View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Elsevier - Publisher Connector

Atmospheric Pollution Research 2 (2011) 436-444

Atmospheric Pollution Research

www.atmospolres.com

Forecasting of air quality in Delhi using principal component regression technique

Anikender Kumar, Pramila Goyal

Centre for Atmospheric Sciences, Indian Institute of Technology Delhi, Hauz Khas, New Delhi-110016 India

ABSTRACT

Over the past decade, an increasing interest has evolved by the public in the day–to–day air quality conditions to which they are exposed. Driven by the increasing awareness of the health aspects of air pollution exposure, especially by most sensitive sub–populations such as children and the elderly, short–term air pollution forecasts are being provided more and more by local authorities. The Air Quality Index (AQI) is a number used by governmental agencies to characterize the quality of the air at a given location. AQI is used for local and regional air quality management in many metropolitan cities of the world. The main objective of the present study is to forecast short–term daily AQI through previous day's AQI and meteorological variables using principal component regression (PCR) technique. This study has been made for four different seasons namely summer, monsoon, post monsoon and winter. AQI was estimated for the period of seven years from 2000–2006 at ITO (a busiest traffic intersection) for criteria pollutants such as respirable suspended particulate matter (RSPM), sulfur dioxide (SO_2), nitrogen dioxide (NO_2) and suspended particulate matter (SPM) using a method of US Environmental Protection Agency (USEPA), in which sub–index and breakpoint pollutant concentration depends on Indian National Ambient Air Quality Standard (NAAQS). The Principal components have been computed using covariance of input data matrix. Only those components, having eigenvalues ≥ 1, were used to predict the AQI using principal component regression technique. The performance of PCR model, used for forecasting of AQI, was better in winter than the other seasons as studied through statistical error analysis. The values of normalized mean square error (NMSE) were found as 0.0058, 0.0082, 0.0241 and 0.0418 for winter, summer, post monsoon and monsoon respectively. The other statistical parameters are also supporting the same result.

Keywords:
Forecasting
Principal component regression (PCR)
Air quality index (AQI)
Meteorological variables
Pollutants

Article History:
Received: 03 November 2010
Revised: 02 February 2011
Accepted: 08 February 2011

Corresponding Author:
Pramila Goyal
Tel: +91-11-26591309
Fax: +91-11-26591386
E-mail: pramila@cas.iitd.ernet.in

© Author(s) 2011. This work is distributed under the Creative Commons Attribution 3.0 License.

doi: 10.5094/APR.2011.050

1. Introduction

With continuous development and increase of population in the urban areas, a series of problems related to environment such as deforestation, release of toxic materials, solid waste disposals, air pollution and many more, have attracted attention much greater than ever before. The problem of air pollution in cities has become so severe that there is a need for timely information about changes in the pollution level. Today forecasting of air quality is one of the major topics of air pollution studies due to the health effects caused by these airborne pollutants in urban areas during pollution episodes. Therefore, the development of effective forecasting models of AQI for major air pollutants in urban areas is of prime importance. With this end in view, there is a need to have a model that would generate the future AQI. Although many forecasting models exist and some are in use, there is still need for developing more accurate models. The Gaussian dispersion models are generally used in most of the air pollution studies. Even though the models have some physical basis, detailed information about the source of the pollutants and other variables are generally not known. In order to overcome these limitations, statistical models are used, which facilitate the prediction of pollutant concen–trations (Finzi and Tebaldi, 1982; Ziomass et al., 1995; Polydoras et al., 1998).

Numerous studies based on the statistical models have been carried out in different regions to identify local meteorological conditions, most strongly associated with air pollutant concen–trations, and to forecast their values (McCollister and Willson, 1975; Aron and Aron, 1978; Lin, 1982; Aron, 1984; Katsoulis, 1988; Robeson and Steyn, 1990). Many of the previous studies (Sanchez et. al., 1990; Mantis et al., 1992; Milionis and Davies, 1994) analyzed the meteorological conditions associated with high pollutant concentrations. These studies usually produced qualitative or semi–quantitative results and shed a light on the relation between the meteorological conditions and pollutant concentrations. Shi and Harrison (1997) developed a regression model for the prediction of NO_x and NO_2 in London. Some non–linear models i.e., Artificial Neural Networks can also be used to forecast the pollutant concentrations (Boznar et al., 1993; Comrie, 1997).

As for the health impact of air pollutants, AQI is an important indicator for general public to understand easily how bad or good the air quality is for their health and to assist in data interpretation for decision making processes related to pollution mitigation measures and environmental management. Basically, the AQI is defined as an index or rating scale for reporting daily combined effect of ambient air pollutants recorded in the monitoring sites. Recently, Van den Elshout et al. (2008) gave a review of existing air quality indices and a proposal of a common alternative. Fuzzy inference systems have also used in modeling of air quality indices by Hajek and Olej (2009). A regression model was also used by Cogliani (2001) for air pollution forecast in cities by an air pollution index highly correlated with meteorological variables. However, when multicollinearity is present, the computations of regression coefficients in regression models become dubious. Principal component analysis (PCA) can be applied to overcome the above limitation. PCA is also a procedure to reduce the number of variables. It is useful when obtained data has a number of variables (possibly a large number of variables), and believed that there is some variables those are correlated with one another. Sanchez et al. (1986) used the principal component factor analysis for studying the spatial and temporal distribution of SO_2 in an urban area. The PCR technique was also used to forecast the long–range forecasting of Southwest monsoon rainfall over India (Rajeevan et al., 2005). The most of the air quality forecasting in Delhi has been done through individual air pollutants whereas the present study was conducted using principal component regression technique with respect to daily AQI. The PCR model is used in the present study to forecast the daily air quality index one day in advance.

