Original Research

IJCRR

IMPROVED ARTIFICIAL NEURAL NETWORK PERFORMANCE ON SURFACE OZONE PREDICTION USING PRINCIPAL COMPONENT ANALYSIS

K. Padma, R. Samuel Selvaraj, S. Arputharaj, B. Milton Boaz

Department of Physics, Presidency College, Chennai, 6005, Tamil Nadu, India

ABSTRACT

Correlated sample data normally creates confusion over ANN (Artificial Neural network) during the learning process. In this work the Principal Component Analysis (PCA) method is used for elimination of correlated terms in data. After application of PCA, the uncorrelated input data were used to train a Multi-Layer Perceptron (MLP) ANN system. The results revealed that the elimination of correlated information by using the PCA method improved the ANN estimation performance. we measured the surface ozone and its influencing factors during the period June 2011- September 2012 at Chennai, a tropical site on the Southeast coast of India situated at 13 04'N 80 17'E. The input data that was used for building the network are the wind speed, temperature, relative humidity, UV radiation had been used in neural networks for the prediction of daily surface Ozone 24 hours in advance.

Key Words: Multi-layer perceptron, Artificial neural network, Principal component analysis, activation function, hidden layers, 'PC variance'

INTRODUCTION

Continuous development of technology and on increasing population in metropolitan cities like Chennai, Delhi, etc., a series of severe problems related to environmental pollution such as air pollution, noise pollution, waste and sewage disposal have attracted much attention than ever before. Among these, air pollution has direct impact on human health through exposure to pollutants at high concentration level existing in ambient areas. Air pollution control is very important to prevent this situation from worsening in the long run. On the other hand, short-term forecasting of air quality is needed in order to take preventive action during episodes of airborne pollution (Wenjian wang, 2002). Ozone is a secondary pollutant and it is not usually emanated straight forwardly from stacks. Process in the formation of Ozone (O_3) is highly multifaceted in nature. The ozone precursors are generally divided into two groups, namely oxides of nitrogen (NOX) and Volatile Organic Components (VOC) like evaporative solvents and other hydrocarbons. In suitable ambient meteorological condition (e.g. warm, sunny, clear day) ultra violet radiation (UV) causes the precursors to interact photo chemically in a set of reactions that result in the formation of ozone (Bandyodhyay et al., 2007). Ozone is a green house as well as secondary air pollutant, interaction between Ozone and climate occurs not only in the Stratosphere but also at the earth's surface. Ozone comes in contacts with life form of the earth's surface and shown it destructive nature. It damages the leaves and affects plant growth thus reduces crop yield and causes noticeable foliage damage. Human health is also affected by high concentration of Ozone (Londhe, et al. 2008. Stathopoulou, et al., 2008).

Therefore, the development of effective prediction models of ozone concentrations in urban areas is important. Management of control and public warning strategies for ozone levels (particularly densely populate area) requires accurate forecasts of the concentration of ambient ozone (Prybutok, 2000). So monitoring daily Ozone level in this big city is important for today and future research whether or not a threshold is exceeded. Such information could be exploited by environmental and

Corresponding Author:
K. Padma, No. G 8. Sam's Avenue Appartments, Valluvar salai, Arumbakkam, Chennai 600 106, Tamil nadu(st), India; Mobile: 91-9790766412; E-mail: padma_manmalai@rediffmail.com

Received: 12.06.2014 **Revised:** 10.07.2014 **Accepted:** 02.08.2014

medical authorities to announce public health warning (Stathopoulo et. al. 2008).

