coordinates are shown in the Table 3.1.

Table 3.1 Coordinates of air pollution measurement stations in Istanbul

DISTRICT	Coordinates							UTM		
	Y				X			Y		X
FATİH	28	57	17		41	0	53	412069.60		4542810.67
EYÜP	28	56	44		41	4	22	411381.33		4549273.66
BEŞİKTAŞ	29	0	36		41	3	14	416771.41		4547124.60
ESENLER	28	53	17		41	2	18	406491.67		4545490.09
SARIYER	29	2	59		41	7	44	420208.05		4555402.39
BAHÇELİEVLER	28	49	36		40	59	56	401285.93		4541199.78
ÜSKÜDAR	29	1	30		41	0	55	417989.25		4542798.67
KADIKÖY	29	2	1		40	59	31	418685.26		4540197.97
KARTAL	29	12	27		40	53	24	433208.71		4528736.27
ÜMRANİYE	29	9	44		41	0	49	429528.69		4542512.06

Pollutants monitored are SO_2, CO, PM_{10}, NO, NOx, NO_2, THC, CH_4, NMHC, O_3. The levels of air pollutant indicators in each district are published on the official IBB website whose address is http://www.ibb.gov.tr/tr-TR/HavaKalitesi/Olcum/. However there are no future predictions of air pollution indicators. Only six pollution levels from good to hazardous are shown for current day and past days with colorful circles. There are six pollution levels shown in the site:

1. Green circles → Good
2. Yellow circles → Moderate
3. Orange circles →Unhealthy for sensitive groups
4. Red →Unhealthy
5. Pink → Very unhealthy
6. Brown →Hazardous

On the Istanbul city map, Figure 3.2, the circles showing the air quality can be seen; and also air pollutant values can be monitored at selected date and at selected district.

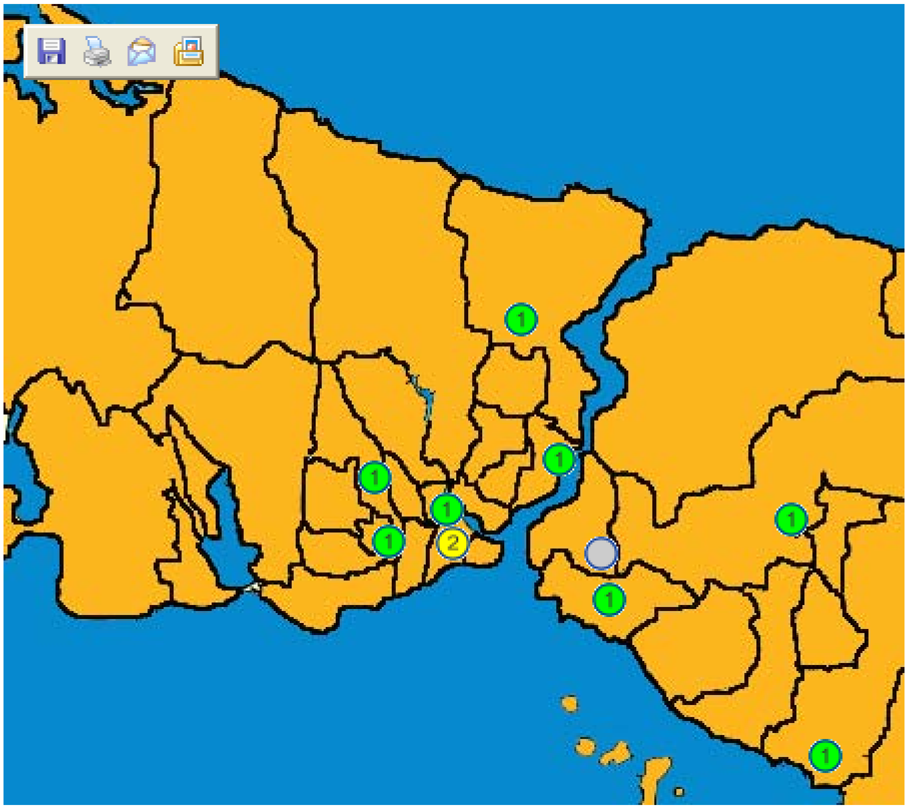


Figure 3.2 Air pollution levels in Istanbul from **www.ibb.gov.tr**

Air pollution data including SO_2, CO, PM_{10}, NO, NOx, NO_2, THC, CH_4, NMHC, and O3 is collected from IBB official site in html format. Collected data is parsed by PHP program and ten air pollutant values for ten districts in Istanbul are inserted into the database. Although there are ten air pollutants measured, air pollution forecasting system forecasts only SO_2, CO and PM_{10}. Forecasting models for other air pollutants have not been developed yet.

3.1.2 Meteorological Air Condition

Meteorological air condition data plays important role in forecasting. Air condition data collected is from www.bbc.co.uk and www.weather.com. Meteorological air condition data is measured by Turkish State Meteorological Service. The published air condition data is in www.bbc.co.uk and www.weather.com is gathered from Turkish State Meteorological Service. In every two hours data in database is updated according to the updates in these web sites. There are seven meteorological air condition values collected from www.bbc.co.uk and eight values from www.weather.com. Recently only the data collected from www.bbc.co.uk is used in forecasts .The data gathered from www.weather.com has not been used in developed systems yet; but it will be used in the next versions of air pollution system. The air condition values and their ranges are shown in the Table 3.2.

Table 3.2 Meteorological air condition data

Meteorological Data	**Dimensions/Ranges**
Air condition	1- Sunny day 2- Partly cloudy (at night) 3- Sunny intervals 4- Black, low level cloud 5- Light rain shower (night) 6- Drizzle 7- Light rain (night) 8- Light rain shower 9- Heavy rain shower (night) 10- Heavy rain shower 11- Heavy rain (day and night) 12- Cloudy with sleet 13- Sleet shower 14- Sleet shower (night) 15- Hail shower 16- Hail shower (night) 17- Cloudy with hail 18- Light snow shower (night) 19- Light snow shower 20- Cloudy with light snow 21- Heavy snow shower (night) 22- Heavy snow shower 23- Cloudy with heavy snow 24- Thundery shower (night) 25- Thundery shower 26- Sandstorm 27- Mist 28- Fog 29- Tropical Storm 30- Hazy
Temperature during day	°C
Temperature during night	°C
Wind Speed	msec^{-1}
Wind Direction	°0-°360
Humidity	%
Pressure	mmHg

Today's seven meteorological data and three days ahead predicted air condition values are collected from web sites. After being parsed, these air condition values are put into the database in to be used in the predictions.

3.2 DATABASE MODULE

It is necessary to store the air pollution and meteorological data taken from web sites and forecasts. Therefore an open source database MySQL is used in air pollution forecasting system to store and manage data.

Air pollution forecasting system has two tables in the database; one for collecting daily measurements of air pollution and meteorological air condition data, other is keeping the estimations for next three days with forecasted air condition values for ten districts in Istanbul.

1. *Table for Daily Measurements:* This table stores daily meteorological weather data and daily air pollutant data in each district in Istanbul. There are two columns as primary key called location and date to identify which district and date the real values belong. There are 15 columns to store daily measured weather conditions which are day temperature, night temperature, pressure, humidity, speed, direction, feels like temperature, direction class, pressure class, ultra violet (UV), UV class, dew point (DEWP), day temperature class, night temperature class There are 10 columns to store daily air pollution indicators levels which are $SO_{2,}$ CO, $PM_{10,}$ NO, NOx, $NO_{2,}$ THC, $CH_{4,}$ NMHC and O_3.

2. *Table for Forecasts:* This table stores the estimated three days into future values of pollutant levels for each day. There are three columns as primary key in this table called location, estimation date, predicted date. Location shows the districts in Istanbul and estimation date column shows on which

day the estimation is done and estimated date column shows for which day the forecasts are done. There are three columns for the predicted values of three air pollutants called estimated SO_2, estimated CO and estimated PM_{10}.

Information is retrieved from the web sites in every two hours to get the most recent data and inserted to the table for daily measurements. Forecasts are done with the data in the table for daily measurements. After forecasts are done, the forecasted air pollution values are inserted into the table for forecasts.

