Predicting tropospheric ozone concentrations in different temporal scales by using multilayer perceptron models

Bilge Özbay a,*, Gülşen Aydın Keskin b, Şenay Çetin Doğruparmak a, Savaş Ayberk a

a Department of Environmental Engineering, Kocaeli University, 41380 Kocaeli, Turkey
b Department of Industrial Engineering, Kocaeli University, 41380 Kocaeli, Turkey

ARTICLE INFO

Article history:
Received 18 December 2010
Received in revised form 14 March 2011
Accepted 26 March 2011
Available online 4 April 2011

Keywords:
Tropospheric ozone concentration
Industrial area
Multilayer perceptron
Principle component analysis
Meteorological factors

ABSTRACT

This study encompasses ozone modeling in the lower atmosphere. It was aimed to develop an appropriate neural network model in order to predict ozone concentrations in various temporal scales as a function of meteorological variables and air quality parameters. All data were collected from Dilovasi, Turkey as this site represents typical industrial regions with major air pollution problems. In the study performance of the multilayer perceptron models were tested for both annual and seasonal periods as meteorological conditions highly influence the ozone levels. Among the various architectures, a network of two hidden layers with fifteen neurons was found to give successful predictions. Modeling efficiency of the developed network was also evaluated for day light and night time data of warming season exhibiting highest ozone levels. Furthermore, principle component analysis was performed by using annual data in order to reduce the number of input variables describing ozone formation. Model run with principle components has also provided satisfying performance.

© 2011 Elsevier B.V. All rights reserved.

1. Introduction

Ozone (O3) formation is a well-known phenomenon resulting from complex chemical reactions of nitrogen oxides and organic species in the presence of solar radiation (Abdul-Wahab et al., 2005). Both precursor emissions and meteorological conditions have important roles in this formation mechanism. Tropospheric O3 is the main index substance of photochemical smog events and it is known as a key precursor of the hydroxyl radicals (OH) controlling the oxidizing power of the atmosphere (Abdul-Wahab et al., 2005). As concentrations of trace species (methane, carbon monoxide, sulfur dioxide, etc.) are effected by OH radicals, ground level O3 influences air quality substantially through a number of pathways (Pouliada et al., 1991). It has also direct negative impacts on environment and public health when present in sufficient quantities (Abdul-Wahab and Al-Alawi, 2002).

In recent years many studies on tropospheric O3 have been receiving extensive attention as it is an important secondary air pollutant influencing air quality (Beck et al., 1992; Chan et al., 1998; Guicherit and Roemer, 2000; Hsu, 2007; Kruzscy et al., 2007; Kulkarni et al., 2011; Oltmans et al., 2006). These investigations have focused on some basic topics such as monitoring and forecasting O3 levels, impact assessment of emission reduction, etc. Especially studies about effective short-term prediction of O3 concentrations have privileged importance for accurate public warning (Ballester et al., 2002). There are two main ways used to forecast tropospheric O3 levels; i: 3-dimensional air quality models which integrate chemistry, transport and dispersion, ii: statistical models which generally directly connect meteorological conditions to level of O3 (Dutot et al., 2007).

Statistical models estimating O3 concentrations are numerous and may be evaluated in four basic groups (Baur et al., 2004). These are regression-based models (Caballero et al., 2007; Sousa et al., 2006), extreme value approaches (Chock and Sluchak, 1986; Smith, 1989), space-time models (Rao et al., 1995, 1997) and neural network models (Capone, 1996, Elkamel et al., 2001). However complex and sometimes non-linear relationships of the variables concerning O3 formation can make many of these statistical models awkward and complicated (Comrie, 1997). In such cases artificial neural networks, particularly multilayer perceptrons (MLP), are commonly preferred due to the superiority of learning complex non-linear interrelationships between independent and dependent variables (Baughman and Liu, 1990; Gardner and Dorling, 1998). Furthermore, it is concluded that neural network applications combined with principle component analysis (PCA) provide simplifications in modeling as PCA is a useful tool reducing number of input variables and allowing us to observe the sources of variation in data sets (Kovač-Andrić et al., 2009).

