International Journal of Engineering & Technology, 7 (4.39) (2018) 139-141

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper



An Analytical and Predictive Approach of Statistical Technique for Air Pollutants

Vijay Kr. Yadav[1]*, P. Goyal[2], V.K. Yadav[3], Akansha Singh[4]

[1] G. L. Bajaj Institute of Technology & Management, INDIA
[2] Indian Institute of Technology Delhi, INDIA
[3] G. L. Bajaj Institute of Technology & Management, INDIA
[4] Sarvottam Institute of Technology & Management,
**Corresponding author E-mail: vijaybit10@gmail.com*

Abstract

Urban air pollution has emerged as an acute problem in recent years because of its detrimental effects on health and living conditions. The prediction of concentration of air pollutants in urban areas has become a major focus area of air quality research today due to their health effects. In the present study statistical model based on neural network (NN) has been developed to predict the pollutants such as NO_X, NO_2 and particulate matters ($PM_{2.5}$ and PM_{10}) for Delhi city at different locations such as ITO (Income tax office), and DTU (Delhi technological university) . Error estimation is also done in this study.

Keywords: *Error estimation, Meteorology, Neural Network, Prediction, Statistical technique.*

1. Introduction

The problem of air pollution has become so important, in urban areas that there is a need for timely information about changes in air pollution level. Air quality prediction is the advance information about the concentration level of air pollutants. It provides air pollution control needed to prevent the situation from becoming worse in the long run. The present work aims to develop a model for forecasting the concentration of NO_2, and $PM_{2.5}$ at ITO, and DTU of Delhi using neural network (NN). For this purpose, this model has been developed using the data of years 2007-2010 that include daily averaged concentration of pollutants, previous day concentration of pollutants and meteorology.
The various research studies exhibit that the performance of NN is generally superior in comparison to other traditional statistical methods, such as multiple regression, classification and regression trees and autoregressive models (Yi and Prybutok, 1996; Gardner and Dorling, 2000; Chaloulakou et al., 2003b). Related work and comparative study with fuzzy inference system (FIS) and their statistical error estimation had done our previous published work (P. Goyal et al., 2012).

2. Data Collection

The concentration of pollutants such as NO_X, NO_2, $PM_{2.5}$ and PM_{10} has been taken from website of Central Pollution Control Board (CPCB) Delhi. The 24 hourly averaged meteorological data wind speed, wind direction, sea level pressure, maximum temperature, minimum temperature, high dew point, low dew point, high humidity, average humidity, rain fall and visibility from Jan 2007 to Dec 2010 has been collected from Indian Meteorological Department (IMD), Delhi. This meteorological data will be used as an input in the present work.

3. Methodology

The NN model has been employed to predict the concentration of NO_2, and $PM_{2.5}$ using daily concentration data of pollutants, previous day concentration of pollutants and meteorology of the years 2007-2010. The input data of the model include daily concentration of pollutants (air quality index-AQI), previous day's pollutants concentration (previous day air quality index-PAQI) and meteorological variables i.e. wind speed, wind direction index (WDI), sea level pressure, average temperature, average dew point, average humidity, rain fall and visibility. Output data includes the daily averaged concentration of pollutants (NO_2, NOx, $PM_{2.5}$ and PM_{10}). The input and output is normalized between -1 to +1 using the minimum and maximum of the time series before any preprocessing. The data of the years 2007-2009 (75% of the total data) has been used for training and the data of the year 2010 (25% of total data) for validation of the model.

3.1. Neural Network (NN)

The artificial neural network represents an alternative methodology to conventional statistical prediction techniques because of their computational efficiency. The models based on NN are mathematical models inspired by the biological neurons (Gardner and Dorling, 1998). NNs are made up of interconnected processing elements called neurons or nodes that are arranged in the layers. These layers include an input layer, one or more hidden layers and an output layer which are connected to each neuron of the next layer by the weights. The number of hidden layer is selected

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

based upon the problem complexity. The total input signal (z) is calculated as:

$$z = \sum_i w_i x_i \quad (1)$$

The total incoming signal is then passed through a non-linear transfer function F to produce the outgoing signal F (z) of the node:

$$F(z) = \frac{1}{1+e^{-z}} \quad (2)$$

The output signal of a hidden node is finally passed to the nodes of the next layer (hidden or output), where a similar procedure takes place. There are several transfer functions available such as pure linear, hyperbolic tangent, sigmoid etc. Transfer function plays a key role in training process of neural network because the NN produce different results sensitive to its transfer function (Wassermann, 1989). The process of optimizing the connection weights is known as training or learning of NN. This is equivalent to the parameter estimation phase in the conventional statistical models. Iterative approaches are used to get the best values for connection weights by minimizing the performance function i.e. error between model output and the provided target values. The trained network is then used for the prediction of pollutants.
In this study, fully connected feed forward neural networks where nodes in one layer are only connected to corresponding nodes in the next layer, have been used. In the model, 9 input nodes are introduced the 9 input parameters. After several experiments it is found that 5 neurons in a hidden layer give the best architecture for the neural network model formed in all the seasons. Thus one hidden layer with 5 hidden neurons has been introduced in between input and output layers Fig.1.