1.1. Description and meteorology of study area

The Delhi city (Latitude 28°35'N, Longitude 77°12'E) is located in the northern part of India and situated between the Great Indian Desert (Thar Desert) of Rajasthan to the west, the central hot plains to the south and the cooler hilly region to the north and east. Delhi has a semi–arid climate with an extremely hot summer, average rainfall and very cold winter. Due to the worst meteorological scenario, the most important season in Delhi is winter, which starts in December and ends in February. This period is dominated by cold, dry air and ground–based inversion with low wind conditions ($u \leq 1$ m s^{-1}), which occur frequently and increases the concentration of pollutants (Anfossi et al., 1990). The summer (March, April, May) is governed by high temperature and high winds, while the monsoon (June, July, August) is dominated by rains and post–monsoon (September, October, November) has moderate temperature and wind conditions.

Delhi, the capital city of India with 13.8 million inhabitants spread over 1 483 km^2 (Aneja et al., 2001). Due to the presence of large number of industries and migration of people from neighboring states, nearly 5.4 million vehicles are running on Delhi roads. The emission of pollutants from these sources deteriorates the ambient air quality. The steep increase in vehicular population (major source of air pollution) has resulted in corresponding increase in pollutants emitted by these vehicles. Presently, more than 1 300 tons of pollutants are emitted by the vehicles in Delhi. Due to the increased level of pollutants, Delhi's air is blamed for 40% of emergency hospital admissions of patients with breathing and heart complaints. The ambient air quality data of Delhi monitored by Central Pollution Control Board (CPCB) shows very high values of suspended particles, SO_2 and NO_x which have been beyond the permissible limits from last several years continuously (Goyal and Sidhartha, 2003). All India Institute of Medical Sciences (AIIMS) reports that there was a massive 900% increase in asthma cases in December 1999 compared to December 1998. Study by Brandon and Hommann (1995), by using the standard US metric, estimated that the 7 490 deaths could be avoided in Delhi by a 141.6 μg m^{-3} reduction in PM_{10}. One, out of every 10 school children in the city, suffers from asthma that is worsening due to vehicular air pollution.

The primary objective of this study is to forecast the daily AQI one day in advance on seasonal basis. A busiest traffic intersection ITO has been chosen to forecast using the air pollutant concentrations monitored at ITO, since it is a continuous air quality monitoring station at the same place. The daily air quality parameters (daily average concentrations of pollutants) namely RSPM, SO_2, NO_2 and SPM used in the present study were measured by CPCB, a regulatory monitoring agency in Delhi. The locations of monitoring stations were categorized on a land use basis (CPCB, 2005) i.e., residential, industrial and traffic intersections. The station that is classified as traffic intersection is ITO. In this study, AQI was calculated using USEPA method in which sub–index and breakpoint pollutant concentrations depend on Indian NAAQS and a PCR technique was also used to forecast the short term i.e., daily AQI through previous day's AQI and meteorological variables.

The 24–hourly averaged surface meteorological variables at Safdarjung airport like daily maximum temperature (t_{max}), minimum temperature (t_{min}), daily temperature range (difference between daily maximum and minimum temperature, t_{range}), average temperature (t_{avg}), wind speed (wsp), wind direction index (wdi), relative humidity (rh), vapor pressure (vp), station level pressure (slp), rainfall (rf), sunshine hours (ssh), cloud cover (cc), visibility (v) and radiation (rd) for Delhi were acquired from the Indian Meteorological Department (IMD), Pune for the period of 2000–2006. There is no meteorological station at ITO. The meteorological station is at Safdarjung airport about 5.7 km from the ITO and this is the only station in the study area (ITO). Figure 1 shows the air pollution monitoring (ITO) and meteorological (Safdarjung airport) stations on the Map of Delhi.

2. Materials and Methods

There are primarily two steps involved in formulating an AQI: first the formation of sub–indices of each pollutant, second the aggregation (breakpoints) of sub indices. Breakpoint concentrations of each pollutant, used in calculation of AQI, are based on Indian NAAQS and results of epidemiological studies indicating the risk of adverse health effects of specific pollutants. It has been noticed that different breakpoint concentrations and different air quality standards have been reported in literature (Environmental Protection Agency, 1999). In India, to reflect the status of air quality and its effects on human health, the range of index values has been designated as "Good (0–100)", "Moderate (101–200)", "Poor (201–300)", "Very Poor (301–400)" and "Severe (401–500)" (Table 1) (Nagendra et al., 2007).

All the values of SO_2, NO_2, RSPM and SPM are in μg m^{-3}.

The formula (EPA, 1999) used to calculate AQI for four criteria pollutants RSPM, SO_2, NO_2 and SPM from 2000–2006 is given below:

$$I_P = \left[\frac{\left(I_{Hi} - I_{Lo}\right)}{\left(BP_{Hi} - BP_{Lo}\right)} \right] \left(C_P - BP_{Lo}\right) + I_{Lo} \tag{1}$$

where I_P is the AQI for pollutant "p", C_P is the actual ambient concentration of the pollutant "p", BP_{Hi} is the breakpoint in Table 1 that is greater than or equal to C_p, BP_{Lo} is the breakpoint in Table 1 that is less than or equal to C_p, I_{Hi} is the sub index value corresponding to BP_{Hi}, and, I_{Lo} is the sub index value corresponding to BP_{Lo}.

The AQI is determined on the basis of AQI of study pollutants and the highest among them is declared as the overall AQI. The formula used here is same as used by USEPA, in which sub-index and breakpoint concentration depends on Indian NAAQS.