Investigating related literature, it is seen that O3 modeling studies were performed for various regions (residential, urban, industrial, etc.) and different temporal scales (annual, seasonal, etc.). Tropospheric O3 in industrialized areas, particularly worths investigation as O3 precursors come from anthropogenic sources including industrial
and vehicular emissions of nitrogen oxides and hydrocarbon species. For this reason, in this study we have evaluated the performance of MLPs for modeling O₃ levels in different temporal scales for a heavily industrialized region Dilovasi, Turkey.

2. Material and method

2.1. Data collection

2.1.1. Studied area

Dilovasi, is a well-known industrial region situated in the northwest of Turkey. It covers approximately 118 km² area and there are five main industrial zones with many factories working on various sectors such as glass, chemistry, motor vehicles, iron and steel industry, etc. Transport facilities have improved with two motorways, railway and many seaports. Developing financial situation has caused continuous migrations to the region. Population was over 50,000 in 2009. Fig. 1, shows the satellite image of the region taken from a height of 1000 m. Residential area lies between latitude 40°78′–40°79′ North and longitude 29°53′ to 29°55′ East whereas industrial plants spread throughout the region.

As a result of heavily industrialization ecology has been destroyed intensely in Dilovasi. Air pollution is one of the major environmental problems of the region and it threatens public health seriously. Industrial zones, busy traffic and increasing residential areas are evaluated as the main sources of this problem. Geographical structure of the region also enhances air pollution in Dilovasi as it is located in a gully with 10 m altitude. Especially in summer season, photochemical smog is commonly reported in the region. Air temperature varied between 7 and 25 °C and July was the warmest month for the studied period (September 2008–August 2009). October, November, May and August were rainless months and highest average rainfall was determined in January 2009 with 0.107 mm. Wind rose given in Fig. 1 summarizes directions and speeds of winds for the studied region. As seen from the wind rose, prevailing winds were north wind, northwester and northeastern. In the wind rose light green, yellow, red and dark blue colors represent the wind speeds in the range of 0.5–2.1 m/s, 2.1–3.6 m/s, 3.6–5.7 m/s and 5.7–8.8 m/s, respectively.

2.1.2. Measurement of the variables

Pollutants measured included methane (CH₄), non-methane hydrocarbons (NMHC), carbon monoxide (CO), nitrogen monoxide (NO), nitrogen dioxide (NO₂), sulfur dioxide (SO₂), ozone (O₃) and particulate materials (PM₁₀). Meteorological parameters monitored simultaneously included wind speed (WS) and wind direction (WD), air temperature (T), rain (R), relative humidity (H), pressure (P) and solar radiation (SR). Concentrations of pollutants were determined with hourly measurements carried out in two stations installed in the region (Station of Dilovasi Industrial Zone Office and Station of Kocaeli Environmental and Forestry Department) for period between September 2008 and August 2009. SO₂, NO, NO₂, CO, O₃ and PM₁₀ were measured in the Station of Kocaeli Environmental and Forestry Department. SO₂ was determined by using Thermo 43 analyzers working according to the UV lamp principle. NO and NO₂ gases were measured with devices working due to chemical reflection method. CO measurements were achieved by using gas filter correlation method. O₃ concentrations were determined according to the active sampling method by using UV adsorption principle measurements. PM was measured using light refractory principle. Concentrations of CH₄ and NMHC were measured in Station of Dilovasi Industrial Zone Office by using Synspec Alpha 115 analyzer working according to FID measurement principle.
2.2. Modeling methodology

2.2.1. Multilayer perceptron models

As multilayer perceptrons (MLPs) have the potential to describe highly non-linear relationships, they are among the commonly preferred methods in atmospheric modeling studies. They consist of a system of simple interconnected neurons, or nodes non-linear mapping between an input vector and an output vector (Fig. 2). The nodes are connected by weights and output signals which are a function of the sum of the inputs to the node modified by a simple non-linear transfer, or activation, function (Coman et al., 2008). Tangent hyperbolic function that produces output values ranging between −1 and +1, is commonly preferred in neural network applications (Neuro Solutions, 2003):

\[ f(x_i) = \tanh(x_i) \]

In Eq. (1) \( \beta \) controls the slope of the function. The tangent transfer function is also known as the hyperbolic tangent function and it can be expressed with the following equation for \( \beta = 1 \):

\[ \tanh(x) = \frac{2}{1 + e^{-2x}} - 1 \]

The output of each neuron is then a smooth non-linear function of the weighted inputs.