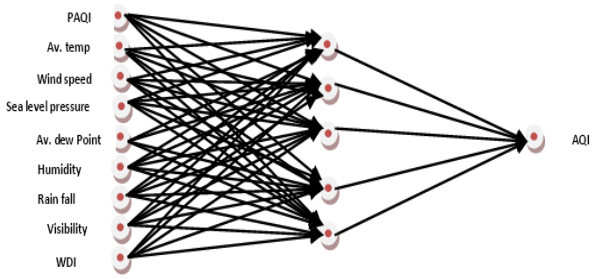


Fig. 1: Architecture of NN to predict pollutants concentration

The process of optimizing the weights in between the neurons of the different layers is known as training or learning process of NN. Firstly, all the connecting weights are initialized by the random numbers. The weights of a network are iteratively modified to minimize the total mean squared error between the desired target and actual output values. Back - propagation learning algorithm used for optimizations of weights.

3.2. Learning Algorithms

The back-propagation algorithm is used for training feed forward neural network. The algorithm uses gradient descent procedure to locate the absolute (or global) minimum of the errors surface (Gardner and Dorling, 1998). In back-propagation, there are two steps in its learning process, first is to propagate the input pattern through the network and other is to adopt the output by changing the weights in the network. It is the error signals that are back-propagated in the network operation to hidden layer. The error in the output layers is used as a basis for adjustment of connection weights between the input and the hidden layers (Boznar et al., 1993). The input signals passing through non- linear activation function given in equation (2).

3.3. Training and Validation of NN

When the learning process has been completed, the performance of the trained network has been validated with an independent dataset of the year 2010. It is important to note that this data should not have been used as a part of the training process. The accuracy of the model has been estimated with the statistical measures.

3.4. Result and Discussion

The model predicts daily averaged concentration of NO_2, and $PM_{2.5}$ has been developed for the whole year at different locations of Delhi viz., ITO, and DTU. Previous day air quality index , air quality index and meteorological data of the years 2007-2009 has been used for training the model, data of year 2010 has been used for validation of the developed model. The models have been developed on MATLAB 13. The models' behavior in both, development (training) and validation steps has been evaluated with the statistical measure viz., root mean square error (RMSE), and mean absolute error (MAE). The results of all locations have been discussed in following points.

3.4.1. Error Estimation

The error estimation, which has been used for estimation of statistical errors of the performance of models, has been given as follows.

Root Mean Square Error

Root mean square error (rmse) is a measure of the differences between values estimated by a model and the observed values and is expressed as:

$$rmse = \sqrt{\frac{\sum (C_{pre} - C_{obs})^2}{n}} \quad (3)$$

Mean Absolute Error

Mean absolute error (mae) is a quantity used to measure how close forecasts or predictions are to observations and calculated by following formula.

$$mae = \frac{\sum |C_{pre} - C_{obs}|}{n} \quad (4)$$

Where: C_{pre} = predicated value, C_{obs} = observed value, and n = number of days.
The figure shows that performance function in training and validation are close in magnitudes implying that the model is optimally trained and not over-trained. Fig. 2, and Fig. 3 represents the plot of observed and model's forecasted results of NO_2 and $PM_{2.5}$ at ITO respectively, have been given in values for the training dataset which reveals that the trend has been well captured by the model. The next step after training the model is to validate it in order to test its performance which has been done using daily data for the year 2010 (total data points 365).
The concentration of NO_2 and $PM_{2.5}$ predicted by the model at ITO has maximum as 288 µg/m^3 (342th day) and 387 µg/m^3 (358th day) and minimum as 107 µg/m^3 (272th day) and 51 µg/m^3 (246th day) respectively. The values of mae, and rmse in validation process are found as 72.96, and 55.67 respectively. The above discussion supports that model is performing conveniently and can be used for forecasting of pollutants concentration.

