AN ARTIFICIAL NEURAL NETWORK-BASED MODEL FOR SHORT-TERM PREDICTIONS OF DAILY-mean PM10 CONCENTRATIONS

G. DEMIR*a, H. OZDEMIR*a, H. K. OZCANb, O. N. UCANC, C. BAYATd

*aFaculty of Engineering, Department of Environmental Engineering, Bahcesehir University, 34 349 Besiktas, Istanbul, Turkey
E-mail: goksel.demir@bahcesehir.edu.tr

*bEngineering Faculty, Environmental Engineering Department, Istanbul University, 34 320 Avcilar, Istanbul, Turkey

*cEngineering Faculty, Electrical-Electronics Engineering Department, Istanbul University, 34 320 Avcilar, Istanbul, Turkey

*dBuyukcekmece Beykent University, 34 500 Istanbul, Turkey

Abstract. Prediction of particulate matter (PM) in the air is an important issue in control and reduction of pollutants in the air. One of the most useful methods to forecast atmospheric pollution is artificial neural network (ANN) because of its high ability to forecast the atmospheric events. In this study ANN technique has been used to predict the PM10 concentration in Istanbul. Meteorological data and PM10 data, which had been collected from Sariyer-Bahcekoy for the one year data, were used. The data were separated into two groups for training and testing the model. The odd days were used for training and the remaining was used for the testing. The transfer function was sigmoid function. In the model, different hidden neuron numbers were altered for proposed ANN structure. We have altered number of neurons for hidden layer between 2 to 10. The prediction of PM10 of the model during the years 2004–2005 follows the actual values with success, with the best calculated correlation coefficient 0.60.

Keywords: PM10, artificial neural networks, prediction.

AIMS AND BACKGROUND
Particulate matter (PM) is the general term used for a mixture of solid particles and liquid droplets found in the air. There is a considerable variety in the chemical makeup and size of particles. Particulate pollutants, divided solids and liquids, include dust, fumes, smoke, fly ash, mist, and spray. Under proper conditions, particulate pollutants will settle out of the atmosphere. Prediction of PM in the air is an important issue in control and reduction of pollutants in the air. In particular, fine particles that are smaller than 2.5 or 10 µm in diameter are defined as PM2.5 or PM10 (Refs 2–4). Particulates may be classified and discussed according to their
physical, chemical or biological characteristics. Physical characteristics include size, mode of information, settling properties and optical qualities. Chemical characteristics include organic or inorganic composition, and biological characteristics related to their classification as bacteria, viruses, spores, pollens, etc. Some of the more common organics found in particulates include phenols, organic acids and alcohols. Common inorganic found in particulates include nitrates, sulphates, and metals such as iron, lead, manganese, zinc and vanadium. The biological particles in the atmosphere include protozoa, bacteria, viruses, fungi, spores, pollens and algae. Microorganisms generally survive for only a short time in the atmosphere because of the lack of nutrients and ultraviolet radiation from the sun. Anthropogenic sources of PM10 are combustion and industrial processes. Vehicular traffic generates dust from the road which has PM10 content and remains suspended for several hours. Black smoke from the exhausts of diesel vehicles is also an important source of PM10 (Ref. 5).

At high concentrations, suspended PM poses health hazards to humans, particularly those susceptible to respiratory illness. The size of the particle is a main determinant of where in the respiratory tract the particle will come to rest when inhaled. Because of the size of the particle, they can penetrate the deepest part of the lungs. Larger particles are generally filtered in the nose and throat and do not cause problems, but PM smaller than about 10 μm referred to as PM10, can settle in the bronchi and lungs and cause health problems. The nature and extent of the ill effects that may be linked to suspended particulates depend upon the concentration of particulates. Approximately 40% of the particles between 1 and 2 μm in size are retained in the bronchioles and alveoli. Particles ranging in size from 0.25 to 1 μm show a decrease in retention, because many particles in this range are breathed in and out again. However, particles smaller than 0.25 μm show another increase in retention because of the Brownian motion, which results in impingement.

EXPERIMENTAL

MODEL DATA

Istanbul is the largest urban center in Turkey and located in the northwest. In the composition of model, the meteorological data used in this study were obtained from the Turkish State Meteorological Service. These measurement results are the daily average results between January 2004 and September 2005 for the Istanbul–Sariyer station.

In the study, 8 different meteorological factors were considered for the prediction of the PM10 concentration with the combination of PM10 data: temperature (T), wind speed (WS), maximum wind speed (MWS), cloud cover (C), relative humidity (RH), rainfall (R), pressure (P), and sunshine (S).
ARTIFICIAL NEURAL NETWORKS

ANN is a nonlinear informational processing device, which is built from interconnected elementary processing devices called neurons. ANN is mathematical representations of a biological neural network. In the mathematical model, neurons and input paths are represented as processing elements and interconnections, respectively. ANN models are inspired by the biological neural system, with capability to learn, store and recall information based on a given training dataset. The most basic and commonly used ANN is the multi-layer perception (MLP). This consists of at least three or more layers, an input layer, an output layer, and a number of hidden layers (Fig. 1). Back propagation (BP) algorithm, as one of the most useful training algorithms for the MLP, is a gradient descent technique to minimise the error through a particular training pattern in which it adjusts the weights by a small amount at a time.

