COMPARISON OF MLR AND ANN MODELS FOR PREDICTING FOG FORMATION CRITERIA IN DELHI, INDIA

Dhirendra Mishra*

Pramila Goyal*

Abstract

Delhi is facing most frequent fog conditions during winter months (November, December, January and February), which might be resulted due to increased concentrations of air pollutants from various sources. In the months of December and January, fog persists for many hours from early evening to late morning. The temperature during these months was reaching at its minimum level with the frequency of western disturbances, causing rainfall over this region. These are the favorable conditions for fog formation in Delhi. In this study principal component analysis (PCA) has been analyzed to find the correlations of different predictor variables (meteorological and pollutants) with visibility. The hourly meteorological variables and concentrations of pollutants were building together to propose the reliable physical multiple linear regressions (MLR) model for forecasting of fog formation in terms of visibility over Delhi. Further, the nature of the internal nonlinear function has been trained and extracted the response of individual predictor variables for forecasting of visibility through artificial neural network (ANN) model. These models have been trained and validated during winter hours of the year 2011-12. The proposed models illustrate how visibility is identifying the key meteorological variables and air pollution levels in Delhi.

KEYWORDS: Fog, Visibility, PCA, MLR, ANN, Meteorological Variables, Delhi.

* Centre for Atmospheric Sciences, Indian Institute of Technology Delhi, Hauz Khas,

New Delhi-110016, India

1. Introduction

Fog is defined as a cloud, formed due to the liquid water droplets or ice crystals suspended in the air at or near the Earth's surface, which reduces the horizontal visibility less than 1km. Light from a distant object viewed by an observer is scattered (by air pollutants) out of the line of sight between the object and observed. This may reduce the brightness contrast between the object and its background. Thus, visibility can be defined as the maximum distance an object or details of a complex pattern can be seen. Occurrence of fogs during late night or early morning over the northern plains of India are common in winter. Meteorological phenomena such as visibility, temperature, precipitation, wind speed or relative humidity can affect the fog formation criteria and reduces visibility in the atmosphere (Goyal et al., 2013). Anthropogenic emissions, e.g. vehicular combustion, construction, mining, agriculture etc. also are causing the degradation in visibility. It has long been known that atmospheric pollution can cause a decrease in the atmospheric visibility and defined as the maximum distance at which the outlines of a target can be recognized against the horizon as background (Horvath, 1995). Indoor air pollution has also been found to be associated with a wide range of adverse health and visibility. Environmental studies have found that air pollutants in the air can scatter or absorb light, therefore reduce the visibility e.g. fine particles ($PM_{2.5}$, particulate matter with aerodynamic diameter less than 2.5 μm) can be effective light scatters because the wavelength of visible light falls in this range. Atmospheric pollutants emitted from the anthropogenic sources e.g. vehicles, industries, power-plants, construction, mining, agriculture etc. can cause the decrease of atmospheric visibility under the favorable meteorological conditions. Ambient aerosols, especially fine particles played a dominant role in the visibility reduction in different regions (Chan et al., 1997). However, the relationship between aerosol concentration and visibility was generally nonlinear (Deng et al., 2008;). When the aerosol concentration was high, the change of the aerosol level had little influence in the visibility and when the aerosol concentration was at a low level, visibility was greatly sensitive to the change of the aerosol level. The pollution concentration should be firstly reduced to a certain degree and then further reduction can lead to a significant increase in visibility (Goyal et al., 2013). Actually, wind roses are an information packed plot providing frequencies of wind direction and wind speed. A wind rose can quickly indicate the dominant wind directions and the direction of strongest wind speeds. High concentrations are observed in light to moderate wind conditions. Due to this nonlinear feature, the improvement of visibility for the Delhi mega city is complicated and needs comprehensive analysis.