2.1. Multiple Linear Regression and Principal Component Regression model

A forecast can be expressed as a function of a certain number of factors that determine its outcome. Multiple linear regression (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):

$$Y= b_1 + b_2 X_2 + ...+ b_k X_k + e \tag{2}$$

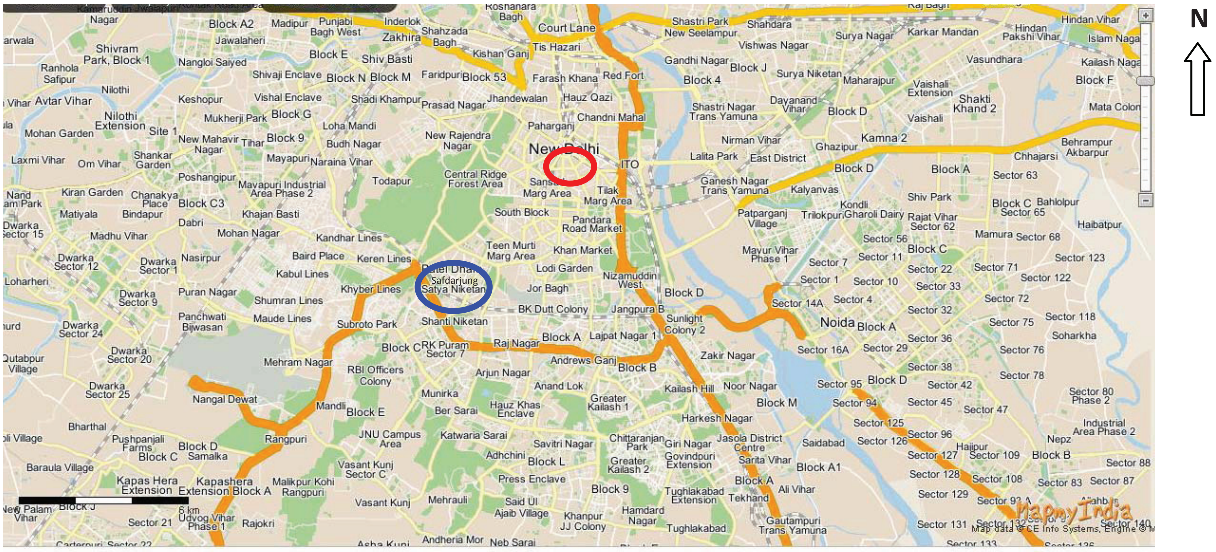
N
New Delhi
ITO
Safdarjung
Satya Niketan
Noida
MapmyIndia
○ ITO (Air quality monitoring station)
○ Safdarjung (Meteorological station)

Figure 1. *Map of Delhi with air pollution monitoring (ITO) and meteorological (Safdarjung airport) stations (Source: http://www.mapmyindia.com/).*

Table 1. *Proposed sub-index and breakpoint pollutant concentrations for Indian-AQI*

SI.No.	Index values	Descriptor	SO_2 (24-h avg.)	NO_2 (24-h avg.)	RSPM (24-h avg.)	SPM (24-h avg.)
1	0-100	Good [a]	0-80	0-80	0-100	0-200
2	101-200	Moderate [b]	81-367	81-180	101-150	201-260
3	201-300	Poor [c]	368-786	181-564	151-350	261-400
4	301-400	Very Poor [d]	787-1 572	565-1 272	351-420	401-800
5	401-500	Severe [e]	>1572	>1272	>420	>800

[a] *Good: Air quality is acceptable; however, for some pollutants there may be a moderate health concern for a very small number of people.*
[b] *Moderate: Members of sensitive groups may experience health effects.*
[c] *Poor: Members of sensitive groups may experience more serious health effects.*
[d] *Very poor: Triggers health alter, everyone may experience more serious health effects.*
[e] *Severe: Triggers health warnings of emergency conditions.*

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 model, AQI is the dependent variable and, previous day's AQI 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.

Equation (2) can be written in the following form:

$$Y = X\,b + e \tag{3}$$

where

$$Y = \begin{bmatrix} Y_1 \\ Y_2 \\ : \\ : \\ Y_n \end{bmatrix},\ X = \begin{bmatrix} 1\,X_{21}\,X_{31}\,....\,X_{k1} \\ .... \quad .. \\ 1\,X_{2n}\,X_{3n}\,....\,X_{kn} \end{bmatrix},\ b = \begin{bmatrix} b_1 \\ b_2 \\ : \\ : \\ b_k \end{bmatrix} \text{ and } e = \begin{bmatrix} e_1 \\ e_2 \\ : \\ : \\ e_n \end{bmatrix}$$

So Y is an nx1, X is an n x k, b is a k x 1 and e is an n x 1 matrix.

After using the minimum square error technique, the solution can be obtained as $b = (X'\ X)^{-1}\ (X'Y)$. Further, the F–test has been performed to determine whether a relationship exists between the dependent variable and the regressors. The t–test is performed in order to determine the potential value of each of the regressor variables in the regression model. The resulting model can be used to predict future observations.

When multicollinearity is present the computation of an inverse matrix $(X'\ X)^{-1}$ becomes dubious. PCA can be applied to overcome this limitation. It is useful when large number of variables are present, and also if there are some variables correlated with each other.

The application of PCA with regression model aims to reduce the collinearity in the datasets which leads to the worst predictions and also determine the relevant independent variables for the prediction of air pollutant concentrations (Sousa et al., 2007). The difference between PCR and MLR is mainly due to input data. PCR model takes PCs of variables as input data and reduces the complexity due to less number of input variables.

Computation of principal components. Principal components can be computed by covariance of input data matrix. In this study, the covariance matrix of the initial data was considered. The eigenvalues of the covariance matrix ”C” are obtained from its characteristic equation:

$$|C-\lambda I|=0 \tag{4}$$

where, λ is the eigenvalue and I is the identity matrix.