In the first step of the modeling study, appropriate data were subdivided into two subsets, one for training purposes (the ‘training’ set) and the other for validation purposes (the ‘test’ set). Considering relevant literature more data (approximately 2/3 of the total) were used for training purpose. Data preparation procedure was repeated for the data sets of each temporal scales. Then input data were scaled from 0 to 1 by normalization process due to following formula:

\[ NI_i = \frac{l(i,j) - \min(j)}{\max(j) - \min(j)} \]

where, \( l \) is the input value, \( NI \) is the standardized value, \( l \) is the number of measurements, and \( j \) is the measured value of the variable (Keskin et al., 2010).

Thirdly network architectures were determined involving decisions on items such as number of hidden layers, number of neurons in each layer, type of network to be used, etc. Backpropagation feed-forward network type was used for all tests in this study.

2.2.2. Principle component analysis

Principle component analysis (PCA) as a multivariate statistical method is used to find a small number of factors from a data set of many correlated variables (Astell et al., 2004). It is commonly employed in air quality studies in order to separate interrelationships into statistically independent basic components (Abdul-Wahab et al., 2005; Sousa et al., 2006; Statheropoulos et al., 1998; Vaidya et al., 2000).

In factor analysis firstly a set of factors are extracted from a data set considering eigen values. In order to make the interpretation of the factors that are considered relevant, the first selection step is generally followed by a rotation of the factors that were retained. Varimax, which was developed by Kaiser (1960), is indubitably the most popular rotation method by far (Abdi, 2003). Obtained factor loads represent the contribution of each variable in a specific principal component. Principle components are calculated by multiplying standardized data matrix with previously calculated weights. In the study, applicability of the PCA was evaluated due to Bartlett’s sphericity test. These calculations were performed by using SPSS 17 statistical program.

2.2.3. Performance indices

The size of the standard error (RMSE) and the correlation coefficients between the actual and the predicted values were used to compare the models. Mathematical expressions for standard error (\( \sigma \)) and determination coefficient (\( R^2 \)) are given below:

\[ \sigma = \sqrt{\frac{\sum (P_i - O_i)^2}{n}} \]

\[ R^2 = \frac{\sum_{i=1}^{n} (O_i - \bar{O})^2 - \sum_{i=1}^{n} (O_i - P_i)^2}{\sum_{i=1}^{n} (O_i - \bar{O})^2} \]

where \( P_i \) and \( O_i \) represent the actual, predicted and average predicted values, respectively (Coman et al., 2008).

3. Results and discussion

3.1. Forecasting annual \( O_3 \) concentrations

In the first step of the study, annually measured data were used for \( O_3 \) modeling. With this aim 3620 data were used for training procedure whereas 1811 were used for simulation. Some test results obtained with different network types are summarized in Table 1.

As seen from the table, the highest prediction efficiency was obtained with Model XI using TRAINR training function, LEARNGDM adaptation function and TANSIG transfer function. 2 hidden layers were used in the model and 1000 epochs provided sufficient training performance. In the model, \( R^2 \) was calculated as 0.95 for the relation between actual and predicted \( O_3 \) levels (Fig. 3(a)).

For forthcoming modeling studies (forecasting seasonal, light and dark hour's \( O_3 \) levels, etc.) were performed by using network type used in Model XI as it provided highest simulation.

3.1.1. Modeling \( O_3 \) levels by using principle components

Principle component analysis was applied in order to reduce the number of components and simplify the modeling process. Varimax rotation process was applied and variables were grouped into four principle components (PCs) considering coefficients in rotated component matrix. PC1 was consisted of NO2, NO, PM10, SO2, NMHC and O3; PC2 was consisted of T, P, CH4, and SR whereas WS and WD were grouped into PC3 and CO and R were grouped into PC4. Training and test data sets were arranged again consisting of 3620 and 1811 measurements, respectively.
Predetermined network type was used for modeling. Using principle components, $R^2$ value of 0.87 was determined for the correlation between predicted and actual O$_3$ levels (Fig. 3(b)). $\sigma$ was calculated as 0.001.