Air pollution that reduces visibility is often called haze or smog. The term smog originally meant a mixture of smoke and fog in the air, but it refers the mixture with air pollutants that can be seen. Smog typically starts in cities or areas with many people and it travels with the wind, it can appear in rural areas also. Delhi, with a population of approximately 22 million, presents a unique study region because of its distinguishing temperature extremes and relatively moderate to high levels of air pollutions are favoring smog formation during winter months. Air pollution is an escalating problem in Delhi fuelled by increasing anthropogenic activities, hasty development, rapid industrialization, transportation, superfluous use of fossil fuel consumption, increasing global power needs etc. The cancellation of hundreds of flights at airports around the world are not caused by the effect of poor visibility when planes they are in the air, but instead when they reach the ground where thick layers of fog observed. A similar situation has been observed at Indira Gandhi International (IGI) Airport, Delhi that during winter months. The climate of Delhi becomes very hot during the month of June, which is followed by monsoon happening somewhere after September. The average temperature of Delhi during summer ranges from 25ºC to 40ºC. May and June are the hottest months of the year. This period marks the expansion of the atmosphere, causing the elevation of mixing height, which gives pollutants more volume to disperse in. Hence it can be said that pollution levels may show lower concentrations and was resulted good visible range. The arrival of monsoon does bring some relief in temperatures, starting from the end of June and lasting till September. Consecutively the SW Monsoon and the retreating NE Monsoon wash out the air of pollutants and gives it the cleanest character with maximum ranges of visibility. The end of monsoon marks the arrival of a transition season. Post monsoon arrives by early or mid October, and is marked by very dry ambiance, warm days and pleasant nights. Maximum temperatures drop below 30°C in late October and there is a gradual fall in average temperature. Minimum temperature drops below 20°C. In comparison to summer, winters are short, which starts from the end of November and continues till February. Temperatures fall substantially down to as low as 3º C to 4º C at the peak during the winter months. These are the favorable conditions for fog formation in Delhi and dense fog enveloping the city during Januray months reduced the visibility consequently. This is the period where the mixing height falls giving a reduced volume of the pollutants to diffuse in the air and show remarkable rise in pollution levels with decrement the visible range. Therefore, visibility was observed best in the summer and worse in the winter than in post-monsoon, probably because of coal burning during the heating period in Delhi. Moreover, the diverse meteorological variables are also influences our environment in to a great extent. Hence a system of continuously knowing the presence of such pollutants in the atmosphere is invariably important in order to check the fog formation criteria in Delhi. Therefore, in this study the influences of air pollutants and meteorological parameters for fog formation in Delhi, India has been studied.

2. Material and methods

2.1. Study area Data

Fog or poor visibility created a nuisance for air travelers at Indira Gandhi International (IGI) Airport, Delhi as it is known as an important gateway to India. It is situated in the capital city of India. The hourly data were collected during foggy months (November, December, January and February 2011-12) at IGI Airport, Delhi to study the fog formation criteria. The hourly air pollutant data (Nitrogen-oxide (NO), Ozone (O_3), Nitrogen dioxide (NO_2), sulfur dioxide (SO_2), carbon monoxide (CO), particulate matter (PM_{10})) and meteorological variables (wind speed (WS), visibility distance (VS), air temperature (Temp), dew point temperature (DT), pressure (PRES), relative humidity (RH) and wind direction index (WDI)) were collected at IGI Airport, Delhi for the present study. Due to the circular nature of wind directions, wind direction index has been used in this study. WDI is defined as;

$$WDI = 1 + \sin(\varphi - \pi/4) \qquad (1)$$

where φ is the wind direction in radians. The collected meteorological and air pollutant data were analyzed based on the continuity and reliability of the available records. The data has been collected from Central Pollution and Control Board (CPCB), Indian Meteorological Department (IMD), New Delhi and Wyoming Weather Web data archive. The data sources included the air quality monitoring stations at IGI Airport, Delhi that is part of Central Pollution Control Board (CPCB), New Delhi. Fig. 1 shows the study locations i.e. IGI Airport in Delhi.

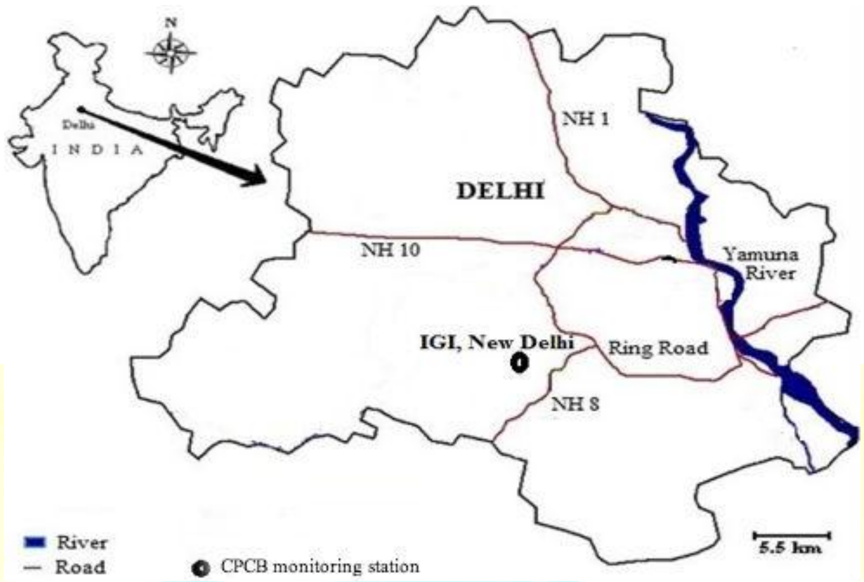


Fig. 1: Map of Delhi with monitoring station (IGI Airport)