ANN-based method still need to improve in order to achieve good prediction performance as effectively and efficiently as expected. In fact, a number of difficulties have been associated with ANN use which hampered their effectiveness, efficiency as well as general acceptability in air quality analysis. These difficulties include susceptibility to chaotic behavior, computationally expensive training, training set problem and topology specification problem etc (Wenjian wang, 2002). In the other training method, meteorological data was preprocessed by the Principal Component Analysis (PCA) method. After applying the PCA, the size of the input sample was reduced from 4 to 3 orthogonal, uncorrelated components. Network's architecture was composed of input layer, one hidden layer and output layer. Standard back propagation with momentum was used for training. In both cases the input data was normalized and scaled to the range (-1, 1). The sigmoidal activation (transfer) function was used in our network of all neurons except those in the input layer.

STUDY AREA

Chennai is situated on the south east coast of India and north east coast of Tamil Nadu. This area is one of the most highly populated urban sites. Chennai lies on the thermal equator and is also a coastal. The latitude and longitude of the center of the city are E80° 14'51" and N13° 03' 40". The geographical location of the experimental site is shown in Fig. 1 and it is located in south Chennai. The different sources of air pollution are classified under the following categories Transport, Industries, Residential in Chennai City. This Urban City can be divided into four areas, North, Central, South and West. The Northern part is primarily an industrial area comprising of petrochemical industries in the Manali area and other general industries in Ambattur. (Fig. 1). Chennai has many industrial areas. This study was conducted at Koyembedu which houses Chennai's moffussil Bus terminus and 100's of Buses and other vehicles ply daily and hence the vehicular emission is very high. This site is surrounded by the number of Industrial areas located within a short radius.

Surface ozone was measured throughout Tamil Nadu during the year 2011 and it was found that Kanniyakumari district had the highest daily average of 17.8 ppbv (Samuel Selvaraj et. al., 2011). Moreover the surface ozone levels studied in Chennai during 2004- 2005 it was found that the hourly values varied from 1 ppbv to 50.27 ppbv (Pulikesi et al., 2005). In the urban area Delhi ozone concentration in the ambient air varied from 9 to 128 ppbv at four different sites during 1989–1990(Varshney et. al., 1992). So these studies have indicated that the effects of O_3 on vegetation were quite severe in India and other parts of Asia. (Emberson D. et al., 2001). From our knowledge of literature survey, there was no measurements have been carried out over Chennai metropolitan area in recent years (Samuel Selvaraj et. al., 2013). Hence through this study, the surface ozone (O_3) concentration was measured in this urban site, Chennai.

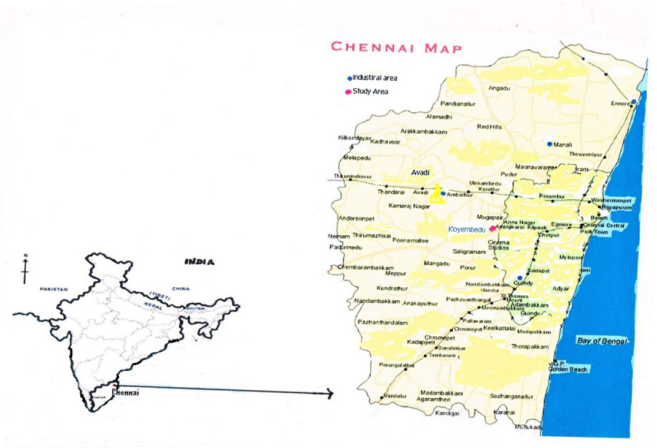

Figure1: Measurement site. (Samuel selvaraj, 2013)

DATA AND METHODOLOGY

The measurement were carried out in the area which has selected to represent the typical residential with high commercial and traffic influenced. Using Aeroqual 200 series Ozone data had measured.(Akram Ali., Sulee et al., Michael Frei et al., Dovile Laurinaviene, 2008). UV irradiation had been measured by UV light meter (UV-3450A series). Surface ozone measurements were carried out daily and ten measurements were made on all days between 08.00 hrs and 17.00 hrs (IST) during the period from June 2011 to September 2012. Furthermore, wind speed, temperature, relative humidity and UV radiation were also measured simultaneously. Here two approaches are applied to predict the surface ozone which are Artificial Neural network without PCA and Artificial Neural network with PCA. The total input data set contains four variables namely wind speed, temperature, UV radiation and relative humidity. The data set is divided into two distinct sets called training and testing sets. The training set is the larger set (90%) used for the network to learn pattern presence in the data and the testing set (10%) were used to evaluate generalization ability of supposedly trained network in both (ANN with PCA , ANN without PCA) method.