3.3 FORECASTING MODULE

Data mining, the process of employing one or more machine learning techniques to analyze and extract knowledge from data, has been used widely in many areas. The amount of data kept in databases is growing and getting more meaningful information from that data is provided by data mining. Neural network is one of the data mining techniques. ANN process information like human biological nervous systems. In this study ANN are used forecasting the air pollution levels.

There is a relationship between air condition and pollutant levels. Temperature, humidity, pressure, wind speed and wind direction have affects on the levels of air pollutants. In recent years, data mining models have been applied to atmospheric pollution modeling in general (Gardner and Dorling, 1998) and air quality problems in particular (Gardner and Dorling, 1999a; Gardner and Dorling, 1999b).

In air pollution forecasting system, mainly ANN are used to forecast next three days air pollution indicators levels. Meteorological air condition data already inserted into database is used as input to neural network. By using recent data, training is done and next days' air condition parameters are used as test data; and then estimations for air pollutants are inserted into database. In Table 3.3 the input and output parameters of ANN are shown. Figure 3.3 refers to data mining module of air pollution forecasting system.

Table 3.3 ANN parameters

Type	Parameters	Units
Input	Condition	[1-29]
Input	Day temperature	°C
Input	Night temperature	°C
Input	Humidity	%
Input	Wind Speed	Km/h
Input	Wind Direction	[0-360]
Input	Pressure	mmHg
Input	Day of week	[1-7]
Output	SO_2	mg/m^3
Output	CO	mg/m^3
Output	PM_{10}	mg/m^3

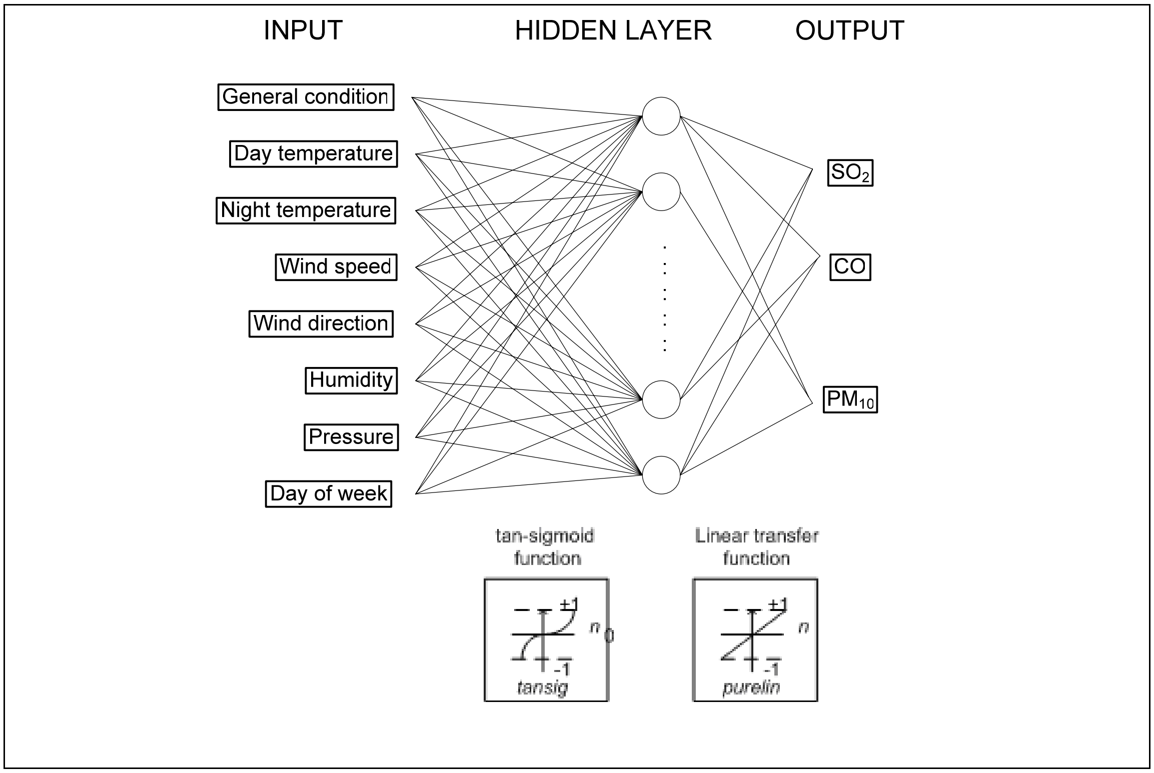
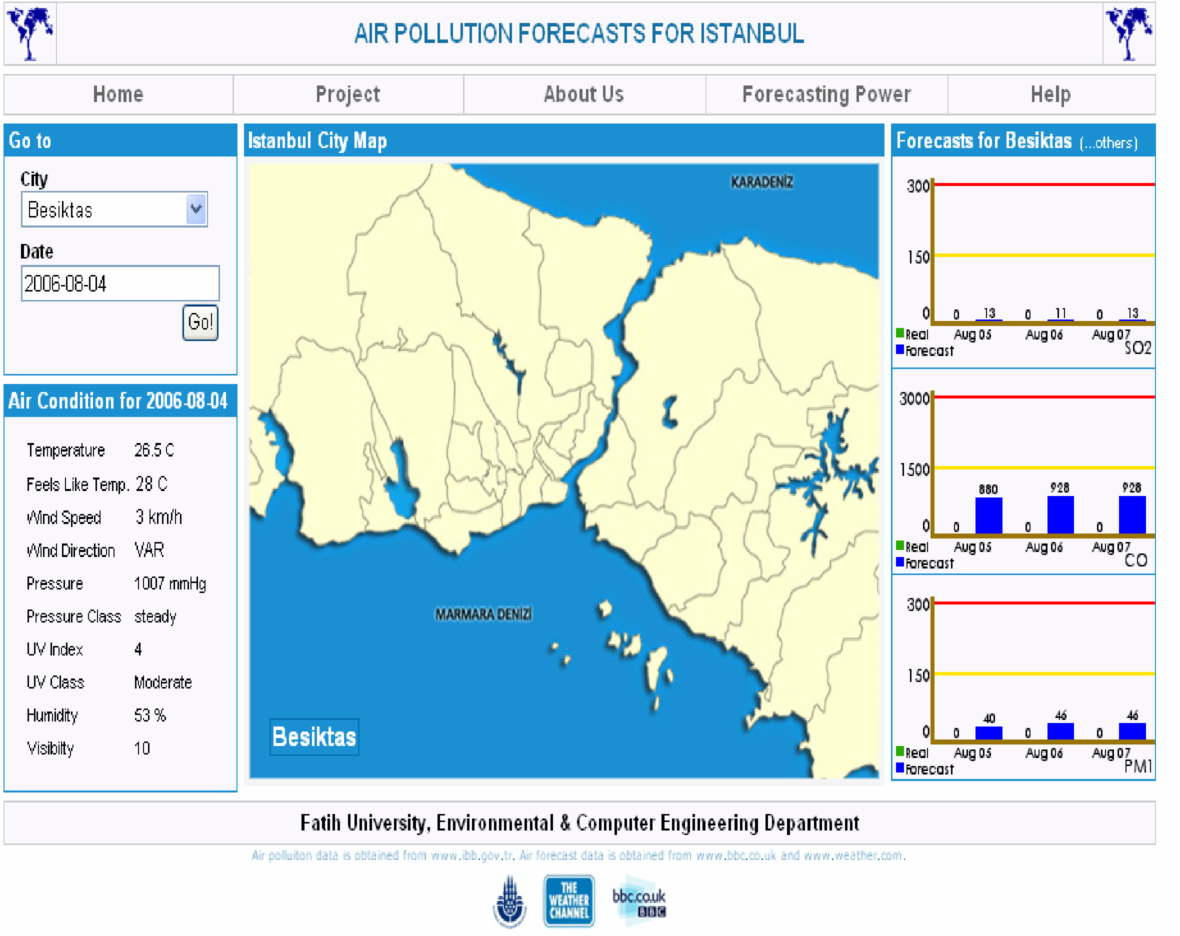


Figure 3.3 Air pollution forecasting neural network

A feed forward back propagation neural network is used in forecasting module. Transfer functions calculate a layer's output from its net input. Hyperbolic tangent sigmoid function and linear transfer function are used as transfer functions. The training function that updates weight and bias values according to Levenberg-Marquardt optimization is used as the training function in ANN. Mean square error performance function is the network performance function used in forecasting. It measures the network's performance according to the mean of squared errors.