2.2. Principal Components Analysis (PCA)

The PCA is a multivariate statistical method widely used in air pollution analysis. The role of PCA is to reduce the number of predictor variables. These predictive variables are obtained from the independent linear combinations and retain the maximum possible variant of the same dataset. PCA is also reducing the collinearity of the datasets, which leads the good predictions of air pollutant concentration. A regression model was used by Kumar et al., 2011 for air pollution forecast in Delhi, megacity by an air quality index highly correlated with meteorological variables. The principal components have been computed using covariance of the input data matrix. Only those components, having eigenvalues ≥ 1, are used in the forecasting models using principal component regression technique. PCA is also a procedure to reduce the number of variables. It is useful when obtained data have a number of variables (possibly a large number of variables) and believed that there are some variables those are correlated with one another. The PCR was also used to forecast the long–range forecasting of Southwest monsoon rainfall over

India (Rajeevan et al., 2005). Thus, the PCA analysis has been used in the analysis for visibility conditions during winter months in Delhi.

2.3. Multiple Linear Regression (MLR) model

A forecast can be expressed as a function of a certain number of factors that determine its outcome. MLR technique includes one dependent variable to be predicted and two or more independent variables. In general, multiple linear regression can be expressed as in Equation (2):

$$Y= b_1 +b_2 X_2+.............+b_k X_k+ e \qquad (2)$$

where Y is the dependent variable, X_2, X_3..........,X_k are the independent variables, b_1, b_2..........,b_k are linear regression parameters. In this study, visibility is the dependent variable and air pollutant concentrations and meteorological variables are independent variables, e is an estimated error term which is obtained from independent random sampling from the normal distribution with mean zero and constant variance. The task of regression modeling is to estimate the b_1, b_2..........,b_k which can be done using minimum square error technique. The similar methodology has been done as Goyal et, al., 2013.

2.4. Artificial Neural Networks (ANN) model

Actually, ANN is a part of the machine learning technique which tries to simulate the way of learning about the human brain. Its function mimics biological neurons in which the structure consists of a group of artificial neurons, which are interconnected, creating networks. The main elements of neuron for learning or information processing are: inputs, weight, summation function, transformation function and output. Similar to the human brain, the process of training the networks need to recognize patterns, develop generalization and learn to improve the performance (Kumar et al., 2013). Technically, the process of ANN is briefly explained as follows. The neuron is the basic information processing unit of an ANN. It consists a set of links, describing the neuron inputs, with weights W_1, W_2, ..., W_m and an adder function (linear combiner) for computing the weighted sum of the inputs i.e. $u = \sum_{j=1}^{m} W_j X_j$. Finally, Activation function φ for limiting the amplitude of the neuron output i.e. $y = \varphi(u + b)$, where 'b' denotes bias. All the processes are depicted in Fig. 2.

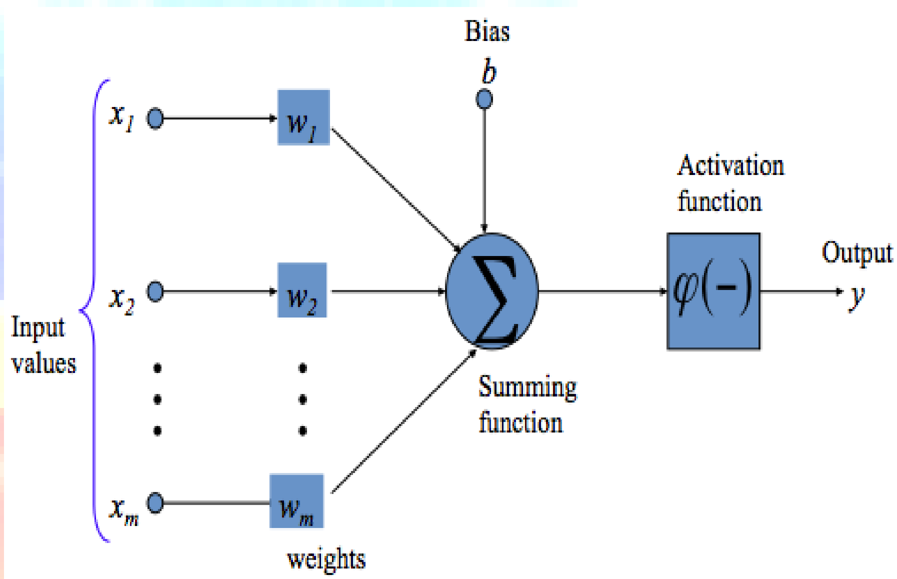
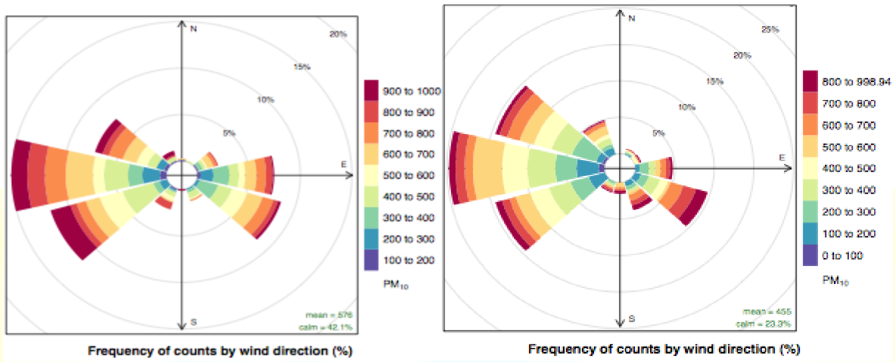