ARTIFICIAL NEURAL NETWORK WITHOUT PCA

ANN is based on principle stating that a system of highly interconnected simple processing elements can learn

complex interrelationships between independent and dependent variables. Yi and Prybutok presented a feed–forward neural network model for predicting ozone concentrations in an urban area. They recognized other precursors of ozone formation must be included in order to improve the prediction of their model (Elkamel et al., 2000). The particular aim is to relate the surface ozone concentration to meteorological variables. A total of 4 variables are used in preparing this model for the prediction of surface ozone concentrations.

ANNs are constructed with many layers so as to be called as multilayer ANNs. First one is input layer has independent variable in statistical literature. Last layer is output layer has contains dependent or response variables. All other unit in this network is called hidden layers. There are two functions governing the behavior of layers. The input function, and The output/activation function. A number of nonlinear functions have been used in the literature as activation functions. However, most common choice is sigmoid function (Andrew C. Comrie, 1999) . The main task of the activation function is to map the outlying values of the obtained neural input back to a bounded interval such as [0,1] or[-1,1]. The sigmoid function has some advantages, due to its differentiability within the context of finding a steepest descent gradient for the back propagation method and moreover maps a wide domain of values into the interval [0,1] (Girish Kumar). In this study we have selected the feed-forward back propagation Multi-Layer Perceptron (MLP) to develop the ANN model with changeable neurons in the hidden layer to get good result with accuracy (Elamparai, 2011). The simulation is carried out in Mat lab using the 'Levenberg Marquardt back propagation' training algorithm.

The performance of an ANN very much depends on its generalization capability, which in turn is dependent upon the data representation. A set of data presented to an ANN consist of correlated information. This correlated data reduce the distinctiveness of data representation and thus, introduce confusion to the ANN model during the learning process and hence, producing one that has low generalization capability to resolve unseen data. This suggests a need for eliminating correlation in the sample data before they are being presented to an ANN. This can be achieved by applying the Principal Component Analysis (PCA) technique onto input data sets prior to the ANN training process as well as interpretation stage. This is the technique examined in this research. The PCA technique was first introduced by Karl Pearson in 1901, but he did not propose the practical calculation method for two or more variables, which were useful for various applications (Junitha, 2008).

ARTIFICIAL NEURAL NETWORK WITH PCA

PCA is a variable compression technique. It transforms a large number of interrelated variables to a new set of uncorrelated PCs which are linear combinations of the original variables (Jolliffe,. 1986). Therefore, each principal component contains information on all meteorological variables. PCA generates the same number of meteorological indices as the original meteorological variable (including both original and quadratic terms) and orders them by the magnitude of variances. In order to reduce the number of predictor variables, the rule of thumb for most previous related studies is to take only the first several PCs with Eigen value greater than or equal to one as predictor variables. The use of the PCA function involves specifying a fraction value corresponding to the desired percentage of the least contribution of the input components. For example, a fraction value of 0.02 means that the input components which contribute less than 2% of the total variation in the data set will be discarded. From this point onwards, this fraction value will simply be referred to as the "PC variance".