For each eigenvalue, a non–zero vector e can be defined as:

$$C\,e=\lambda\,e \tag{5}$$

where the vector *e* is called the characteristic vector or eigenvector of the covariance matrix C associated with its corresponding eigenvalue. The eigenvectors derived from the covariance matrix represent the mutually orthogonal linear combination of the matrix. Their associated eigenvalue represent the amount of total variance, which is explained by each of the eigenvectors. By retaining only the first few pairs of eigenvalue–eigenvector, or principal components, a substantial amount of the total variance can be explained while explaining the higher order principal components which explain minimal amounts of the total variance and can be viewed as noise. Variance explained by i^{th} PC is given by:

$$\text{The variance}_i=\frac{\lambda_i}{\sum_n \lambda_n} \tag{6}$$

The PC associated with the greatest eigenvalue, the first PC (PC1), represents the linear combination of the variables accounting for the maximum total variability in the data. The second PC explains the maximum variability that is not accounted by the PC1 and so on. All components with eigenvalues$\geq$1 should be retained. The rationale behind this method is an eigenvalue of 1, represents amount of variance, explained by the original variables, and components of eigenvalue<1 explain less variance than the original variables. After getting the PC’s, the initial data set is transformed in to the orthogonal set by multiplying the eigenvectors to the initial data set. Now this transformed data set is used as input to the multiple linear regression technique.

$$Y=\beta_0+\beta_1(PC_1)+\beta_2(PC_2)+\ldots\ldots\ldots\ldots\ldots\ldots\ldots+\beta_n(PC_n)+e \tag{7}$$

where β_0, β_1, β_2……β_n are the coefficients in the model equation. The coefficients of regression model have been estimated using the least squares method. Further, the F–test has been performed to determine whether a relationship exists between the dependent variable and the regressors. The t–test is performed in order to determine the potential value of each of the regressor variables in the regression model. The resulting model can be used to predict future observations.

Principal Components were computed using the data for the years 2000–2005 and also used as an input to regression model to form PCR model. The same process was adopted for all four seasons. The previous day’s AQI and meteorological variables (15 variables, as mentioned in Section 1) for the years 2000–2005 were used as the input to PCR model. The covariance matrix of the given input is determined. The PCs have been determined on the basis of the variance explained by the eigenvalues of the covariance matrix. Only those principal components whose eigenvalues ≥ 1 based on the analysis of 15 variables, were used to forecast one daily air quality index. The application of PCA with regression models reduces the collinearity of the datasets, which can lead to worst predictions and also determines the relevant independent variables for the prediction of AQI. The architecture of the PCR model to forecast AQI has been shown in Figure 2. The difference between the PCR and MLR is due to the input variables. Consequently, the network architecture will be less complex in PCR due to the decreased number of input variables.

Once this process has been completed, the performance of PCR model has been validated with an independent data, observed for the year 2006, that has been transformed to the new data set using the previously determined weights of principal components. It is important to mention that 2006 data was not used to build the model. The accuracy of the model was analyzed through the statistical parameters.

3. Results and Discussion

The forecasting of daily AQI, based on previous day’s AQI and meteorological variables, was done using MLR and PCR model on the seasonal basis for the period of 2000–2005 and validated through the daily AQI of 2006.

The regression models for different seasons, summer, monsoon, post monsoon and winter were developed using the MLR technique on the basis of daily data of 2000–2005 using the procedure discussed in Section 2. The regression equations based on MLR technique are obtained as Equations (8), (9), (10) and (11) for summer, monsoon, post monsoon and winter season, respectively.

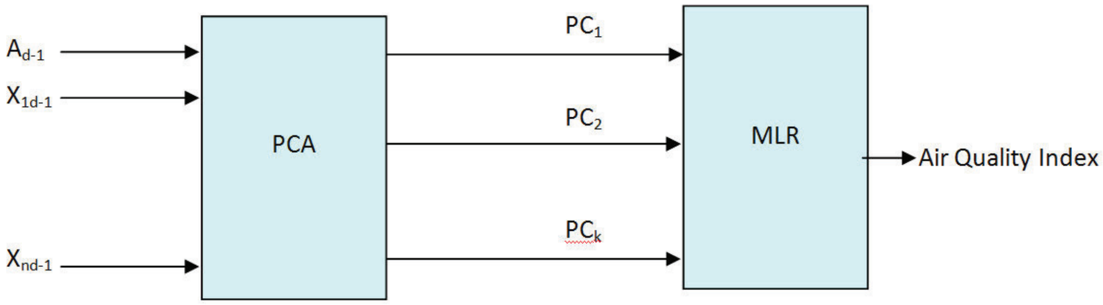

Figure 2. *Architecture of PCR model for the forecasting of AQI.*

$$[AQI]=131.26+0.503\times[AQI_{d-1}]-0.462\times[rh]+1.689\times[t_{max}]-6.131\times[cc] \quad (8)$$

$$[AQI]=2493.64+0.599\times[AQI_{d-1}]-1.736\times[rh]-22.89\times[v]-3.44\times[t_{min}] \quad (9)$$

$$[AQI]=361.33+0.537\times[AQI_{d-1}]+1.72\times[slp]-1.67\times[vp]+2.697\times[ssh]-20.49[v]+1.49[t_{max}] \quad (10)$$

$$[AQI]=1728.11+0.503\times[AQI_{d-1}]-15.60\times[v]-7.98\times[cc]+4.50\times[wsp]-1.19[rh]+0.84[rf] \quad (11)$$

Equations (8)–(11) show that previous day's air quality index is the common variable for all seasons.