Fig. 2: A schematic model of ANN

The typology of most neural networks generally involves a collection of neurons that are configured in two or more layers. Therefore, the research combines some neurons into multilayer structures to have the power of pattern recognition and prediction. For this reason, this research employs a multilayer feed-forward networks, which are the most common type of neural network currently in use. The multi layer feed-forward network comprises of an input layer, hidden layer and output layer. Specifically, the input layer is a layer that is directly connected to outside information. All data in the input layer will be fed-forwarded to the hidden layer as the next layer. Meanwhile, the hidden layer functions as feature detectors of input signals and releases them to the output layer. Finally, the output layer is considered as a collector of the features detected and as a producer of the response. On the networks, the output from output layer is the function of the linear combination of hidden unit's activation; whereas the hidden unit's activation function is in the form of a non-linear function of the weighted sum of inputs.

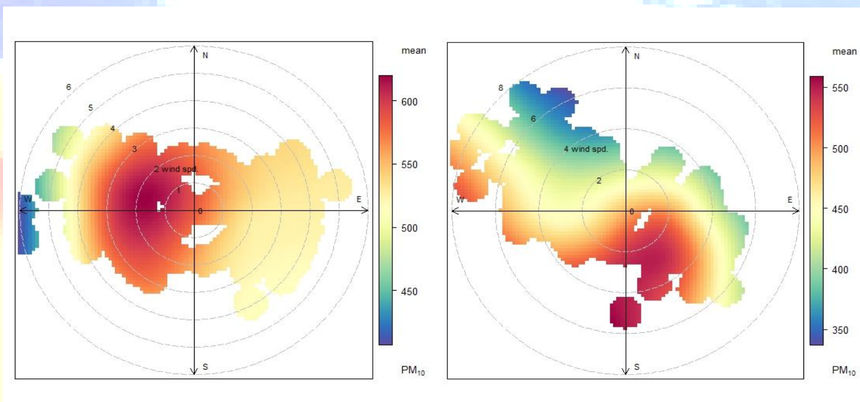
3. Results and Discussions

The fog formation conditions in terms of visibility distances have been studied during the winter months in Delhi. The hourly meteorological and air quality data sets for 2011-12 of winter months were used for training and validation of the forecasting ability of the model. The hourly fog conditions on an average 221 hours and dense fog during 87 hours in winter months of 2011-2012 (Nov, Dec, Jan and Feb) have been observed. The most vulnerable timing of occurrences of fog was 11 PM night until 10 AM the next morning. However, it also occurs at 8 PM in the evening and persists till 11 AM in the morning, which disturbed the flight operations at the Airport. The wind rose diagram for PM_{10} pollutant in December 2011 (Fig. 3 (a)) shows the prevailing winds were from the west and WSW with some significant frequencies also observed from east. High concentrations are observed in light to moderate wind conditions (concentration polar plots) that causes fog formation. The wind at IGI Airport in Januarary, 2012 blow from the west much of the time. In fact, the 3 spokes around the west direction (WNW, W and WSW) comprise more than 50% of all hourly wind direction. The pollution rose diagram indicates prevailing winds are from the east with significant frequencies also observed from north-west and southeast during the study period. Fig. 4 shows the concentration polar plots, which revolving the above facts about wind and concentrations in Delhi.



(a) (b)

Fig. 3: Wind rose diagram for PM_{10} at IGI Airport, Delhi for (a)Decmeber 2011 and (b) Jauaray 2012.



(a) (b)

Fig. 4: Concentration polar plot for PM_{10} at IGI Airport, Delhi during (a) Decmeber 2011 and (b) Januaray 2012.