The PCA method is potentially very well suited for neural networks training methods. Training is more effective when performed on uncorrelated and orthogonal data. Moreover the network is smaller the faster the training and in several cases the better generalization properties. The PCA transformation is based on the following autocorrelation matrix:

$$R_{xx} = \frac{1}{n}\sum_{k=1}^{n}\left(x_k x_k^T\right) \qquad (1)$$

Where n is the number of vectors in the input set, x_k is the k-th vector. Eigenvectors of matrix R_{xx} corresponding to eigenvalues sorted in the decreasing order point out the principal components. The first principal component is responsible for the highest percentage of the variance of the sample the second one – for the next highest variance, and so on (Joliffe I.T., 1986). By choosing the required number of principal components one is able to build matrix W of the PCA transformation:

$$W = [w1,w2,...,w_M]T,$$

Where M is the required number of components and wk (for k=1,…,M) are the principal components itself. Two types of PCA data processors had been implemented for the purpose. The first one is called the PCA pre-processor, which is responsible for pre-processing raw data, to eliminate correlation in the training samples. The second is called PCA postprocessor, used to transform the validation and test datasets according to their principal components. The implementation and simulation were car-

ried out with the aid of built-in functions supported by MATLAB Neural Network Toolbox (Junitha, 2008). Each MLP's performance was calculated based on the Mean Absolute Error (MAPE) .

RESULT AND DISCUSSION

The data were then divided into three datasets; the training, validation and test. The training set was used to train the ANN the validation set was used for early-stopping of the training process and the test set was used to evaluate the ANN performance after completion of the training process. In case of using the back propagation algorithm in ANN without PCA the average Root mean square error was equal to 1.4 ppbv. The respective average mean absolute percentage errors within the test set - was equal to10.5% (see Table 1).

$$\mathrm{RMSE} = \frac{1}{n}\sum_{n=1}^{n}(Ri - Pi)^2 \qquad (2)$$

$$\mathrm{MAPE} = \frac{1}{n}\sum_{n=1}^{n}\frac{|R_i - P_i|}{|R_i|}\times 100 \qquad (3)$$

Here 'n' is the total number of observed value and R_i, P_i are i^{th} real and predicted values respectively. Where RMSE, is root mean square error and MAPE is Mean absolute percentage error and the correlation ratio between predicted and real changes were calculated from the below equation.

$$\delta = \frac{\sum_{i=0}^{n}(Ri^t - Rt^{t-1})(Pi^t - Rt^{t-1})}{\sqrt{\sum_{i=0}^{n}(Ri^t - Rt^{t-1})((Pi^t - Rt^{t-1})^2)}} \qquad (4)$$

N denotes the size of the test set, Pi t and Ri are the i-th predicted, real values at time t, respectively. The experiment can be repeated several times with vary the number of neurons in the hidden layer. The best result of ANN without PCA given below and the correlation coefficient δ was equal to 0.5.

Table 1: Results of ANN without PCA algorithm.

Error	Value(ppb)	Value%
Average	1.4	10.5
Minimum	0.1	2
Maximum	3.8	14.5

In case the PCA was initially used for pre-processing the data (and reducing the input dimension from 4 to 3). The average Root mean square error was equal to 1.15 ppbv. The correlation ratio δ was equal to 0.619 which is better than plain back propagation ANN network and % of error is 8.2 which are lesser than ANN without PCA (see Table 2).

Table 2: Results of ANN with PCA algorithm.

Error	Value (ppb)	Value%
Average	1.1	8.2
Minimum	0.02	1.0
Maximum	2.8	13

Closer look at the principal components defined in our experiment – in case of no data compression, i.e. with all 4 principal components - revealed some interesting properties of the way the input factors were "combined" into components. That is most of the input variables are closely related to another variable. First, the sums of absolute values of all weights incoming to every principal component (from the input vector) were calculated. Intuitively, these sums should be greater for the first few components (the most relevant ones) than for the less significant components. It can be clearly seen that the rough estimation of the relative importance of the components presented in Fig. 2. It can be also seen from the figure the last input component which has low weight value relatively low contribution was discarded.

Fig. 3. shows the best result of predicted values of surface ozone using ANN model without PCA with 6 neuron used in the hidden layer in neural network is compared with the observed surface ozone data.