3.4 WEB SITE MODULE

In air pollution forecasting system, forecasts for the air pollution indicators are displayed in the web site whose address is http://airpol.fatih.edu.tr. Apache web server is used as web server. There is an Istanbul map and graphs that show the air pollutant levels for the selected district of Istanbul at selected date. A screenshot of the web site is shown in Figure 3.4.



AIR POLLUTION FORECASTS FOR ISTANBUL

City: Besiktas

Date: 2006-08-04

Air Condition for 2006-08-04

Temperature	26.5 C
Feels Like Temp.	28 C
Wind Speed	3 km/h
Wind Direction	VAR
Pressure	1007 mmHg
Pressure Class	steady
UV Index	4
UV Class	Moderate
Humidity	53 %
Visibilty	10

Fatih University, Environmental & Computer Engineering Department

Air polluiton data is obtained from www.ibb.gov.tr. Air forecast data is obtained from www.bbc.co.uk and www.weather.com.

Figure 3.4 Air pollution forecasting system web site

By clicking on any district in the Istanbul city on the map, forecasts for next three days will be shown on the graphs on the right of the page. There are red and yellow lines on the graphs. The yellow lines show the warning level and red lines show the danger level for human health.

Also today’s meteorological air condition values including temperature, wind speed, wind direction, pressure, pressure class, UV index, UV class, humidity and visibility are displayed on the left side of the page. It is possible to get the past forecasts for any district by writing a past date in the date box on left top of the page. For the past forecasts, there are two columns for the estimated date on the graphs. Navy blue columns show the forecasted values and green columns show the actual air pollutant concentrations for the estimated date.

There are links to project, about us, forecasting power and help pages from the home page. On the project page, general information about the air pollution forecasting system is presented. On the about us page, people included in air pollution system project are introduced. On the forecasting power page, the average error percentages of last 15 forecasts are published. There are help topics on the help page.

CHAPTER 4

EXPERIMENTS, RESULTS AND EVALUATION

It is necessary to find the accuracy of estimates in data mining. In this study, after the air pollution forecasting system is developed, the accuracy of forecasts is calculated and in order to have higher accuracy in estimates some experiments are done.

After the first version of air pollution forecasting system is improved, ANN model is designed differently in each experiment to get lower error percentages in forecasts. Experiments are made by using the actual data that contains meteorological air condition levels and air pollution concentrations.

This chapter introduces the experiments, the results and evaluation of the experiments run over the air pollution forecasting system. At first, in section 4.1 the environment of the experiments is explained. In section 4.2, air pollution data set is introduced. In section 4.3, the experiments run over are presented. In section 4.4 the evaluation of the experiments is presented.

4.1 ENVIRONMENT

The model of ANN is designed in each experiment according to the input parameters. ANN are implemented with the neural network toolbox in Matlab. Matlab, the language of technical computing, integrates computation, visualization, and programming in an environment where problems and solutions are expressed in familiar mathematical notation (Mathworks, 2006).

4.2 AIR POLLUTION DATASET

Meteorological and air pollution data which are collected daily and inserted into the database from related web sites are used in the experiments as test and training data. There are seven meteorological data including air condition, day temperature, night temperature, pressure, wind direction, wind speed, humidity and three air pollutants including SO_2, PM_{10}, CO data in experiments. Format of test data that includes the meteorological air condition data used in ANN is shown in Table 4.1.

Table 4.1 Test data format

Predicted Day's	General Condition	Day temperature	Night temperature	Wind Speed	Wind direction	Pressure	Humidity

Table 4.2 shows the training data format of ANN in air pollution forecasting system. It consists of seven meteorological data and air pollution data for each day. In ANN, training is done with the training data and weight values are calculated and estimations are done according to the test data.

While estimating next days air pollutant indicator's values for next day, predicted meteorological air condition values for next three days are selected as test data from the table where estimations are inserted. Real meteorological and air pollutant data are selected from table where real values are inserted and used as training data in the forecasts.

Table 4.2 Training data format

	INPUT PARAMETERS							OUTPUT PARAMETERS		
Today	General Condition	Day temperature	Night temperature	Wind Speed	Wind direction	Pressure	Humidity	SO_2	PM_{10}	CO
Yesterday	General Condition	Day temperature	Night temperature	Wind Speed	Wind direction	Pressure	Humidity	SO_2	PM_{10}	CO
The day before yesterday	General Condition	Day temperature	Night temperature	Wind Speed	Wind direction	Pressure	Humidity	SO_2	PM_{10}	CO
	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .	. . .
Days before yesterday	General Condition	Day temperature	Night temperature	Wind Speed	Wind direction	Pressure	Humidity	SO_2	PM_{10}	CO

The experiments are run over for one year period from 01.08.2005 to 31.07.2006. The statistical values containing minimum, maximum, average, median, standard deviation, mode, number of null values and percentage of missing values for input and output parameters in ANN for Besiktas from 01.08.2005 to 31.07.2006 are shown in Table 4.3. Besiktas district is used in experiments because it has lower missing data percentage than other districts.

Table 4.3 General statistical values of parameters in data set

	Condition	Day Temperature	Night Temperature	Wind Direction	Wind Speed	Pressure	Humidity	SO2	PM10	CO
Minimum	1	1	-9	0	1	997	26	0	9	311
Maximum	29	37	24	360	47	1033	100	63	206	2846
Average	4	18	11	51	12	1017	66	18	53	960
Median	1	17	11	0	10	1016	66	16	53	883
Standard deviation	5	8	7	100	8	6	15	12	37	451
Mode	1	14	22	0	8	1016	61	16	0	0
Number of null values	79	79	79	79	79	79	79	30	37	31
% of missing values	22%	22%	22%	22%	22%	22%	22%	8%	10%	8%

There are missing meteorological and air pollution data in the dataset. The missing meteorological data is replaced by the mean of nearest two days method. The missing data of each day is calculated as the mean of the day before data and the day after data.

After experiments are done and error percentage is calculated by absolute percent error formula defined below:.

Given some value *a* and an approximation *b* of *a*, the **absolute error** is

$$\epsilon = |a - b| \tag{4.2.1}$$

The **percent error** is

$$\delta = \frac{|a - b|}{|a|} \times 100\% \tag{4.2.2}$$

4.3 EXPERIMENTS

Some experiments are done on the dataset to find the higher accuracy in forecasts. The experiments and their results are explained briefly from section 4.3.1 through .section 4.3.5.

4.3.1 Finding Optimum Training Dataset Size

This experiment is performed to figure out how the forecasts are affected by changing the number of recent days used in training data. The objective of this experiment is finding the number of optimum training data set size for forecasts.

In order to estimate air pollutant concentrations, predicted meteorological air conditions for next three days are used as test data. For training data, real air condition values collected until current estimation date are used. By changing the number of days in the training data from 3 to 15 training data set size is changed. For each training data set forecasts are done. Pseudo code of this experiment is shown in Figure 4.1.

```
for each pastday from 3 to 15
        select test data: predicted air conditions for next day
        select training data: select recent past days real data until today
        train
        insert new estimations into database

        select test data: predicted air conditions for the day after next day
        select training data: select recent past days real data until today and  next day's predicted data
        train
        insert new estimations into database

        select test data: predicted air conditions for two days after next day
        select training data: select recent past days real data until today and next two days' predicted data
        train
        insert new estimations into database
end
```

Figure 4.1 Pseudo code of finding optimum training dataset size

This experiment is run over for from 01 August 2005 to 31 July 2006 for Besiktas district. Eight neurons are used in this experiment. Average error percentages for each training data set size are shown in the Figure 4.2 through Figure 4.4. When recent three days used in training data, lower error percentages in forecasts are determined.