A comprehensive PCA analysis of fog formation conditions in Delhi due to air pollutant data, namely PM_{10}, NO_2, SO_2, O_3 and CO with meteorological variable temperature, humidity, visibility, wind speed, wind direction index, dew point temperature and sea level pressure has been studied. The correlation matrix of hourly visibility (VS), meteorological variables and concentration of air pollutants have been studied (Table 1) and analyzed through the scatter plots (Fig. 5) respectively. Table 1 reflects that the correlations between most of the variables are not poor as their values are more than 0.02, which represents that the correlation coefficient is significant. Since the critical value of the correlation coefficient of the 99- percentile confidence interval using the student t - test for the study period is found to be 0.02, which is statistically significant. Correlation coefficients computed between various fog hours and observed that visibility is positively correlated with temperature, Ozone, wind speed while negatively correlated with PM_{10}, relative humidity, dew point temperature, NO_2, SO_2 and CO.

Table 1: Correlation matrix of visibility, meteorological variables and air pollutants in Delhi.

	VS	O_3	NO_2	SO_2	CO	PM_{10}	RH	Temp	WDI	PRES	WS	DT	VS_{h-1}
VS	1												
O_3	0.108	1											
NO_2	-0.215	-0.432	1										
SO_2	-0.024	0.049	0.278	1									
CO	-0.163	0.013	-0.12	-0.208	1								
PM_{10}	-0.442	-0.326	0.365	0.213	0.079	1							
RH	-0.307	-0.576	0.027	-0.058	0.196	0.645	1						
Temp	0.137	0.571	-0.124	0.079	-0.175	-0.282	-0.557	1					
WDI	0.001	0.066	0.071	-0.002	-0.036	-0.053	-0.178	0.059	1				
PRES	-0.059	-0.06	0.012	-0.002	-0.022	0.092	0.157	-0.21	-0.053	1			
WS	0.322	0.378	-0.622	-0.161	-0.02	-0.433	-0.276	0.321	-0.018	-0.018	1		
DT	-0.433	-0.044	0.099	-0.03	0.024	0.376	0.242	0.332	0.049	-0.094	-0.185	1	
VS_{h-1}	0.603	0.122	-0.126	0.013	-0.178	-0.33	-0.281	0.139	-0.002	-0.067	0.24	-0.275	1

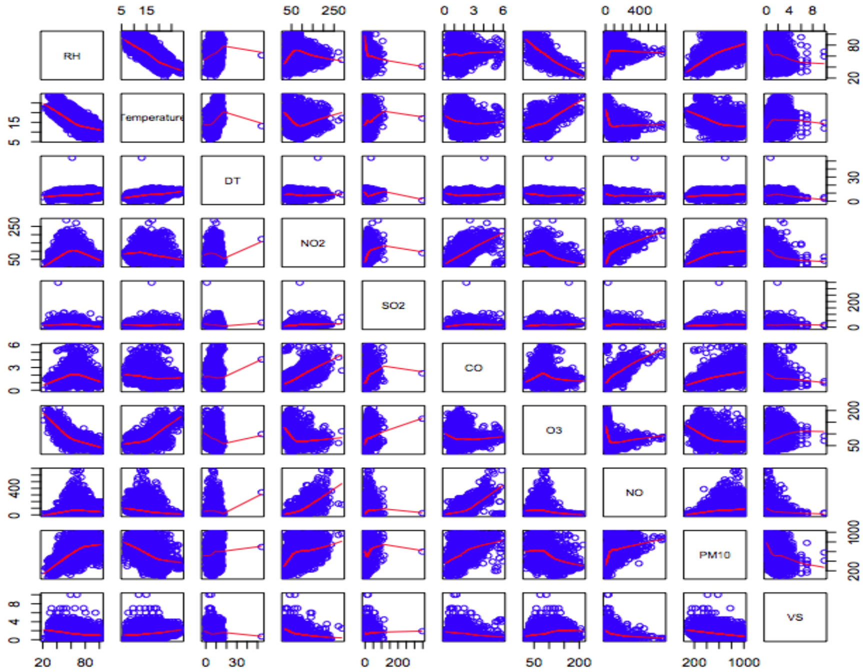


Fig. 5: Scatter plot matrix of input variables

Delhi's winter was marked by very dense fog, which dramatically reduces visibility and makes days colder by cutting off the sunlight. In January average temperatures begin to rise very gradually, though the rise is almost contained by the cold northeastern winds which result due to very heavy snowfall that occurs in the Himalayas during this part of the month. Sulfates and nitrates are formed in the atmosphere from sulfur dioxide (SO_2) and nitrogen dioxides (NO_2) emissions contribute to visibility impairment. According to EPA, sulfate particles are accounting for 50 to 70 percent of the visibility reduction in the Delhi. It is to be expected that visibility in certain wind directions, notably the East and Southeast, will be lower than in other directions, resulting from a combination of most polluted air masses, lighter winds and possibly less atmospheric mixing in more stable conditions. The visibility generally decreased despite an overall increase in relative humidity throughout the study period. Fig. 5 represents the pairwise scatter plot between the variables, where each row and column defines a single scatter plot. This matrix provides the relationships between the variables and the nature of these relationships. It can be noticed that visibility has the negative correlations with PM_{10}, NO, CO and relative humidity and positive correlations with temperature and O_3. All the other remaining variables are showing mixed relations.