The best achieved result of predicted values of surface ozone using ANN model with PCA and 6 neuron which is

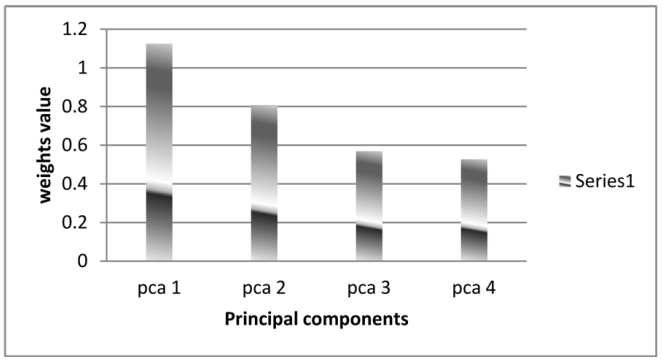

Figure 2: Sum of absolute values of all input weights to principal components PC 01,, PC 05.

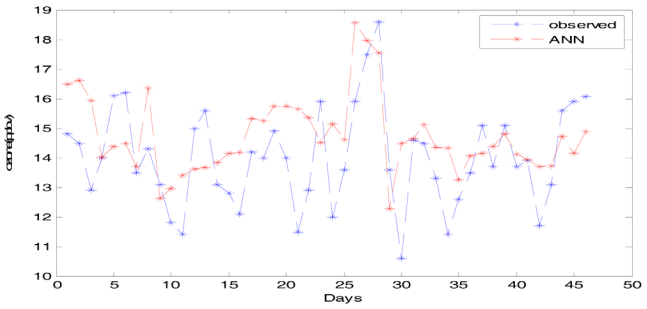

Figure 3: Predicted and observed values of surface ozone using the back propagation in ANN algorithm without PCA .

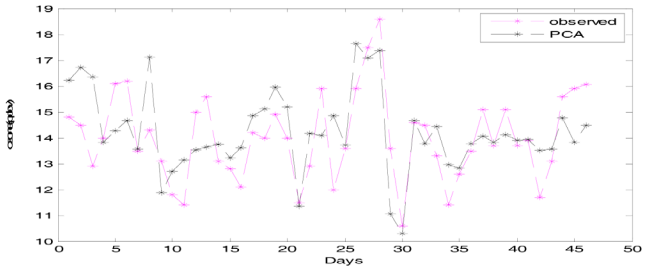

Figure 4: Predicted and observed values of surface ozone using the back propagation in ANN algorithm with PCA .

used in the hidden layer in neural network is compared with the observed surface ozone data. This ANN model with PCA provides a very good prediction of daily surface ozone in one day advance as shown in fig.4.

CONCLUTION

In this work, Mat lab tools are used to predict the daily surface ozone data in an one day advance. The 400 days surface ozone data used for training the network while 45 day data are used for testing the network. The result obtained with average error of 8.2% (1.1ppbv) and correlations between predicted and real changes is very encouraging and provide a further exploration of the issue that is combine our approach with expert systems. One of the important conclusions of this research is applicability of the PCA method as a supporting tool for data pre – processing in the problem considered this issue deserve further investigation and reduces the complexity of network and provide good result.

AKNOWLEDGEMENT

The authors wish to thank the Tamil Nadu Pollution Control Board, Chennai for providing valuable guidance and for checking the accuracy of the instruments. We also thank our co-researchers who have given their valuable suggestions and guidance enabling us to release this research paper and their help is gratefully acknowledged. Authors acknowledge the immense help received from the scholars whose articles are cited and included in references of this manuscript. The authors are also grateful to authors / editors /publishers of all those articles, journals and books from where the literature for this article has been reviewed and discussed.