### 3.2. Forecasting seasonal O$_3$ concentrations

As mentioned before, one of the primary objectives of the study is to evaluate the performance of MLPs in O$_3$ modeling comparatively for different temporal scales. So investigating different seasonal periods is essential due to well-known effect of T on O$_3$ concentrations. Fig. 4, shows the monthly average O$_3$ concentrations and T levels for the studied period. As seen from the figure, there is a strong relation between T and O$_3$ with $R^2 = 0.608$. Considering the tendencies in T levels; April, May, June, July and August months were evaluated in warming period whereas November, December, January, February and March were evaluated as cooling period in the study. Raw data were used in these applications.

#### 3.2.1. Warming period

Efficiency of MLP was evaluated for warming period by using predetermined network architecture. In the test, 1734 and 921 samples were used for training and simulation, respectively. The graph of obtained prediction efficiency is presented in Fig. 5(a).

The relation between actual and predicted O$_3$ levels was described with $R^2$ of 0.96 and $\sigma$ of the model was 0.03.

#### 3.2.2. Cooling period

In order to model O$_3$ levels in cooling period, 1244 and 691 samples were used for training and simulation processes, respectively. Predicted and actual O$_3$ levels are shown in Fig. 5(b).

Determination coefficient of 0.88 was obtained for the relation between actual and measured O$_3$ levels in the model.

### 3.3. Forecasting daylight and night O$_3$ concentrations

SR is an effective meteorological factor on tropospheric O$_3$ levels as O$_3$ formation is explained with complex photochemical reactions.

---

**Table 1**

Properties of chosen neural network models tested for modeling annual ozone data.

<table>
<thead>
<tr>
<th>Model no</th>
<th>Training function</th>
<th>Adaptation function</th>
<th>Transfer function</th>
<th>Number of hidden layers</th>
<th>Number of epochs</th>
<th>$R^2$</th>
<th>$\sigma$</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>TRAINGDM</td>
<td>LEARNGDM</td>
<td>TANSIG + TANSIG</td>
<td>2</td>
<td>1000</td>
<td>0.40</td>
<td>0.005</td>
</tr>
<tr>
<td>II</td>
<td>TRAINGDM</td>
<td>LEARNGDM</td>
<td>TANSIG + TANSIG</td>
<td>2</td>
<td>3000</td>
<td>0.82</td>
<td>0.003</td>
</tr>
<tr>
<td>III</td>
<td>TRAINGDM</td>
<td>LEARNGDM</td>
<td>TANSIG + TANSIG + TANSIG</td>
<td>3</td>
<td>3000</td>
<td>0.49</td>
<td>0.004</td>
</tr>
<tr>
<td>IV</td>
<td>TRAINGDM</td>
<td>LEARNGDM</td>
<td>LOSIG + LOGSIG</td>
<td>2</td>
<td>1000</td>
<td>0.24</td>
<td>0.004</td>
</tr>
<tr>
<td>V</td>
<td>TRAINLM</td>
<td>LEARNGDM</td>
<td>LOSIG + LOGSIG</td>
<td>2</td>
<td>1000</td>
<td>0.26</td>
<td>0.009</td>
</tr>
<tr>
<td>VI</td>
<td>TRAINLM</td>
<td>LEARNGDM</td>
<td>TANSIG + LOGSIG</td>
<td>2</td>
<td>1000</td>
<td>0.74</td>
<td>0.03</td>
</tr>
<tr>
<td>VII</td>
<td>TRAINLM</td>
<td>LEARNGDM</td>
<td>TANSIG + TANSIG</td>
<td>2</td>
<td>1000</td>
<td>0.85</td>
<td>0.002</td>
</tr>
<tr>
<td>VIII</td>
<td>TRAINLM</td>
<td>LEARNGDM</td>
<td>LOSIG + LOGSIG + LOGSIG</td>
<td>3</td>
<td>1000</td>
<td>0.81</td>
<td>0.002</td>
</tr>
<tr>
<td>IX</td>
<td>TRAINLM</td>
<td>LEARNGDM</td>
<td>TANSIG + TANSIG + TANSIG</td>
<td>3</td>
<td>1000</td>
<td>0.68</td>
<td>0.03</td>
</tr>
<tr>
<td>X</td>
<td>TRAINLM</td>
<td>LEARNGDM</td>
<td>LOSIG + LOGSIG</td>
<td>2</td>
<td>1000</td>
<td>0.93</td>
<td>0.02</td>
</tr>
<tr>
<td>XI</td>
<td>TRAINR</td>
<td>LEARNGDM</td>
<td>TANSIG + TANSIG</td>
<td>2</td>
<td>1000</td>
<td>0.95</td>
<td>0.02</td>
</tr>
<tr>
<td>XII</td>
<td>TRAINR</td>
<td>LEARNGDM</td>
<td>TANSIG + TANSIG + TANSIG</td>
<td>3</td>
<td>2000</td>
<td>0.55</td>
<td>0.04</td>
</tr>
</tbody>
</table>