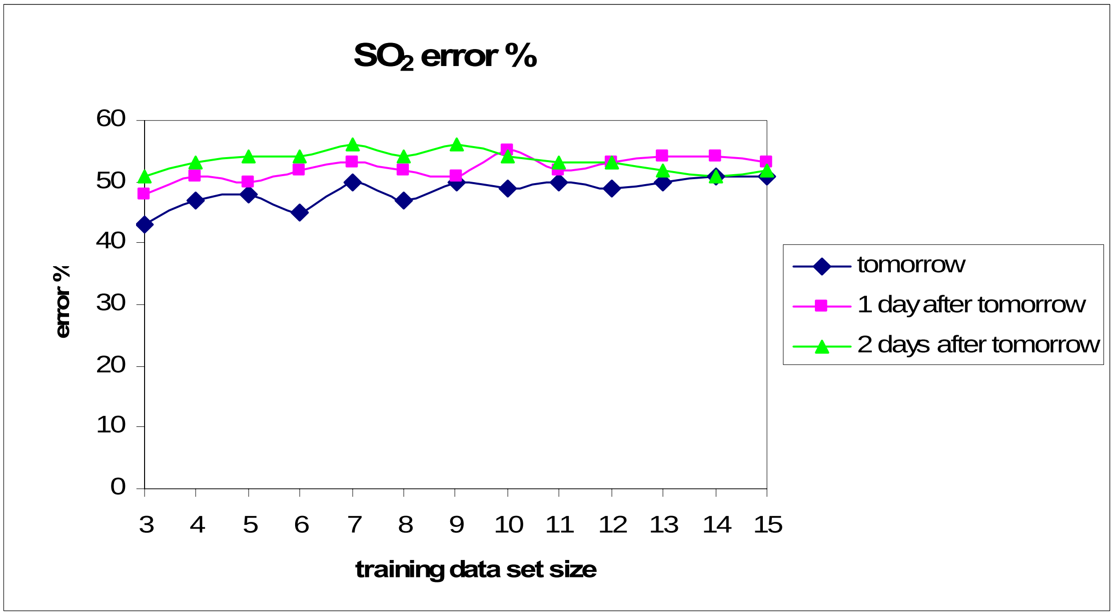


Figure 4.2 SO_2 error percentages for past days 3 to 15

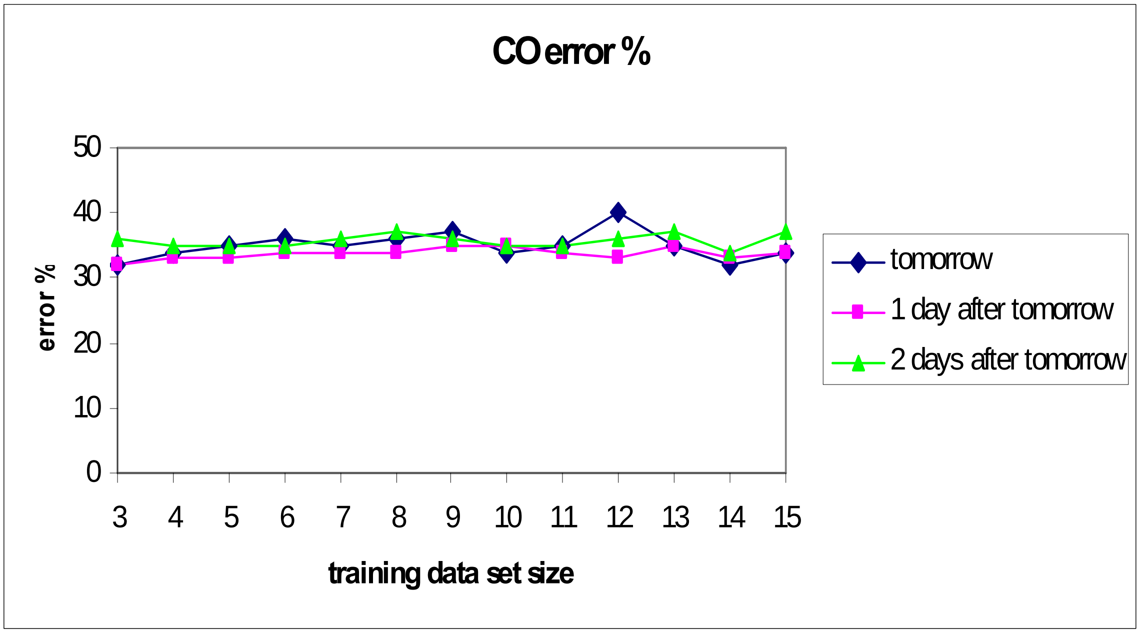


Figure 4.3 CO error percentages for past days 3 to 15

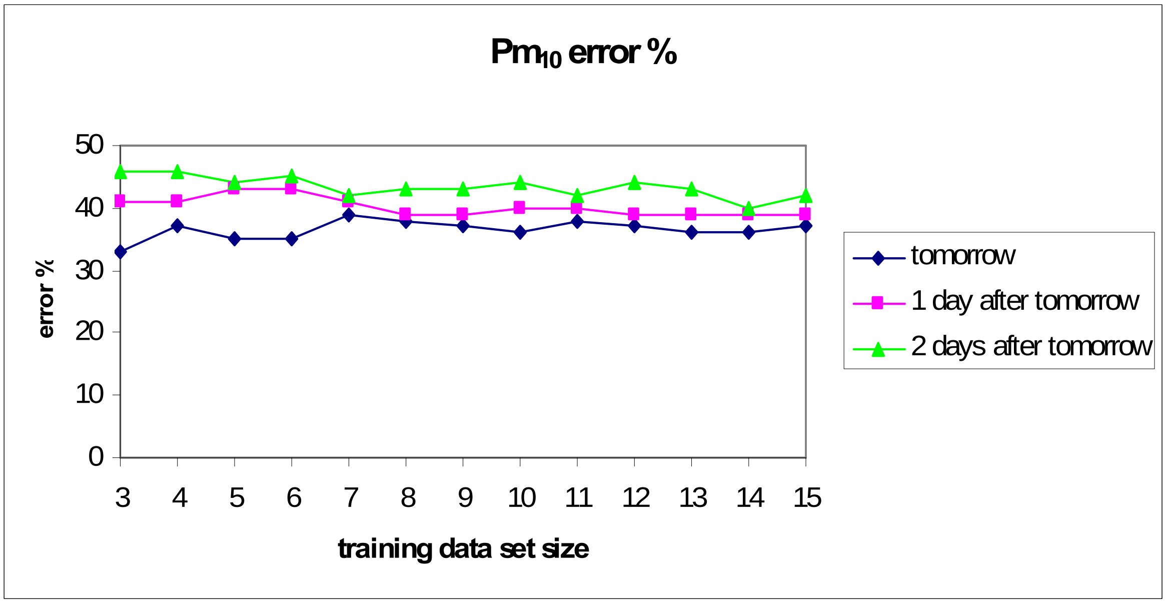


Figure 4.4 PM_{10} error percentages for past days 3 to 15

4.3.2 Arithmetic Mean

According to the finding the optimum training data set size experiment, error percentages were lower when last three days data is used as training data in ANN. Since using only three days data for ANN is surprising, a validation by taking only arithmetic mean of past three days as forecast is done in order to understand whether neural network is really essential for forecasting. Arithmetic mean of the recent 3 days for each air pollutant is calculated for Besiktas district as forecasted air pollutant concentration for next day. One year data is used to forecast next days air pollutant levels. The error percentages are shown in Table 4.4. It is validated that error percentages calculated by the arithmetic mean of recent three days are higher than ANN.

Table 4.4 Average error percentages of arithmetic mean experiment

Avg SO_2 error %	**Avg PM_{10} error %**	**Avg CO error %**
62	44	38

Another validation technique is using the last actual air pollution value as forecasted value for next day. Using last true value as the prediction for tomorrow is essentially the most primitive prediction. Today’s data is used as the forecast for tomorrow in this experiment for the same dataset. The error percentages are shown in the Table 4.5. The forecasts by using the last true value as the prediction for tomorrow are better than arithmetic mean but worse than ANN. By the way, it is validated that ANN makes better forecasts than other techniques.

Table 4.5 Average error percentages of using the last true value as prediction for tomorrow experiment

Avg SO_2 error %	**Avg PM_{10} error %**	**Avg CO error %**
58	43	40

4.3.3 Three Days Data in One Record

In time series analysis, when making estimations recent information play more important role than past information. By combining one day before and two days before meteorological air condition values and air pollutant values, one row contains three days’ data. By the way, recent days’ information is taken care of more. The objective in this experiment is using the most recent data many times in forecasts.