REFERENCES

1. Akram Ali, Factors affecting on response of Broad bean and corn to air quality and soil CO_2 flux rates in Egypt. Water Air soil Pollute.195, 311-323, 2008.
2. Andrew C. Comrie. Comparing neural network and regression models for Ozone forecasting. Air waste Management. Assoc. 47,653-663, 1997,.
3. Bandyopathyay.G , S.Chattopadhyay,. Single hidden artificial neural network models versus multiple linear regression model forecasting the time series ozone. Int. J. Environ. Scxi. Tec., 4(1).,141-149,. 2007
4. Dovile Laurinaviene., Ground level Ozone Air pollution in Vilnius City. Environmental Researech, Engineering and Management, 2008, No. 3(49), 21-28.
5. Elkamel.A, S.Abdul-Wahab., W. Bouhamra, E. Alper., Measurement and prediction of Ozone levels around heaviloy industrialized area. A neural network approach.Adance in Environment Research 5, 47-59, 2000.
6. Elampari. K, Chithambarthanu T. Diurnal and seasonal variations in surface ozone levels at tropical Semi - Urban site, Nagercoil, India, and relationships with meteorological conditions. International Journal of Science and Technology, 2011, volume1, No.2.
7. Emberson. D., M.R. Ashmore, F. Murray, J.C.I. Kuylenstierna, K.E. Percy, T. Izuta, Y. Zheng, H. Shimizu, B.H. Sheu, C.P. Liu, M. Agrawal, A. Wahid, N.M. Abdel-Latif, M. van Tienhoven, L.I. de Bauer, M. Domingos, Impacts of air pollutants on vegetation in developing countries,Water Air Soil Pollut. 2001, 107–118.
8. Girish Kumar Jha ., Artificial Neural Networks., Indian Agriculturral Research Institute.Pusa , New Delhi
9. Joliffe. I.T, Principal component Analysis, Springer Verlag,1986, 533 – 536.
10. Junitha Mohamad-Saleh, Brian S. Hoyle., Improved Neural Networ performance using principal component analysis on matlab. International journal of the computer,vol.16. No.2. pp 1-8, 2008.
11. Lodhe A.L, D.B. Jadhav,P.S. Buchunde and M.J.Kartha. Surface Ozone variablility in the urban and nearby rural locations of tropical India. Currennt science, Volumw 95, No.12, 25. A13. 2008).
12. Michael Frei, Juan Pariasca Tanaka and Matthias Wissuwa. Genotypic variation in tolerance to elevated ozone in rice; dissection of distinct genetic factors linked to tolerance mechanisms, Journal of Experimental Botany, 2008, 13, 3741-3752.

13. Pulikesi M.,P. Baskaralingam, Ramamurthi.V., Sivanesan. Studies on surface ozone in Chennai. Research journal of chemistryand environment, 2005, Vol. 9., issue.
14. Samuel Selvaraj R., Milton Boaz, B., Sachithananthem C.P., Padma, K., Steephen Rajkumar S. Inbanathan., Kanmani RajaselviG. Indira and VImalpriya S.P., Measurement of surface ozone in the year 2011 at different sites over Tamil Nadu, India. Indian Journal of science and Technology, 2011, vol. 5, No. 2.
15. Samuel selvaraj R., Padma K, Miltoin Boaz B, Seasonal variation of surface ozone and its association with meteorological parameters, UV radiation, rainfall, cloud cover, over Chennai, India. Current science, vol. 105, no. 5, 10 september 2013.
16. Su Lee and shih-Wei Tsai, Passive sampling of ambient ozone by solid phase micro extraction with on – fiber derivatization, Analytica Chimica Acta, 2008, Vol.610, Issue 2, 149-155.
17. Varshney C.K., M. Aggarwal, Ozone pollution in the urban atmosphere of Delhi, Atmos. Environ. 1992, 26B, 3, 291–294.
18. Wenjian Wang, Zongben Xu, Jane Weizhen Lu,. Three improved neural network models for air quality forecasting. Engineering computations, vol. 20, No. 2, 2003.