---

**Fig. 3.** Predicted O$_3$ levels versus measured O$_3$ levels for (a) the annual data set and (b) principle components.

**Fig. 4.** Monthly average O$_3$ (µg/m$^3$) and T (°C) levels (Sep. 2008–Aug. 2009).
involving nitrogen oxides and volatile organic carbons (Kovač-Andrić et al., 2009).

Fig. 6 exhibits average hourly SR and O3 values for the studied region. As seen from the figure, SR reaches highest values in midday hours. Similarly O3 levels show increase tendency until 3 p.m. Decreasing SR lowers O3 concentrations in dark hours. In order to investigate this affect on O3 modeling, dark and light hours’ data of warming season, exhibiting highest O3 levels, were evaluated individually. With this aim, time interval between 6 a.m and 8 p.m was evaluated as light hours while night time was represented by remaining hours.

1037 and 533 samples were used respectively for training and simulating the model predicting light time O3 levels (Fig. 7(a)). Model efficiency was described with $R^2$ of 0.94 and $\sigma$ was calculated as 0.04.

In the model predicting dark hours’ O3 levels, training process was run with 719 samples. Simulation was performed with 366 samples and modeling efficiency of 0.95 (Fig. 7(b)) was obtained with 0.004 standard error.

4. Conclusions

In the present study, efficiency of MLP was evaluated for modeling tropospheric O3 levels of a heavily industrialized region in different temporal scales. Among the various models developed for predicting annual O3 data, best efficiency was obtained by using perceptron model with two hidden layers (Model XI). TRAINR training function, LEARNCDM adaptation function and TANSIG transfer function were used in the chosen model. In the study, it was obtained that an increasing number of hidden layers decrease modeling efficiency. Forthcoming prediction studies for other relevant periods were performed by using this network architecture. Calculated determination coefficients were 0.95, 0.96 and 0.88 for annual, warming and cooling periods, respectively. $R^2$ value of 0.96 is particularly valuable as highest O3 levels are recorded in warmer periods all over the world. Considering the effect of SR on O3 formation, light and dark hours’ data were also modeled individually. $R^2$ values were calculated as 0.94 and 0.95 respectively for light and dark periods. Furthermore principle component analysis was performed in the study by using
annual data in order to determine the principle factors explaining $O_3$ formation. Neural network model run with principle factors has also provided a satisfying modeling efficiency ($R^2 = 0.87$).

As MLPs show the superiority of modeling non-linear relationships in presence of large sized data sets, they provide efficient results in modeling and predicting the ground level concentrations of $O_3$ in all relevant temporal scales. But we should demonstrate that the efficiency of the model is directly depended on the selection of appropriate functions (learning and transfer functions, number of hidden layers, epochs, neurons etc.) for the present data set.

Expanding this kind of studies will provide a significant contribution for improving early warning systems especially for industrialized regions where photochemical smog events are frequently noted.

Acknowledgement

This study was funded by the University of Kocaeli Research Fund under Project No. 2009-013.

References