In this experiment we selected three recent days’ data in one row and predicted air pollutant indicator levels. Pseudo code of this experiment is shown in Figure 4.5.

```
select  test data: next days Predicted data + yesterday data + two days before data
select train data: all data until today in three parts format
train
insert into database

select  test data: next+1 days Predicted data + next Predicted data + today data
select train data: all data until today+1 in three parts format
train
insert into database

select  test data: next+2 days Predicted data + next+1 days data + next days data
select train data: all data until today+2 in three parts format
train
insert into database
```

Figure 4.5 Pseudo code of three days data in one row experiment

This experiment is run over by using one year data from 1.08.2005 to 31.07.2006 for Besiktas district. 30 neurons are used in ANN. Error percentages of this experiment can be shown in Table 4.6.

Table 4.6 Three days data in one row experiment error percentages

	Avg SO$_2$ error %	**Avg PM$_{10}$ error %**	**Avg CO error %**
Tomorrow	54	36	27
The day after tomorrow	68	61	53
Two days after tomorrow	79	64	52

4.3.4 Using the Day of Week Parameter in Neural Network

Each day of week, air pollution changes according to some parameters like traffic density, closed factories on weekends etc. The average air pollution concentration on each day of week shows the air pollution density in each day. So we get the average air pollution concentrations for each day of week in Besiktas from January 2005 to December 2005 as shown in Table 4.7.

Table 4.7 Average of each day's pollutant

	Avg SO_2	Avg PM_{10}	Avg CO
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

The values in Table 4.7 show that, air pollution is highest on Thursday and lowest on Sunday. The objective in this experiment is considering the day of week while forecasting air pollution. Numbers from 1 to 7 is assigned to each day according to the pollution level from Sunday to Thursday. By using these numbers as another input parameter in ANN, experiment run over the data set from August 2005 to July 2006. In the experiment, the last 3 days' values used in training data. 9 neurons used ANN for Besiktas district. Table 4.8 shows the average error percentages when day of week parameter is used as an input in neural network. Table 4.9 shows the average error percentages when day of week parameter is not used as an input in neural network for Besiktas district.

Table 4.8 The day of week 1-7 is used as input parameter in neural network recent 3 days in training data

	Avg SO_2 error %	**Avg PM_{10} error %**	**Avg CO error %**
Tomorrow	43	35	28
The day after tomorrow	45	39	32
Two days after tomorrow	48	40	36

Table 4.9 The day of week 1-7 is not used as input parameter in neural network recent 3 days in training data

	Avg SO_2 error %	**Avg PM_{10} error %**	**Avg CO error %**
Tomorrow	43	33	32
The day after tomorrow	48	41	32
Two days after tomorrow	51	46	36

Average error percentages show that using day of week parameter gives lower error percentages for SO_2, PM_{10} and CO. Day of week parameter can be used as input to forecast air pollution indicators.

4.3.5 Location based Data Mining

There are ten different air pollution indicators measurement stations in different ten districts in Istanbul. The coordinates and map of these districts are shown in the Figure 4.6. Each district's air pollution has high or low correlations with other districts and the pollution in each district affects other districts as shown in Table 4.10. By, determining the correlations and distance between districts, experiments are run over the dataset.

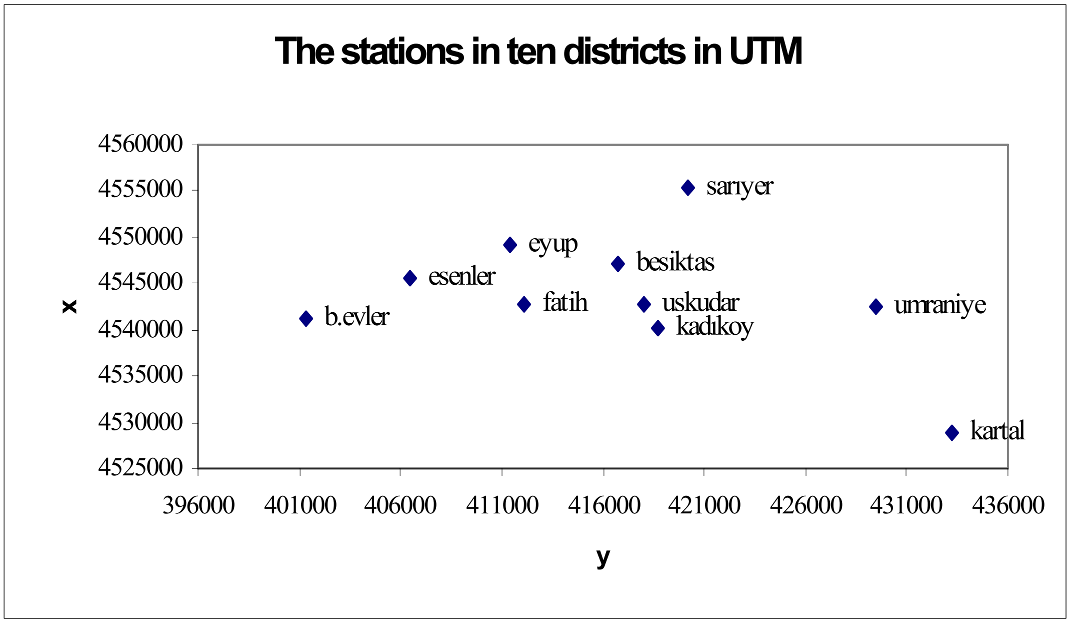


Figure 4.6 Ten stations in Istanbul

Table 4.10 Correlations of 10 districts in Istanbul

SO_2	**Besiktas**	**Esenler**	**Kadikoy**	**Kartal**	**Sarachane**	**Sariyer**	**Umraniye**	**Uskudar**	**Yenibosna**
Besiktas	1,00	0,53	0,59	0,45	0,34	0,52	0,66	0,58	0,71
Esenler	0,53	1,00	0,39	0,51	0,30	0,50	0,53	0,42	0,58
Kadikoy	0,59	0,39	1,00	0,53	0,62	0,40	0,67	0,67	0,61
Kartal	0,45	0,51	0,53	1,00	0,44	0,46	0,60	0,52	0,59
Sarachane	0,34	0,30	0,62	0,44	1,00	0,18	0,58	0,47	0,42
Sariyer	0,52	0,50	0,40	0,46	0,18	1,00	0,48	0,34	0,52
Umraniye	0,66	0,53	0,67	0,60	0,58	0,48	1,00	0,61	0,63
Uskudar	0,58	0,42	0,67	0,52	0,47	0,34	0,61	1,00	0,59
Yenibosna	0,71	0,58	0,61	0,59	0,42	0,52	0,63	0,59	1,00

PM_{10}	**Besiktas**	**Esenler**	**Kadikoy**	**Kartal**	**Sarachane**	**Sariyer**	**Umraniye**	**Uskudar**	**Yenibosna**
Besiktas	1,00	0,26	0,35	0,21	0,26	0,38	0,27	0,30	0,35
Esenler	0,26	1,00	0,65	0,42	0,14	0,67	0,65	0,60	0,50
Kadikoy	0,35	0,65	1,00	0,42	0,38	0,64	0,60	0,73	0,60
Kartal	0,21	0,42	0,42	1,00	0,17	0,44	0,47	0,39	0,25
Sarachane	0,26	0,14	0,38	0,17	1,00	0,33	0,22	0,47	0,35
Sariyer	0,38	0,67	0,64	0,44	0,33	1,00	0,65	0,69	0,55
Umraniye	0,27	0,65	0,60	0,47	0,22	0,65	1,00	0,67	0,53
Uskudar	0,30	0,60	0,73	0,39	0,47	0,69	0,67	1,00	0,60
Yenibosna	0,35	0,50	0,60	0,25	0,35	0,55	0,53	0,60	1,00

CO	**Besiktas**	**Esenler**	**Kadikoy**	**Kartal**	**Sarachane**	**Sariyer**	**Umraniye**	**Uskudar**	**Yenibosna**
Besiktas	1,00	0,70	0,58	0,53	0,28	0,77	0,64	0,43	0,67
Esenler	0,70	1,00	0,62	0,47	0,34	0,73	0,70	0,47	0,70
Kadikoy	0,58	0,62	1,00	0,46	0,43	0,67	0,51	0,36	0,74
Kartal	0,53	0,47	0,46	1,00	0,37	0,55	0,53	0,23	0,56
Sarachane	0,28	0,34	0,43	0,37	1,00	0,31	0,41	0,10	0,61
Sariyer	0,77	0,73	0,67	0,55	0,31	1,00	0,75	0,45	0,64
Umraniye	0,64	0,70	0,51	0,53	0,41	0,75	1,00	0,36	0,66
Uskudar	0,43	0,47	0,36	0,23	0,10	0,45	0,36	1,00	0,38
Yenibosna	0,67	0,70	0,74	0,56	0,61	0,64	0,66	0,38	1,00

4.3.5.1 Forecasting One District with the Nearest District's Values

The objective in this experiment is to take into consideration the effect of air pollution in one district to air pollution in another district. In this experiment, forecasts are made using the air pollution indicator levels of nearest district as input parameter in ANN. Since places close to each other have high correlations, we used Besiktas and Uskudar districts in this experiment. While estimating Besiktas' air pollution indicators for next days, Uskudar's air pollution indicator levels are used as input parameters. Recent 3 days is used as train data. This experiment is run over from 01.08.2005 to 31.07.2006 with 10 neurons in neural network. Results are compared with forecasts for Besiktas without using Uskudar's air pollutants as input parameter as shown in Table 4.11. This experiment gives higher error percentages according to without using Uskudar's indicators as shown in Table 4.12. In this experiment, it is indicated that using the nearest city as another input parameter is not efficient.