The daily AQI of the year 2006 has been forecasted using the above equations, which has been compared with observed AQI of 2006. The statistical evaluation of forecasted and observed AQI values is shown in Table 2. Results indicated that the MLR model is performing satisfactorily in all seasons and gives better results in winter with respect to the NMSE and Root Mean Square Error (RMSE). It shows a minor difference in coefficient of determination compared to the other seasons. Fractional bias shows that the model is under–predicting in summer, post monsoon and winter in training as well as in validation and is over–predicting in monsoon season.

The PCR models for summer, monsoon, post monsoon and winter, based on the daily data for the years 2000–2005 were developed as discussed in Section 2 and were analyzed statistically.

As a first step, data for the years 2000–2005 was used to calculate the covariance matrix for all four seasons. The predictor variables were transformed into principal components through the eigenvalue matrix of variables that would explain most of the total variation in the data. Table 3 represents the eigenvalues and amount of variance, explained by each principal component with eigenvalues≥1. Rest of the components having eigenvalues<1, explaining less variance than any of original variables were ignored. Table 3 also shows that only 5 PCs have eigenvalues≥1 with a cumulative variance of 69.98 in summer and 4 PCs have eigenvalues≥1 with cumulative variances of 60.79, 68.35 and 66.62 in monsoon, post monsoon and winter, respectively. Communalities of each original variable in all four seasons are shown in Table 4, using the first 5 PCs in summer and 4 PCs in monsoon, post monsoon and winter seasons. This Table reflects that the most relevant original variables for PCR are average daily temperature, relative humidity, daily minimum temperature and daily average temperature in summer, monsoon, post monsoon and winter, respectively.

Table 2. *Comparison of MLR model predicted and observed values in years 2000-2005 and year 2006*

S.N.	Season	2000-2005				2006			
		RMSE	NMSE	Coefficient of determination	Fractional Bias	RMSE	NMSE	Coefficient of determination	Fractional Bias
1	Summer	35.62	0.0114	0.4983	0.0004	35.41	0.0112	0.5495	0.0509
2	Monsoon	53.18	0.0405	0.6362	-0.0003	56.53	0.0427	0.4203	-0.0022
3	Post Monsoon	40.63	0.0174	0.6680	0.0336	48.98	0.0260	0.5014	0.0309
4	Winter	40.03	0.1227	0.4590	0.0005	31.78	0.0080	0.3936	0.0446

Table 3. *Eigenvalues and explained variance of the computed PCs for summer, monsoon, post monsoon and winter seasons*

Seasons	Principal Component	Eigenvalue	% of Variance	Cumulative variance (%)
Summer	1	4.7032	31.35	31.35
	2	2.1172	14.11	45.47
	3	1.5657	10.44	55.91
	4	1.1039	7.36	63.27
	5	1.0076	6.72	69.98
Monsoon	1	5.1592	34.39	34.39
	2	1.5240	10.16	44.55
	3	1.3698	9.13	53.69
	4	1.0658	7.11	60.79
Post-monsoon	1	5.7652	38.43	38.43
	2	2.2458	14.97	53.41
	3	1.1746	7.83	61.24
	4	1.0676	7.12	68.35
Winter	1	4.1107	27.40	27.40
	2	2.7014	18.01	45.41
	3	2.0011	13.34	58.75
	4	1.1804	7.87	66.62

Table 4. *Communalities of each original variable for summer, monsoon, post-monsoon and winter seasons*

Variable	Summer	Monsoon	Post-Monsoon	Winter
AQI $_{d-1}$	0.6370	0.6802	0.6958	0.7135
t_{avg}	0.9597	0.8871	0.8585	0.9272
rh	0.7880	0.8949	0.7630	0.8267
vp	0.6849	0.5720	0.8899	0.7395
rf	0.5502	0.6015	0.2131	0.5305
wsp	0.7752	0.5598	0.7816	0.6079
wdi	0.7095	0.4124	0.6415	0.2377
rd	0.2247	0.2154	0.4218	0.4333
t_{max}	0.9463	0.8414	0.8516	0.9085
t_{min}	0.9503	0.7778	0.9116	0.8872
ssh	0.6525	0.6481	0.7016	0.7469
slp	0.7816	0.1948	0.4904	0.4580
v	0.5867	0.6419	0.6439	0.6962
cc	0.4232	0.4081	0.5845	0.4273
t_{range}	0.8271	0.7825	0.8038	0.8526

The loadings (or coefficients) of each input variable corresponding to all 5 PCs in summer and 4 PCs in monsoon, post monsoon and winter are given in Tables S1, S2, S3 and S4, respectively (see the Supporting Material, SM). In this case, only 5 new variables (PCs) were used instead of original 15 variables in summer and 4 new variables were used for the remaining three seasons.

The PCR models for all four seasons based on the transferred data for 2000–2005 were developed and analyzed statistically. The T–test was used to test the significance of the variables. Insignificant/ statistically invalid variables were removed from the model equation. It was observed that 3 PCs (PC1, PC2 and PC3) lied in 95% confidence interval and these variables were retained in the model equations. It was also observed that two PCs (PC1 and PC2), one PC (PC1) and two PCs (PC2 and PC4) lied in 95% confidence interval in monsoon, post monsoon and winter, respectively. Thus, the forecasting Equations (12, 13, 14, and 15) using PCR technique for four seasons are shown below:

$$[\mathrm{AQI}] = 450.53 - 0.773 \times \mathrm{PC1} + 1.040 \times \mathrm{PC2} + 0.969 \times \mathrm{PC3} \tag{12}$$

$$[\mathrm{AQI}] = 257.59 + 1.639 \times \mathrm{PC1} - 0.398 \times \mathrm{PC2} \tag{13}$$

$$[\mathrm{AQI}] = -260.605 + 1.856 \times \mathrm{PC1} \tag{14}$$

$$[\mathrm{AQI}] = -162.504 + 1.462 \times \mathrm{PC2} + 0.617 \times \mathrm{PC4} \tag{15}$$

Figures 3a, 3b, 3c, and 3d show graphical presentation of four different seasons. The coefficient of correlation (R) between observed and forecasted values for the years 2000–2005 were found as 0.70, 0.79, 0.80 and 0.64 in summer, monsoon, post monsoon and winter, respectively. The daily AQI of the year 2006 was forecasted using the Equations (12)–(15).