Table1: Statistical errors of NN models

Statistical Errors	Training data	Validation data
mae	39.36	72.96
rmse	59.78	55.67

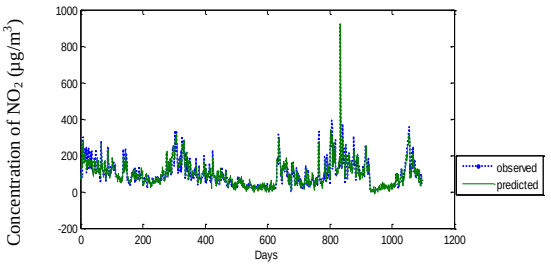


Fig.2: NN model predicted values of NO_2 in 2007- 2009 at ITO

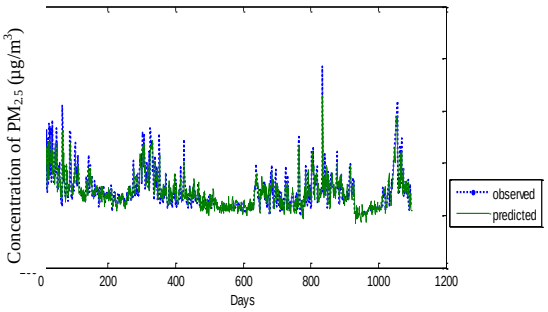


Fig.3. NN model predicted values of $PM_{2.5}$ in 2007- 2009 at ITO

Fig. 4, and Fig.5 represents the plot of observed and model's predicted results of NO_2 and PM_{10} at DTU respectively, have been given in values for the training dataset which explained that the trend has been well captured by the model.

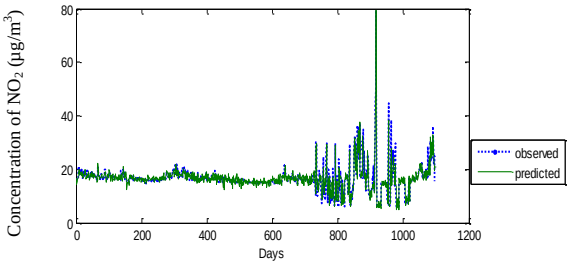


Fig. 4: NN model predicted values of NO_2 in 2007- 2009 at DTU

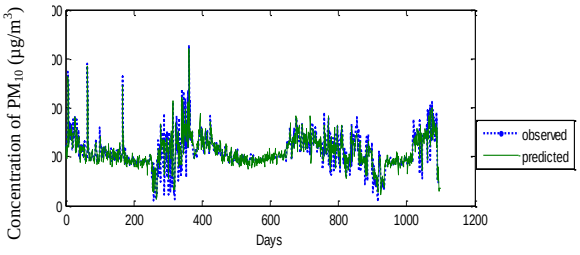


Fig. 5: NN model predicted values of PM_{10} in 2007- 2009 at DTU

4. Conclusions

A predictive approach based on NN with back propagation algorithm has been used in the present study. These networks have been used almost exclusively for the prediction and forecasting studies of air pollutants concentration. A three layer network architecture (9:5:1) has been obtained after detailed analysis of the data. The first layer consists of 9 input neurons corresponding to each input variable. One output node is introduced in the third layer i.e. daily concentration of NO2, PM_{10} and $PM_{2.5}$ at different location of Delhi viz., ITO, and DTU. Between the input and output layer, one hidden layer has been introduced with 5 hidden neurons. The models behaviour in both i.e. development (training) and testing steps has been evaluated with the help of statistical measures (rmse), which shows a consistency between predictions and real observations.

Acknowledgement

A wide portion of this study was carried out at IIT Delhi during my stay, and rest of portion of this study was equally done by all of us at G. L. Bajaj Institute of Technology and Management, India.

References

[1] Boznar, M., Lesjak, M., and Mlakar, P., "A neural network based method for short term predictions of ambient SO2 concentrations in highly polluted industrial areas of complex terrain", *Atmos. Environ.*,1993, 27B:221–230.

[2] Chaloulakou A, Grivas G, Spyrellis N., "Neural network and multiple regression models for PM10 prediction in Athens A comparative assessment", *J Air Waste Manage Assoc*, 2003b; 53: 1183–90.

[3] Gardner, M. W., Dorling, S. R., "Artificial Neural Networks (The Multilayer Perceptron) - a Review of Applications in the Atmospheric Sciences", *Atmospheric Environment 32*, 1998, 2627-2636.

[4] Gardner, M. W., Dorling, S. R, "Neural network modeling and prediction of hourly NO_x and NO2 concentration in urban air in London", *Atmospheric Environment 33*, 1999,709 – 719.

[5] Gardner, M. W., Dorling, S. R, "Statistical surface ozone models: an improved methodology to account for non-linear behavior", *Atmospheric Environment 34*, 2000, 21-34.

[6] Goyal P., Kumar A., Yadav Kr. Vijay., "Forecasting of air pollutants in Delhi using different statistical techniques", *Indian Journal of air Pollution Control*, Vol. xii, no.2, 2012, p57-66.

[7] Wassermann R.,Tao T.Y., Whitesides M., "Structure and reactivity of alkyalsiloxane monolayers formed by reaction of alkyltrichlorosilanes on silicon subtracts", *Langmuir*, 1989, pp1074-1087.

[8] Yi J., Prybutok V.R., "A neural network model forecasting for prediction of daily maximum ozone concentration in an industrialized urban area", E*nvironmental pollution 92*,1996, 349-357.