The fog formation criteria using hourly average data of November, December and January (2011-12) were analyzed to find a linear regression visibility model amongst various parameters. A MLR model equation (3) was derived with the help of hourly time series data to explain the variation in visibility due to meteorological variables and air pollutants:

$$\begin{aligned}[VS] = {} & 2115.26 + 0.43 \times [VS_{h-1}] - 121.43 \times [DT] - 0.532 \times [PM_{10}] - 6.67 \times [O_3] - 3.97[NO_2] \\ & - 0.03 \times [CO] + 66.675 \times [Temp]\end{aligned} \tag{3}$$

where, VS- visibility (meters), VS_{h-1}- previous hour visibility (meters), DT- dew point temperature (^{0}C), Temp- temperature (^{0}C). The above equation was validated over available hourly time series data of Febuarary, 2012 at IGI Airport, Delhi. The predicted and observed fog occurrences conditions in terms of visibility are plotted in Fig. 6. The quantitative measure of regression model's performance has been analyzed through statistical performance measures that characterize the agreement between model prediction and observations of visibility.

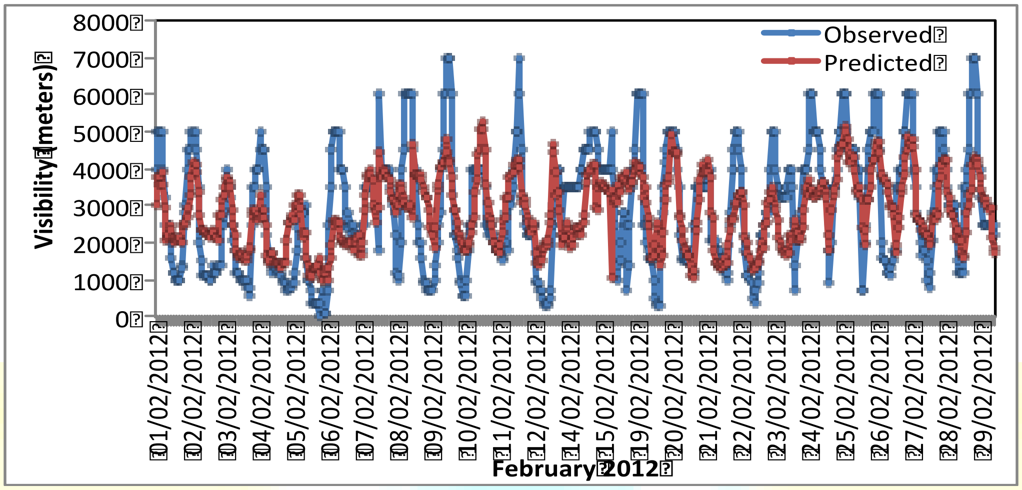


Fig. 6: Comparison between observed and predicted visibility (meters) using MLR model.

For the development of the ANN model for visibility, all predictor variables were normalized to the range [0,1] by linear scaling techniques. In other words, the input and output data would be converted to values between zero and one. The ANN model was developed using Matlab 7.12 (Licensed IIT Delhi). The relationship between the prediction variables and visibility has been well identified. The schematic of Artificial Neural Network structure (15:6:1) is used in winter with hyperbolic activation function. Hourly data for November, December of the year 2011 and January 2012 have been chosen for training and hourly data of February 2012 used for the validation of the model. The graph is plotted for predicted and observed visibility during validations periods in Fig. 7.