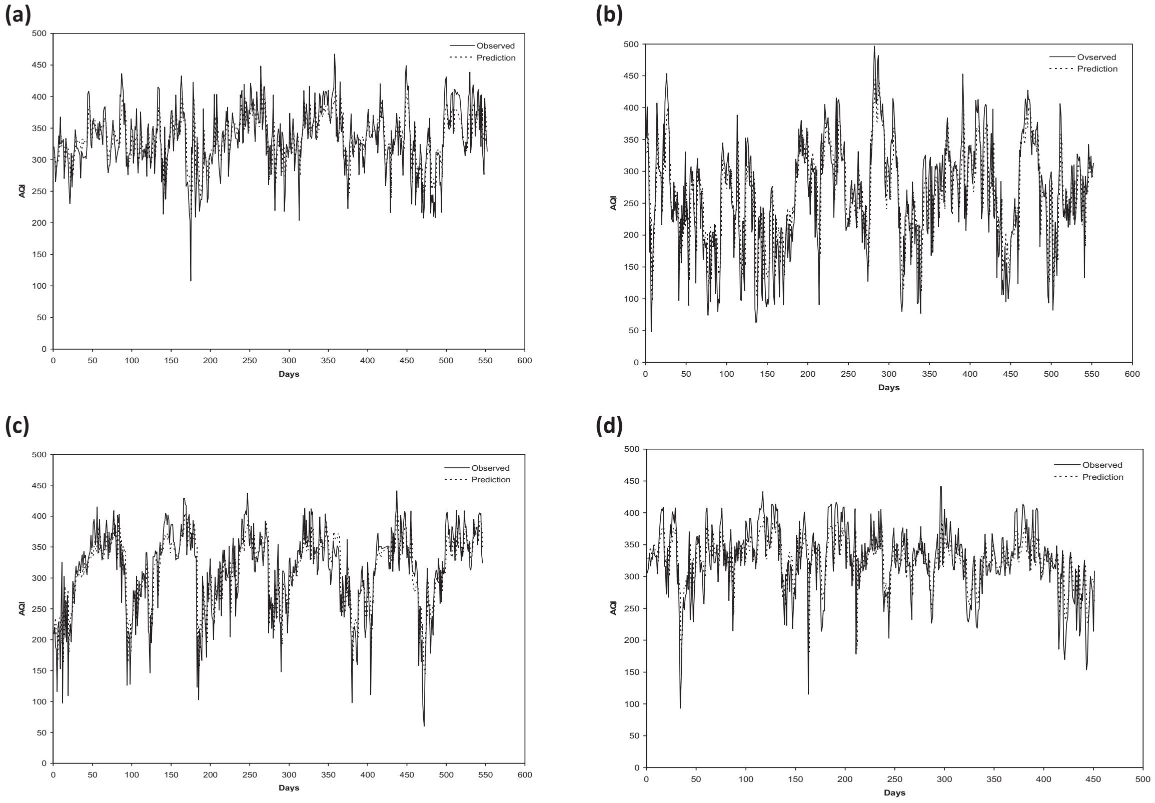

Figure 3. *Comparison of observed and model predicted values of daily AQI in* ***(a)*** *Summer,* ***(b)*** *Monsoon,* ***(c)*** *Post Monsoon and* ***(d)*** *Winter seasons during the years 2000-2005.*

The comparison of forecasted and observed AQI values for the year 2006 are shown in Figures 4a, 4b, 4c, and 4d for summer, monsoon, post monsoon and winter, respectively. Figures 3 and 4 show that maximum observed AQI is 497 and the minimum is 48 in monsoon season of 2003 and 2000 respectively, whereas the predicted maximum AQI is 447 and the minimum is 96.5 in monsoon for the periods of 2003 and 2000. The observed values for 2006 were not included in model development for forecasting AQI of 2006. Figure 4 also shows that there is one day shifting between the predicted and observed values of AQI. The reason for this shifting may be due to the uncertainties involved in the air quality data for the years 2000–2005 that was validated with the data for 2006.

Statistical evaluation between observed and predicted values for 2000–2005 and 2006 was made for different seasons in Table 5. The NMSE and coefficient of determination (R^2) are found as (0.0082, 0.5767) in summer, followed by (0.0418, 0.4225) in monsoon; (0.0241, 0.5155) in post monsoon and (0.0058, 0.5625) in winter seasons during 2006. This shows that forecasted AQI could be explained by the selected input variables as approximately 58% in summer, 57% in winter, 52% in post monsoon and 42% in monsoon seasons. Fractional bias shows the under–prediction of PCR model in all the seasons in training as well as in validation. However, the overall performance of the PCR model was found better in comparison to the MLR model and also model's performance was found to be better in winter compared to other seasons.

4. Conclusions

In the present study, the daily AQI at ITO was forecasted using the MLR and PCR models based on the previous day's AQI and meteorological variables.

The statistically error analysis of model evaluation for all four seasons shows that model is performing satisfactorily in all the seasons but is performing better in winter than the other seasons. The use of PCs based models was found useful due to elimination of collinearity problems in MLR and reduction of the number of predictors. It is also found that the performance of the PCR model was found better in comparison to the MLR model in 2006 validation period. Finally, it could be concluded that the air quality forecasting would be helpful to concerned authorities in providing the necessary information to the general public, to protect their health and take necessary precautionary measures.