Fig. 1. Structure of ANN network with four hidden layer

To develop an ANN model, the network is processed through two stages: training/learning stage and testing/validation stage. In the training stage, the network is trained to predict an output based on input data. In the testing stage, the network is tested to stop or continue training it, and it is used to predict an output. In our study, training and testing data sets belong to same time periods but they consist of different data. The data were separated into two groups for training and testing the model. The odd days were used for training and the remaining was used for the testing (odd and even numbered days). The transfer function was sigmoid function.

In this study, ANN model was run on WEKA (Waikato Environment for Knowledge Analysis). WEKA is a collection of machine learning algorithms for solving real-world data mining problems. It is written in Java and runs on almost
any platform. The algorithms can either be applied directly to a dataset or called from your own Java code. Some of program details are given in Figs 2 and 3.

**Fig. 2.** Explorer window

**Fig. 3.** Changing the parameters
**Correlation coefficient.** The correlation coefficient is a measure of how well trends in the predicted values follow trends in past actual values. It is a measure of how well the predicted values fit with the real-life data. The correlation coefficient is a number between 0 and 1. If there is no relationship, between the predicted values and the actual values, the correlation coefficient is zero (0) or very low (the predicted values are no better than random numbers). As the strength of the relationship between the predicted values and actual values increases so does the correlation coefficient. A perfect fit gives a coefficient of 1.0. Thus the higher the correlation coefficient the better.

\[ r = \frac{n \sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} \sqrt{n(\sum y^2) - (\sum y)^2}} \]

where \( r \) is the correlation coefficient between two random variables \( x \) and \( y \), and \( n \) – the number of pairs of data.

**Mean absolute error (MAE).** The mean absolute error is a weighted average of the absolute errors \( e_i = f_i - y_i \), where \( f_i \) is the prediction and \( y_i \) the true value. Note that it can also include the relative frequencies as weight factors.

\[ \text{MAE} = \frac{1}{n} \sum_{i=1}^{n} |f_i - y_i| = \frac{1}{n} \sum_{i=1}^{n} |e_i| . \]

**Root mean square error (RMSE).** The root mean square error (RMSE) (also root mean square deviation (RMSD)) is measure of the differences between values predicted by a model and the values actually observed from the thing being modelled or estimated. The RMSE of a model \( \hat{\theta} \) with respect to the estimated parameter \( \theta \) is defined as the square root of the mean squared error:

\[ \text{RMSD}(\hat{\theta}) = \sqrt{\text{MSE}(\hat{\theta})} = \sqrt{E((\hat{\theta} - \theta)^2)} . \]

**Relative absolute error**

\[ E_i = \frac{\sum_{j=1}^{n} |P_{ij} - T_j|}{\sum_{j=1}^{n} |T_j - \bar{T}|} \]

where \( P_{ij} \) is the value predicted by the program \( i \) for sample case \( j \) (out of \( n \) sample cases); \( T_j \) – the target value for sample case \( j \); and \( \bar{T} \) – given by the formula:

\[ \bar{T} = \frac{1}{n} \sum_{j=1}^{n} T_j . \]
For a perfect fit, the numerator is equal to 0 and \( E_i = 0 \). So, the \( E_i \) index ranges from 0 to infinity, with 0 corresponding to the ideal.

**Root relative squared error.** Root relative squared error is found both by taking the squares of numerator and the denominator than the square root of the whole expression.

\[
E_i = \sqrt{\frac{\sum_{j=1}^{n} (P_{ij} - T_j)^2}{\sum_{j=1}^{n} (T_j - \bar{T})^2}}.
\]

**RESULTS AND DISCUSSION**

In this paper, PM10 concentrations are predicted using ANN approach for actual measured values. The neuron number of hidden layer is an important factor where the number cascade connected hidden layer is not so effective. In our study, we have compared various hidden layers with different neuron numbers for proposed ANN structure. We have altered number of neurons for hidden layer between 1 to 10. ANN model outputs are given in Figs 4–7.

In the learning trials, learning rate and the momentum are given different values between 0.1 and 0.6. Towards these values, by using different hidden layers, 10 trials are made. Because more than these values resulted with over learning problem, trials are optimised to these values. Produced weight factors for the learning process are applied to a different test set and finally PM10 values are predicted.

**Fig. 4.** Daily changes of PM10 values, and predicted with ANN modelling test method during the years 2004–2005 (hidden layer=1)
After the determination of daily PM10 value by ANN modelling technique with WEKA, performance of the model is calculated by comparing the produced model predictions with the observed results. For this reason, by using the statistical performance evaluation parameters that WEKA calculated the obtained values for each run are given in Table 1.
It is seen from the data of Table 1 that the highest correlation coefficient was achieved with hidden number 1 \((r=0.60)\). The lowest model results were obtained with hidden neuron 10 \((r=0.48)\).

### CONCLUSIONS

In this work, closeness between the predicted and observed values of PM10 is studied at the end of the learning and test trials. If the results are examined, the best successful model performance appears to be in the trial which the hidden neuron number is 1. Proposed model results show that the prediction of PM10 of the model during the years 2004–2005 follows the actual values with success, which can be seen with the calculated correlation coefficient 0.60. Graphical illustrations can be found in figures, which are showing the closeness between the observed values and the model prediction. As a result, with the use of 8 different meteorological parameters, it is determined that ANN modelling technique can be used in the prediction of particulate matter. The performance of the proposed model network can further be improved by using different data sets.

### REFERENCES


Received 18 June 2009
Revise 31 July 2009