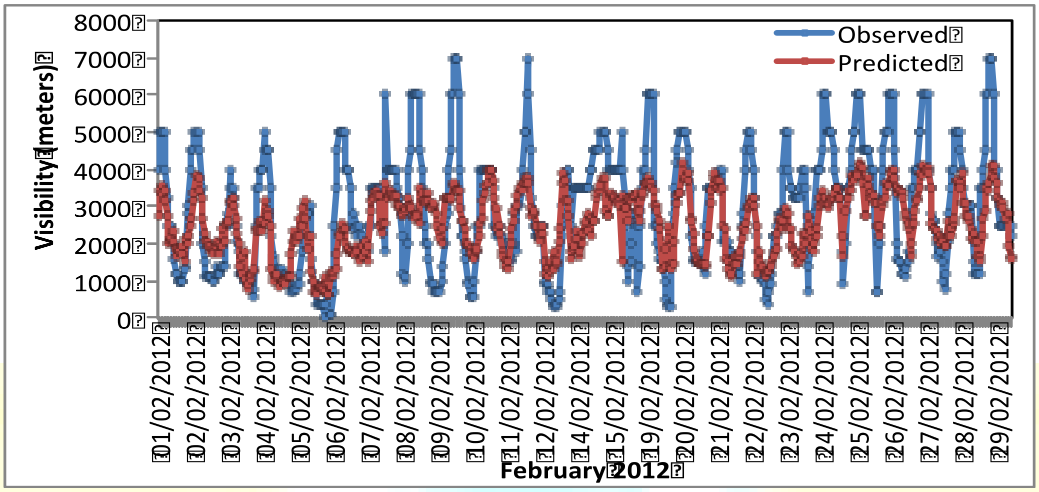


Fig. 7: Comparision between observed and predicted visibility (meters) using ANN model.

Now, the statistical performance computed between observed and predicted visibility from above both models i.e. MLR and ANN are shown in Table 2. The measure of model performance is the correlation coefficient (R), index of agreement (IOA), normalized mean square error (NMSE) and fractional bias (FB). The measures were calculated as follows (Chang et. al., 2004):

$$R = \frac{\overline{(V_O - \bar{V}_O)(V_P - \bar{V}_P)}}{\sigma_{V_P}\sigma_{V_O}} \tag{4}$$

$$IOA = 1 - \frac{\sum_{i=1}^{n}(V_{Pi} - V_{Oi})^2}{\sum_{i=1}^{n}(|V_{pi} - \bar{V}_o| + |V_{Oi} - \bar{V}_O|)^2}$$

(5)

$$NMSE = \frac{\overline{(V_O - V_P)^2}}{\bar{V}_O . \bar{V}_P} \tag{6}$$

$$FB = \frac{(\bar{V}_P - \bar{V}_O)}{0.5(\bar{V}_P + \bar{V}_O)} \quad (7)$$

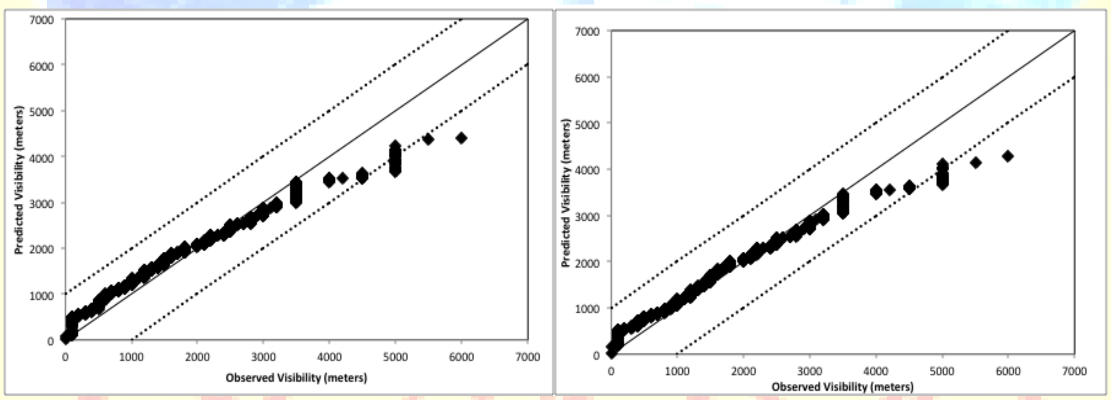
where, V_P is the model predictions, V_O is the observations, overbar $\bar{V}$ is the average over the dataset, and σ_V is the standard deviation over the data set.

Table 2: Statistical performance measure for visibility by taking MLR and ANN models.

Statistical Measures	Ideal Value	MLR		ANN	
		Training	Validation	Training	Validation
R	1	0.72	0.8	0.85	0.79
IOA	1	0.95	0.98	0.96	0.98
NMSE	0	0.23	0.08	0.22	0.09
FB	1	-0.04	0.07	0.04	0.06

From Table 2, the comparison of R, IOA, NMSE and FB values for MLR and ANN models are in good agreement with observed values and approximately near to corresponding ideal values. It further strengthens that a better forecast was resulted from ANN model. Thus, the best agreement was obtained for ANN model, while the performance of MLR was somewhat poorer.