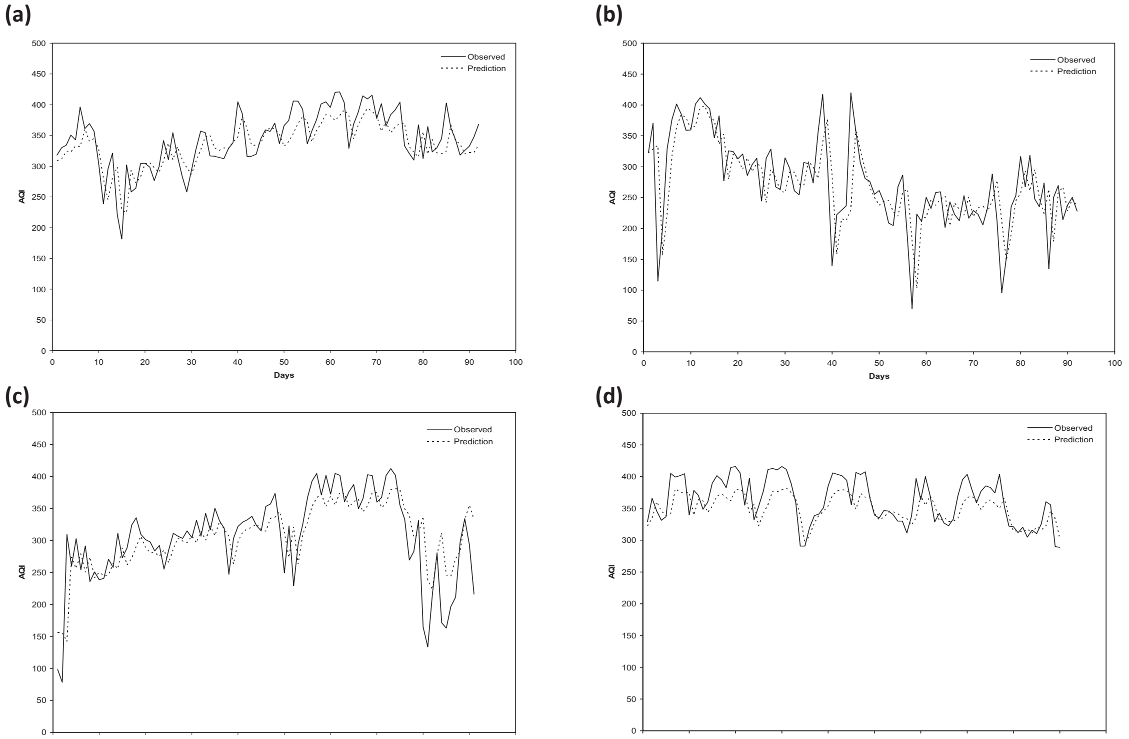

***Figure 4**. Comparison of observed and model predicted values of daily AQI in **(a)** Summer, **(b)** Monsoon, **(c)** Post Monsoon and **(d)** Winter seasons during the year 2006.*

***Table 5.** Comparison of PCR model predicted and observed values in years 2000-2005 and year 2006*

S.N.	Season	2000-2005				2006			
		RMSE	NMSE	Coefficient of determination	Fractional Bias	RMSE	NMSE	Coefficient of determination	Fractional Bias
1	Summer	35.91	0.0116	0.4902	0.0003	30.90	0.0082	0.5767	0.0229
2	Monsoon	53.94	0.0417	0.6241	0.0002	55.62	0.0418	0.4225	0.0094
3	Post Monsoon	40.44	0.0166	0.6467	0.0003	47.40	0.0241	0.5155	0.0019
4	Winter	41.52	0.1273	0.4096	0.0001	27.19	0.0058	0.5625	0.0360

Appendix

The statistical measures used for statistical evaluation of the performance of models were given by Chang and Hanna (2004) as follows:

Coefficient of Correlation (R). Coefficient of correlation (R) is relative measure of the association between the observed and predicted values. It can vary from 0 (which indicates no correlation) to $\underline{+}$1.0 (which indicates perfect correlation). A value of R close to 1.0 implies good agreement between the observed and predicted values, i.e. good model performance.

$$R=\frac{\overline{\left(C_o-\bar{C}_o\right)\left(C_p-\bar{C}_p\right)}}{\sigma_{C_p}\sigma_{C_o}}$$

Coefficient of Determination (R^2). Coefficient of determination (R^2), which is the square of coefficient of correlation, determines the proportion of variance that can be explained by the model.

Root Mean Square Error (RMSE). RMSE, is a measure of the differences between values predicted by a model and the observed values and is expressed as follows:

$$\mathrm{RMSE}=\sqrt{\overline{\left(C_o-C_p\right)^2}}$$

Normalized Mean Square Error (NMSE). NMSE, as a measure of performance, emphasizes the scatter in the entire data set and is defined as follows:

$$NMSE=\frac{\overline{\left(C_o-C_p\right)^2}}{\bar{C}_o.\bar{C}_p}$$

The normalization by $\bar{C}_o.\bar{C}_p$ ensures that NMSE will not be biased towards models that over–predict or under– predict. Ideal value for NMSE is zero. Smaller values of NMSE denote better model performance.

Fractional Bias (FB). It is a performance measure known as the normalized or fractional bias of the mean concentrations:

$$\mathrm{FB}=\frac{\left(\bar{C}_o-\bar{C}_P\right)}{0.5\left(\bar{C}_o+\bar{C}_P\right)}$$

where C_p are the model predictions, C_o are the observations, Overbar $\left(\bar{C}\right)$ is the average over the dataset, and σ_C is the standard deviation over the data set.

