Analyzing and predicting the criteria pollutants over a tropical urban area by using statistical models

***S. Dey[1], P. Sibanda[1], S. Gupta[2] and A. Chakraborty[3]**

[1]School of Mathematics, Statistics and Computer Science, University of KwaZulu-Natal, Private Bag X01 Scottsville 3209, Pietermaritzburg, South Africa

[2]Department of Environmental Science, The University of Burdwan, Golapbag, Burdwan 713104, West Bengal, India

[3]Center for Rivers, Oceans, Atmosphere and Land Sciences (CORAL), Indian Institute of Technology, Kharagpur,
Kharagpur -721302, West Bengal, India

*Corresponding author: sharadiadey1985@gmail.com

Abstract

Modeling and prediction of criteria pollutants over the urban areas is essential for the formulation and improvisation of urban air quality management strategies. Various statistical techniques have been employed worldwide for accurate prediction of the air pollutants. This study focuses on the analysis and prediction of the criteria pollutants over a tropical urban area (Durgapur, 23° 30′ 34.58″ N and 87° 21′ 03.42″ E) performed by using statistical models *viz.* multiple linear regression (MLR) and principal component regression (PCR). Multiple linear regression analyses have been performed using the original variables and principal components (PCs) as the inputs. On the basis of the performance indicators, MLR model is found to perform better than the PCR in most cases. The R^2 values obtained by MLR are 0.962, 0.945, 0.898, 0.937, 0.603, 0.874, 0.871, 0.837, 0.858, 0.868, 0.842 and 0.825 for PM_{10}, $PM_{2.5}$, sulphur dioxide, nitrogen dioxide, carbon monoxide, ammonia, ozone, benzene, benz(a)pyrene, arsenic, lead and nickel respectively which are greater than the respective R^2 values obtained by PCR model. Results of the two models reveal that use of PCA could not enhance the MLR performance. The predictive equations proposed by the statistical models suggest that the meteorological parameters (temperature, relative humidity, wind speed and cloud cover) have significant influence on the concentration of the criteria pollutants.

Key words: PCR, MLR, Performance Indicators, criteria pollutants

1. Introduction

Escalating air pollution and deteriorating air quality status of urban areas is a matter of concern worldwide. In this era of rapid urbanization and industrialization, air pollutants containing toxic substances like particulate matters, heavy metals, polycyclic hydrocarbons (PAH), volatile organic compound (VOC) and other gaseous substances (like SO_2, NO_2, CO, NH_3, tropospheric O_3 etc.) have an increasing impact on urban air quality. Actually, air pollution risk is a function of the hazard of the pollutant and exposure to the pollutant. Carbon monoxide, lead, nitrogen dioxide, ozone, particulate matter, and sulfur dioxide have identified as criteria pollutants by Clean Air Act (CAA) of 1970. Central Pollution Control Board (CPCB) has identified 12 health based parameters [namely particulate matters (PM_{10} & $PM_{2.5}$), benzene, benzo(a)pyrene, nitrogen dioxide (NO_2), sulphur dioxide (SO_2), carbon monoxide (CO), ammonia (NH_3),ozone (O_3), lead (Pb), nickel (Ni) and arsenic (As)] for assessing the air quality status across the country in 2009 under the provision of Air (Prevention & Control of Pollution) Act, 1981.

The complexities and difficulties in continuous measurement of air pollutant concentrations have led to the development of modeling techniques which enable the researchers to predict the pollutant concentration with acceptable accuracy [1]. Accurate knowledge of pollutant sources, emission inventories and proper description of the physico - chemical processes are essential for minimizing biasness and errors of the outputs of the deterministic models. These are quick and easy empirical techniques for predicting the ambient air pollutant concentration as a function of several input parameters. In air quality modeling, one of the most common models available for predicting outdoor and indoor air pollutant concentrations are statistical regression methods [2]. Statistical models are suitable for the description of the complex site-specific relationship between air pollutants and explanatory variables, and they often make predictions with a higher accuracy than mechanistic models [3]. Multiple linear regression (MLR) is a widely used multivariate statistical technique for expressing the dependence of a response variable on several independent (predictor) variables. Awang *et al.* [4] compared the multivariate methods (MLR and PCR) for predicting the surface O_3 concentration during daytime, nighttime and critical conversion time in Shah Alam, Malaysia. The concentration PM_{10}, $PM_{2.5}$, CO and CO_2 concentrations and meteorological variables (wind speed, air temperature, and relative humidity) were employed by Elbayoumi *et al*. [5] for predicting the annual and seasonal indoor concentration of PM_{10} and $PM_{2.5}$ at Gaza Strip (Palestine) using multivariate statistical methods. Luvsan *et al.* [6] used multiple linear regression models for exploring the association of concentration of SO2 with temperature, relative humidity and wind speed in Mongolia. Sayegh et al.[7] employed several approaches including linear, non-linear, and machine learning methods are evaluated for the prediction of urban PM_{10} concentrations in the City of Makkah, Saudi Arabia.

In the present work, we predict the concentration of various criteria pollutants by using multiple linear regression (MLR) and principal component regression (PCR) models, the performance of both the statistical models is evaluated in terms of the performance indicators. Deterministic models require a large number of input data which are difficult to provide whereas statistical models are relatively simple and sufficiently reliable tools for predicting the concentration of different air pollutants. Moreover, application of multivariate statistical methods for the prediction of the air pollutants is a new piece of work over this eastern part of India.