Table 4.11 Forecasting a district with the nearest district's values

	Avg SO_2 error %	**Avg PM_{10} error %**	**Avg CO error %**
Tomorrow	47	36	41
The day after tomorrow	49	39	50
Two days after tomorrow	55	40	53

Table 4.12 Forecasting by not using nearest district's values

	Avg SO_2 error %	**Avg Pm_{10} error %**	**Avg CO error %**
Tomorrow	43	33	32
The day after tomorrow	48	41	32
Two days after tomorrow	51	46	36

4.3.5.2 Forecasting a District with Nearest Two Districts' Values

In this experiment, the levels of air pollution indicators of Besiktas are forecasted by using nearest two districts air pollution indicators as input parameters to neural network. Saraçhane and Üsküdar are used as nearest two districts. Recent 3 days used for training data. This experiment is run over from 01.08.2005 to 31.07.2006 with 10 neurons in neural network. The results show that this experiment has higher error percentages than the experiment without using nearest districts as shown in Table 4.13.

Table 4.13 Forecasting a district with nearest two districts' values

	Avg SO_2 error %	**Avg PM_{10} error %**	**Avg CO error %**
Tomorrow	46	42	36
The day after tomorrow	49	43	37
Two days after tomorrow	53	46	39

4.3.5.3 Forecasting Air Pollution with Three Districts' Values Propagated with Distance

The previous experiments that use the air pollution concentrations of nearest districts do not have high accuracy in forecasts. Therefore, we proposed to consider the distance of districts in forecasts. In this experiment by using the distance of three districts to a district a reverse ratio is found between the air pollutants. We calculated reverse ratio because the most distant district has least effect

The districts that enclose Besiktas are used in this experiment. They are Sarachane, Uskudar and Sariyer. First of all, the distance of stations in each district to station in Besiktas is calculated and then the percentage for each district is calculated as shown in Table 4.14.

Table 4.14 Distances to Besiktas

	Distance to Besiktas	**% of reverse proportion**
FATİH	6381	0,32
ÜSKÜDAR	4494	0,45
SARIYER	8963	0,23

The formula of SO_2 in Besiktas is:

-
- New attribute SO_2 = Sarachane SO_2 *0,32 + Uskudar SO_2 *0,45 +Sariyer SO_2 *0,23

This calculated value is given as an input to neural network and forecasts are made by using 3 recent days in train data for Besiktas district. Dataset is one year period from 01.08.2005 to 31.07.2006. The error percentages are shown in Table 4.15.

Table 4.15 Error percentages of forecasting air pollution with three districts values propagated with distance

	Avg SO_2 error %	**Avg PM_{10} error %**	**Avg Co error %**
Tomorrow	46	40	27
The day after tomorrow	48	42	32
Two days after tomorrow	48	43	38

4.3.5.4 Forecasting Air Pollution with Three Districts Values Using Angles

The objective in this experiment is considering the direction of a district with distance. The direction of a district can be calculated by angles.

In this experiment, nearest three districts' air pollutant values are used as inputs for ANN. Three districts Fatih, Üsküdar and Sariyer are chosen as the previous experiment. Than the sinus and cosines values of angles ß1, ß2 and ß3 are calculated according to Besiktas by taking Besiktas as middle point as shown in Figure 4.7. Sum of sinus of air pollution values and sum of cosines of air pollution values of three districts' are calculated and they are used as inputs for ANN as shown in Table 4.16.

In this experiment 10 neurons are used for using recent three days in train data. This experiment is run over from 01.08.2005 to 31.07.2006 for Besiktas district. Error percentages of forecasting air pollution with three districts values using angles are shown in Table 4.17.

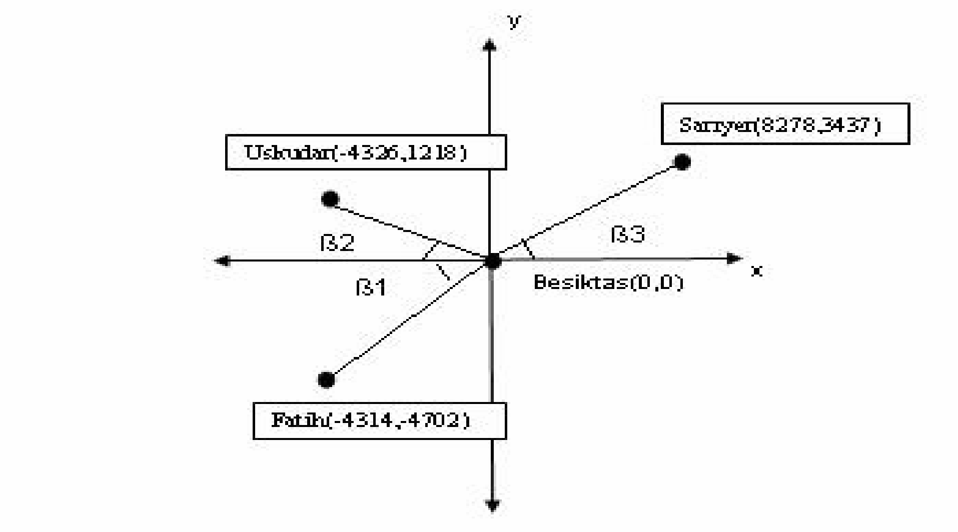


Figure 4.7 Four districts on the x-y plane

Table 4.16 Distances, sinus, cosines values of three districts

	Distance to Besiktas	sin	cos
FATİH **ß1**	6381	0,74	0,68
ÜSKÜDAR **ß2**	4494	0,27	0,96
SARIYER **ß3**	8963	0,38	0,92

Table 4.17 Error percentages of forecasting air pollution with three districts values using angles

	Avg SO_2 error %	**Avg PM_{10} error %**	**Avg Co error %**
Tomorrow	45	36	29
The day after tomorrow	46	36	35
Two days after tomorrow	49	41	38

4.4 EVALUATION

Nine experiments run over the dataset from 01.08.2005 to 31.07.2006. The experiment names and numbers are seen in Table 4.18.

Table 4.18 Experiment names and numbers

Experiment Number	Experiment Name
1	Finding optimum training data set size
2	Arithmetic mean
3	Last true value
4	3 days data in 1 row
5	Day of week parameter
6	Nearest 1 district
7	Nearest 2 district
8	Three districts values using distance
9	Three districts values using angles

Generally forecasts are made for only next day in the literature; but in this study the day after next day and two days after next day are also forecasted. It is more meaningful to evaluate the experiments by next day's error percentages rather than error percentages of one or two days after. The average error percentages of experiments for each air pollution indicator are shown in Figure 4.8 through Figure 4.10.