Supporting Material Available

The weights of the PC's for summer season (Table S1), The weights of the PC's for monsoon season (Table S2), The weights of the PC's for post monsoon season (Table S3), The weights of the PC's for winter season (Table S4).This information is available free of charge via the Internet at http://www.atmospolres.com.

References

Aneja, V.P., Agarwal, A., Roelle, P.A., Phillips, S.B., Tong, Q.S., Watkins, N., Yablonsky, R., 2001. Measurements and analysis of criteria pollutants in New Delhi, India. *Environment International* 27, 35-42.

Anfossi, D., Brisasca, G., Tinarelli, G., 1990. Simulation of atmospheric diffusion in low wind speed meandering conditions by a Monte Carlo dispersion model. *Il Nuovo Cimento* 13C, 995-1006.

Aron, R., 1984. Models for estimating current and future sulphur dioxide concentrations in Taipei. *Bulletin of Geophysics* 25, 47–52.

Aron, R., Aron, I. M., 1978. Statistical forecasting models: I. Carbon monoxide concentrations in the Los Angeles basin. *Journal of Air Pollution Control Association* 28, 681–684.

Boznar, M., Lesjak, M., Mlakar, P., 1993. A neural-network-based method for short-term predictions of ambient SO_2 concentrations in highly polluted industrial areas of complex terrain. *Atmospheric Environment Part B-Urban Atmosphere* 27, 221-230.

Brandon, C., Hommann, K., 1995. The cost of inaction: valuing the economy-wide cost of environmental degradation in India. *Proceedings of the Symposium on Global Sustainability,* United Nations University, Tokyo.

Central Pollution Control Board, 2005. PARIVESH highlights 2004-Delhi: CPCB.

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

Cogliani, E., 2001. Air pollution forecast in cities by an air pollution index highly correlated with meteorological variables. *Atmospheric Environment* 35, 2871-2877.

Comrie, A.C., 1997. Comparing neural networks and regression models for ozone forecasting. *Journal of the Air and Waste Management Association* 47, 653-663.

EPA, 1999. Air Quality Index Reporting; Final Rule, Federal Register, Part III, 40 CFR Part 58.

Finzi, G., Tebaldi, G., 1982. A mathematical model for air pollution forecast and alarm in an urban area. *Atmospheric Environment* 16, 2055-2059.

Goyal, P., Sidhartha, 2003. Present scenario of air quality in Delhi: a case study of CNG implementation. *Atmospheric Environment* 37, 5423-5431.

Hajek, P., Olej, V., 2009. Air quality indices and their modelling by hierarchical fuzzy inference systems. *WSEAS Transactions on Environment and Development* 10, 661-672.

Katsoulis, B.D., 1988. Some meteorological aspects of air pollution in Athens, Greece. *Meteorology and Atmospheric Physics* 39, 203-212.

Lin, G.Y., 1982. Oxidant prediction by discriminant analysis in the south coast air basin of California. *Atmospheric Environment* 16, 135-143.

Mantis, H.T., Repapis, C.C., Zerefos, C.S., Ziomas, J.C., 1992. Assessment of the potential for photochemical air pollution in Athens - a comparison of emissions and air pollutant levels in Athens with those in Los Angeles. *Journal of Applied Meteorology* 31, 1467-1476.

McCollister, G. M., Wilson, K. R., 1975. Linear stochastic models for forecasting daily maxima and hourly concentrations of air pollutants. *Atmospheric Environment (1967)* 9, 417-423.

Milionis, A.E., Davies, T.D., 1994. Regression and stochastic models for air pollution. 1. Review, comments and suggestions. *Atmospheric Environment* 28, 2801-2810.

Nagendra, S.M.S., Venugopal, K., Jones, S.L., 2007. Assessment of air quality near traffic intersections in Bangalore city using air quality index. *Transportation Research Part D: Transport and Environment* 12, 167-176.

Polydoras, G.N., Anagnostopoulos, J., Bergeles, G.C., 1998. Air quality predictions: dispersion model vs Box-Jenkins stochastic models. An implementation and comparison for Athens, Greece. *Applied Thermal Engineering* 18, 1037-1048.

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 – 411 005.

Robeson, S.M., Steyn, D.G., 1990. Evaluation and comparison of statistical forecast models for daily maximum ozone concentrations. *Atmospheric Environment Part B-Urban Atmosphere* 24, 303-312.

Sanchez, M.L., Pascual, D., Ramos, C., Perez, I., 1990. Forecasting particulate pollutant concentrations in a city from meteorological variables and regional weather patterns. *Atmospheric Environment*

Part A-General Topics 24, 1509-1519.

Sanchez, M.L., Casanova, J.L., Ramos, M. C., Sanchez, J.L., 1986. Studying the spatial and temporal distribution of SO_2 in an urban area by principal component factor analysis. *Atmospheric Research* 20, 53-65.

Shi, J.P., Harrison, R.M., 1997. Regression modelling of hourly NO_X and NO_2 concentrations in urban air in London. *Atmospheric Environment* 31, 4081-4094.

Sousa, S.I.V., Martins, F.G., Alvim-Ferraz, M.C.M., Pereira, M.C., 2007. Multiple linear regression and artificial neural networks based on principal components to predict ozone concentrations. *Environmental Modelling* and Software 22, 97-103.

Van den Elshout, S., Leger, K., Nussio, F., 2008. Comparing urban air quality in Europe in real time - a review of existing air quality indices and the proposal of a common alternative. *Environment International* 34, 720-726.

Ziomas, I.C., Melas, D., Zerefos, C.S., Bais, A.F., Paliatsos, A.G., 1995. Forecasting peak pollutant levels from meteorological variables. *Atmospheric Environment* 29, 3703-3711.