The Quantile-Quantile (Q-Q) plots have been used extensively to evaluate the results of a model in many studies and have been considered as a useful tool for characterizing the model performance (Venktaram et al., 2001). Here, Q-Q plots are used to compare the predicted and observed visibility distance at IGI Airport, Delhi (Fig. 8). In Q-Q plots, the sorted predicted visibility distances are plotted against the sorted observed values (i.e. independent of time and position) to determine whether the observed and predicted visibility distance datasets come from populations with a common distribution. If the two sets come from a population with the same distribution, the points should fall approximately along the 1:1 reference line. The greater the departure from this reference line, the greater the evidence for the conclusion that the two data sets have come from populations with different distributions. The values close to the 1:1 line indicate a good fit between the predictions and observation data. The values above the 1:1 solid line shows the level of over-prediction and values lying below it under-prediction. The values on dashed line shows the predicted values are half of the corresponding observed values. Here both models showing good agreement during the study periods at IGI Airport, Delhi and ANN model showing better in comparison to MLR because the maximum quartiles are near to 1:1 line. Thus, ANN model is showing better results in comparison to MLR model.



(a) (b)

Fig. 8: Quantile-Quantile (Q-Q) plots between observed and predicted visibility for (a) MLR and (b) ANN models

4. Conclusions

In this study the fog formation criteria i.e. visibility has been studied and observed that visibility was not only influenced by the meteorological variables, but also by the concentration of air pollutants. It was observed that under the certain meteorological conditions, the increased concentrations of pollutants were causing the formation of fog. A novel approach, based on regression techniques, PCA was employed in selecting the input variables. In such a way, the selection of the input variables in the models was based on the objective to study the influences of air pollutants and meteorological variables on visibility due to fog formation in Delhi. The regression models MLR and ANN have been used for hourly forecasting of visibility conditions at IGI Airport, Delhi. These models were built on measured meteorological variables and concentrations of the pollutants concerned. Generally a good agreement between the proposed models and the observed visibility distances confirms the relationship between the physical state of the atmosphere and fate of pollutants. The poorer agreement between the modeled and observed visibility can be attributed to the irregularity of the process affecting fog formation (such as traffic, dust storm, resuspension etc.) or due to omission of the relevant variables (e.g. solar radiation, boundary layer height, etc.). Results of regression analysis have suggested that the degradation of visibility in Delhi was mainly due to increment of air pollutions (PM_{10}, NO_2 and CO) concentrations in the ambient atmosphere of Delhi. However, relative humidity and dew point temperature were the main meteorological variables are favoring for fog formation or visibility reduction. In this study, the measured data were employed as meteorological input. However, the employed input meteorological variables are generally available for routine weather prediction models. Thus, the developed model can be used for operational forecasts of visibility and fog. In summary, it can be concluded that the most prominent factors affecting the investigated fog conditions in Delhi were either high concentrations of criteria pollutants or some meteorological variables. It is concluded that the exposure of air pollutants in the atmosphere has been linked with the deterioration of visibility with fog formations.

Acknowledgement

The authors would like to thanks Central Pollution Control Board, New Delhi for providing the air pollution data and Indian Meteorological Department, New Delhi for meteorological data used in this study.

References

Chan, Y.C., Simpson, R.W., Mctainsh, G.H., Vowles, P.D., Cohen, D.D., Bailey, G.M., 1997 . Characterisation of chemical species in $PM_{2.5}$ and PM_{10} aerosols in Brisbane, Australia. Atmospheric Environment 31, 3773–3785.

Chang, J.C., Hanna, S.R., 2004. Air quality model performance evaluation. Meteorology and Atmospheric Physics 87, 167- 196.

Deng, X., Tie, X., Wu, D., Zhoud, X., Bi, X., Tan, H., Li, F., Jiang, C., 2008. Long-term trend of visibility and its characterizations in the Pearl River Delta (PRD) region, China. Atmospheric Environment 42, 1424–1435.

Goyal P., Kumar, A., Mishra, D., 2013. The impact of air pollutants and meteorological variables on visibility in Delhi. Environmental Modeling & Assessment, DOI 10.1007/s10666-013-9380-4.

Horvath, H., 1995. Estimation of the average visibility in Central Europe, Atmospheric Environment 29, 241-246.

Kumar, A., Goyal, P., 2011. Forecasting of air quality in Delhi using principal component regression technique. analysis. Atmospheric Pollution Research 2, 436- 444

Kumar, A., Goyal, P., 2013. Forecasting of air quality index in Delhi using neural network based on principal component analysis. Pure Appllied Geophysics 170, 711–722.

Rajeevan, M., Pai, D. S., Rohilla, A. K., 2005. New statistical models for long range forecasting of southwest monsoon rainfall over India. NCC Research Report, National Climate Centre, India Meteorological Department, Pune – 411005.

Venkatram, A., Brode, R., Cimorelli, A., Lee, R. Paine, R., Perry, S., Peters, W., Weil, J. and Wilson, R., 2001. A Complex Terrain Dispersion Model for Regulatory Applications. Atmospheric Environment 35, 4211–4221.