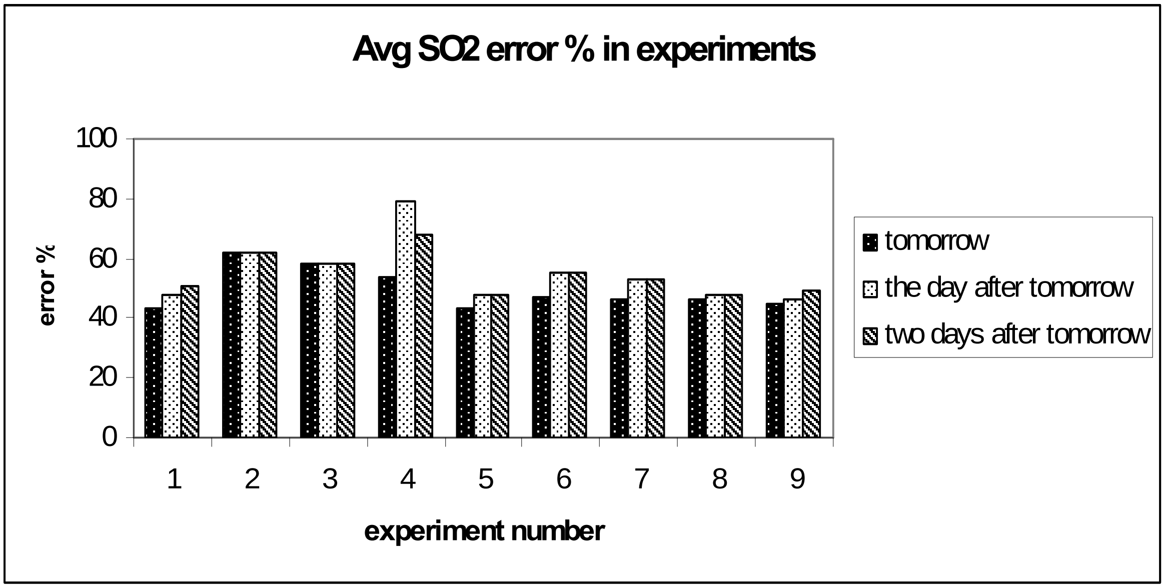


Figure 4.8 Average error percentages of SO_2

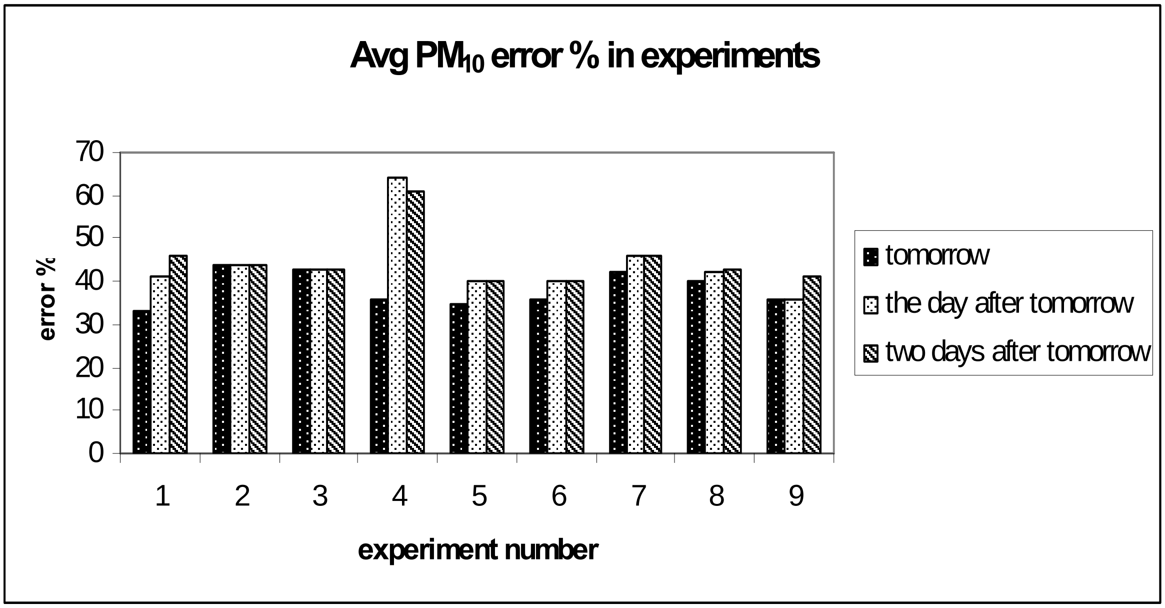


Figure 4.9 Average error percentages of PM_{10}

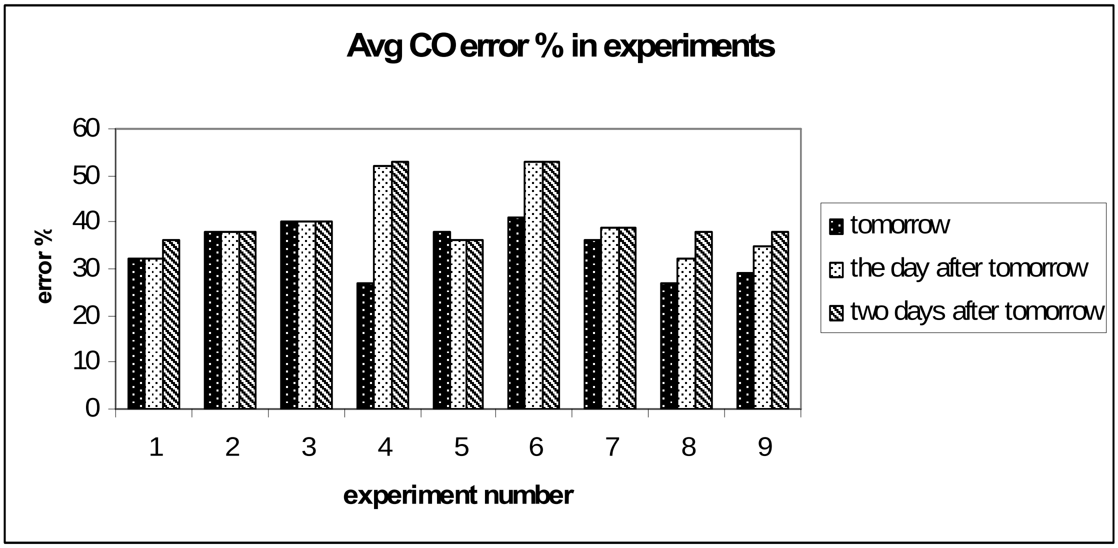


Figure 4.10 Average error percentages of CO

Figure 4.8 shows the error percentages of SO_2. In experiment 5, using the day of week parameter, the error percentages are the lowest. Error percentages of SO_2 are higher than other two air pollutants. Since SO_2 has lower values than other PM_{10} and CO, absolute error percent is higher.

Figure 4.9 shows the error percentages of $PM_{10.}$ The minimum error percentage while forecasting tomorrow is in experiment 1. However experiment 5 has lower error percentages in forecasting the day and two days after tomorrow.

Figure 4.10 is the graph of the error percentages of CO for each experiment. The minimum error percentage while forecasting tomorrow is in experiment 4 three days data in one row experiment. Experiment 8 has lower error percentages for estimating the day after and two days after tomorrow. Since CO has high values, CO absolute error percentage is lower than other two elements.

Experiment 1, finding optimum training dataset size, showed that using recent 3 days as train data is appropriate. Experiment 2 and experiment 3 showed that ANN is required for forecasts. Experiment 4, 3 days data in one row, is appropriate for only forecasting tomorrow. In experiment 5 it is indicated that using the day of week parameter lowers the error percentages. Experiments 6 and 7 showed that nearest cities does not affect the forecasts. Experiments 8 and experiment 9 indicated that geographical models can be used in forecasts because they have high accuracy.

CHAPTER 5

CONCLUSIONS

Air pollution is an important problem of humans and it causes many health problems. It is necessary to develop warning systems to evaluate and control air pollution especially in metropolitan cities.

In this thesis, the web-based air pollution forecasting system is developed for ten districts in Istanbul. The air pollution forecasting system based on four modules: Data collection, database, forecasting and website modules. Three air pollution indicators concentrations SO_2, PM_{10} and CO are predicted by using the relationship between meteorological air conditions and air pollution with ANN. The forecasted air pollution concentrations are presented in the project web site with graphs showing the warning and danger levels for air pollution indicators.

After developing the system, the accuracy of forecasts are determined and to get high accuracy for air pollution forecasts some experiments are done. Totally nine experiments are run over the dataset. Each experiment has different ANN model. All of the experiments run over for one year period from 01.08.2005 to 31.07.2006.

The experiments indicated that using the day of week parameter provides lower error percentages. Also using large data set in training data is not necessary in forecasts. In addition location based forecasting models can be used by calculating the distances and angles between locations.

In the future, web services for the air pollution forecasting system can be developed. Also the ANN model can be enhanced by using the other parameters gathered from www.weather.com. The air pollution forecasting system can be applied in other cities. The last but not least, new ANN models can be developed to forecast other seven air pollution indicators that are not forecasted.

REFERENCES

Akkoyunlu A. and F. Ertürk, "Evaluation of Air Pollution Trends in İstanbul", *Int. J. of Environ. Poll.*; Vol. 18, No. 4, pp. 388-398, 2003.

Antonic, O., Hatic, D., Krian, J., Bukocev, D., "Modelling groundwater regime acceptable for the forest survival after the building of the hydro-electric power plant", *Ecol. Model.*, Vol. 138, No. 1, pp. 277–288, 2001.

Bose, N.K., Liang, P., *Neural Network Fundamentals with Graphs, Algorithms and Applications*, Tata McGraw Hill Publishing Company Limited, New Delhi, India, 1998.

Boznar, M., Lesjak, M., Malker, 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.*, Vol. 27B, No. 2, pp. 221–230, 1993.

Comrie, A.C., "Comparing neural networks and regression model for ozone forecasting", *J. Air Waste Manage. Assoc.*, No. 47, pp. 653–663, 1997.

Defra, The Air Quality Strategy for England, Scotland, Wales and Northern Ireland,2001 http://www.defra.gov.uk/environment/airquality/strategy/index.htm

Dimopoulos, I., Chronopoulos, J., Sereli, A.C., Lek, S., "Neural network models to study relationships between lead concentration in grasses and permanent urban descriptors in Athens city (Greece)", *Ecol. Model.,* Vol. 120, No. 2, pp 157–165, 1999.

Dockery, D.W., Pope III, C.A., Xu, X., Spengler, J.D., Ware, J.H.,Fay, M.E., Ferris, B.G., Speizer, F.E., "An association between air pollution and mortality in six U.S. cities", *The New England J. of Medicine*, Vol. 329, No. 24, pp. 1753-1759,1993.

MathWorks Products Documentation,
http://www.mathworks.com/access/helpdesk/help/helpdesk.html

Gevrey, M., Dimopoulos, I., Lek, S., "Review and comparison of methods to study the contribution of variables in artificial neural network models" *Ecol. Model.,* Vol. 160, No. 3, pp. 249–264, 2003.

Gardner, M. W. and Dorling, S. R., “Artificial neural networks (the multilayer perceptron) - a review of applications in the atmospheric sciences”, *Atmos. Environ*, Vol. 32, pp. 2627-2636, 1998.

Gardner, M. W. and Dorling, S. R., “Neural network modelling and prediction of hourly NOx and NO2 concentrations in urban air in London”, *Atmos. Environ*, Vol. 33, No. 5, pp. 709-719, 1999.

Gardner, M. W. and Dorling, S. R., “Statistical surface ozone models: an improved methodology to account for non-linear behaviour”, *Atmos. Environ.*, Vol. 34, No. 1, pp. 21-34, 2000.

Heymans, J.J., Baird, D., “A carbon flow model and network analysis of the northern Benguela upwelling system”, *Ecol. Model.*, Vol. 126, pp. 9–32, 2000.

Hornik, K., Stinchcombe, M., White, H., “Multi layer feed forward networks are universal approximators”, *Neural Netw.*, Vol. 2, pp. 359–366, 1989.

Karaca, F., Alagha O., Nikov A., “NN-AirPol: A neural-networks-based method for air pollution evaluation and control”, *J. of Environ. Pol.*, Vol. 28, No. 3/4, pp. 215-221, 2006.

Karul, C., Soyupak, S., Cilesiz, A.F., Akbay, N., Germen, E., “Case studies on the use of neural networks in eutrophication modeling”, *Ecol. Model.*, Vol. 134, pp. 145–452, 2000.

Künzli, N., Kaiser, R., Medina, S., Studnicka, M., Chanel, O., Filliger, P., Herry, M., Horak, F., Puybonnieux-Texier, V., Quénel, P., Schneider, J., Seethaler, R., Vergnaud J. C., and Sommer, H., “Public-health impact of outdoor and traffic-related air pollution”, *Eur. Assess,* Vol. 356., pp. 795-801, 2000.

Lawrence Berkeley ELSI Lab, http://www.lbl.gov/Education/ELSI/Frames/pollution-defined-f.html

Lek, S., Delacoste, M., Baran, P., Dimopoulos, I., Lauga, J., Aulagnier, S., “Application of neural networks to modelling nonlinear relationships in ecology”, *Ecol. Model.,* Vol. 90, pp. 39–52, 1996.

Longhurst, J.W.S., Lindley, S.J.,Watson, A.F.R., Conlan, D.E., “The introduction of local air quality management in the United Kingdom: a review and theoretical framework”, *Atmos. Environ.,* Vol. 30, pp. 3975–3985, 1996.

Manel, S., Dias, J.M., Ormerod, S.J., “Comparing discriminant analysis, neural networks and logistic regression for predicting species distributions: a case study with a Himalayan river bird”, *Ecol. Model.*, Vol. 120, No. 2/3, pp. 337–347, 1999.

McDonnell, W.F., Nishikawa, N., Peterson, F.F., Chen, L.H., Abbey, D.E., "Relationships of mortality with the fine and coarse fractions of long term ambient PM_{10} concentrations in nonsmokers", *J. of Exposure Analysis and Environ. Epidemiology*, Vol. 10, pp. 427-436, 2000.

Monteiro A, Lopes M, Miranda AI. Carlos Borrego, Vautard R, "Air Pollution Forecast in Portugal: a demand from the new air quality framework directive", *Int J Environ Pol,* Vol. 5, pp. 1-9, 2005.

Nagendra, S.M.S., Khare, M., "Artificial neural network based line source models for vehicular exhaust emission predictions of an urban roadway", *J. Transport. Res. D Transport Environ.*, Vol. 9, pp. 199–208, 2004.

Nikov A., F. Karaca, O. Alagha, A. Kurt, H. Hakkoymaz, "AirPolTool: A web based tool for Istanbul air pollution forecasting and control", 3rd Int. Symposium on Air Quality Management at Urban, AQM2005, September 2005.

Nunnari, G., Nucifora, A.F.M., Randieri, C., "The application of neural techniques to the modelling of time-series of atmospheric pollution data", *Ecol. Model.*, Vol. 111, No. 2/3, pp. 187–205, 1998.

Olden, J.D., Jackson, D.A., "Illuminating the black box: a randomization approach for understanding variable contributions in artificial neural networks", *Ecol. Model.*, Vol. 154, No. 1, pp. 135–150, 2002.

Park, Y.S., Cereghino, R., Compin, A., Lek, S., "Applications of artificial neural networks for patterning and predicting aquatic insect species richness in running waters", *Ecol. Model.*, Vol. 160, No. 3, pp. 265–300, 2003.

Paruelo, J.M., Tomasel, F., "Prediction of functional characteristics of ecosystems: a comparison of artificial neural Networks and regression models", *Ecol. Model.*, Vol. 98, No. 2, pp. 173–186, 1997.

Roiger R, Geatz M, *Data Mining A Tutorial-based Primer*, Addison Wesley, 2005.

Ryan, M.,Muller, C., Di, H.J., Cameron, K.C., "The use of artificial neural networks (ANNs) to simulate NO2 emissions from temperate grassland ecosystem", *Ecol. Model.,* Vol. 175, No. 2, pp. 189–194, 2004.

Saral A. and Ertürk F., "Prediction of Ground Level SO2 Concentration Using Artificial Neural Networks", *2nd Int. Symposium on Air Quality Management at Urban, Regional, and Global Scales,* pp. 388-395, 2001.

Schwartz, J., "Air pollution and hospital admissions for respiratory disease", *Epidemiology,* Vol. 7, pp. 20-28, 1996.

Tittanen, P., Timonen, K.L., Ruuskanen, J., Mirme, A., Pekanen, J., “Fine particulate air pollution, resuspended road dust and respiratory health among symptomatic children”, *European Respiratory J.*, Vol. 13, pp. 266-273, 1999.

Zurada, J.M., *Introduction to Artificial Neural Systems,* West Publishing Company, Mumbai, India, 1997.