2. Method

2.1 Description of the study area

Durgapur (chosen urban area) is situated in the Burdwan district of West Bengal, India. It is located on the bank of River Damodar. This area is covered with Red and Yellow Ultisols soil and the topography of this area is undulating, with an average elevation of 65 m MSL. This area experiences a transitional climate between the tropical wet and dry climate and the more humid subtropical climate.

2.2 Data used

The data of concentration of all the criteria pollutants such as ammonia, arsenic, benzene, benzo(α)pyrene, carbon monoxide, lead, nickel, nitrogen dioxide, ozone, sulphur dioxide, PM_{10} and $PM_{2.5}$ at Bidhannagar, India (23° 30′ 34.58″ N and 87° 21′ 03.42″ E) were collected for the duration of June, 2013 to May, 2015 from the archived data set of WBPCB (Bidhannagar unit of Durgapur). These parameters are monitored twice a week at this location by WBPCB [www.wbpcb.gov.in]. The data of meteorological parameters [Temperature (T), relative humidity (RH), wind speed (WS) and cloud cover (CC) are collected from the NOAA Air Resources Laboratory (ARL) website. (http://ready.arl.noaa.gov/READYamet.php).

The air pollutants and the meteorological parameters data were divided into two sets: model development set and the model validation set. The model development set comprises of the 24 average values of criteria air pollutants and meteorological parameters recorded from June, 2013 to December, 2014 while the data set of January, 2015 to May, 2015 is used for data validation. The accuracy and errors in the MLR and PCR models were evaluated in terms of performance indicators (PIs)

2.3 Statistical analysis

Data analysis was carried using the statistical software XLSTAT 2015. Step wise multiple regression (MLR) and Principle Component Regression (PCR) analyses have been used for finding the predictive equations of the criteria pollutants.

2.3.1 Principal component analysis (PCA)

Among multivariate techniques, Principal Components Analysis (PCA) is designed to classify variables based on their correlations with each other. The goal of PCA and other factor analysis procedures, is to consolidate a large number of observed variables into a smaller number of factors (components) that can be more readily interpreted as these underlying processes. It is often used as an exploratory tool to identify the major sources of air pollutant emissions [8] [9]. For physical interpretation of the components, loadings of variables on the component are estimated. Loading represents the degree and direction of relationship of the variables with a factor. An analysis of the PC loadings on the chosen variables allows the identification of the PCs as pollution sources affecting the data. The number of factors (PCs) is selected such that the cumulative percentage variance explained by all the chosen factors is more than 70%. As the normalized variables each carry one unit of variance, so the factors with eigen value more than 1 are chosen for the study. The factors with eigen values less than one are discarded as they are assumed to contain less information [10]. To undertake PCA, the XLSTAT 2015 statistical software was used, specifying the principal components method with varimax rotation [11]. The rotation of the component axis is performed so that components are clearly defined by high loadings for some variables and low loadings for others, facilitating the interpretation in terms of original variables.

The principal components of the predictor variables are obtained using a data reduction technique by means of finding linear combinations of the original variables. In general, PCs are expressed by the following equation

$$PC_i = A_{1i} V_i + A_{2i} V_2 + \ldots\ldots + A_{ni} V_n \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

where,

PC_i is principal component i and

A_{ni} is the loading (correlation coefficient) of the original variable V_n.

As the scores of high loading components with an eigen value greater than or equal to 1 account for most of the variations in the data set, it is ideal to use them as independent or predictor variables in regression analysis. Thus, principal component regression (PCR) establishes relationship between dependent variables and the selected PCs of the independent variables [12].

2.3.2 Multiple Linear Regression (MLR)

Multiple linear regression attempts to model the relationship between two or more explanatory variables (independent variables) and a response variable (dependent variable) by fitting a linear equation to observed data. This multivariate statistical technique finds wide application in the field of atmospheric science, especially air pollution studies. The MLR

technique has the capability of exploring the contribution of selected variables to chosen air pollutant concentration. The general equation of MLR is expressed as [12]

$$y = b_0 + \sum_{k=1}^{p} \mathbf{b_i}\, \boldsymbol{x_i} + \xi \quad \text{.....(2)}$$

Where,
b_i is the regression coefficient,
x_i is the independent variable, and
ξ is the stochastic error associated with the regressions.

2.3.3 Principal Component Regression (PCR)

Principal Component Regression (PCR) is a combination of Principal Component Analysis (PCA) and Multiple Linear Regression (MLR). The PCs obtained in PCA are used as the inputs in MLR. The selected variables with high loadings from PCA ensure the inclusion of the majority of the original variances in the statistical model and they are ideal for use as independent variables in MLR **[12]**. The use of PCs as the independent variables of MLR reduces the problem of multicolinearity.

2.3.4 Performance Indicators (PIs)

The performance of MLR and PCR models are assessed on the basis of the performance indicators (PIs). Good prediction models should have minimal errors (closer to 0 for NAE and RMSE) and high accuracy (closer to 1 for IA, PA, and R^2). The following PIs are used in this study -

- Normalized Absolute error (NAE) - It measures the average difference between predicted and observed values in all cases divided by observed values **[5]** and is expressed as:

$$\text{NAE} = \frac{\sum_{i=1}^{n} |P_i - O_i|}{\sum_{i=1}^{n} O_i} \quad \text{.....(3)}$$

where *n* is the sample size, P_i is the predicted concentration of the criteria pollutant and O_i is the observed value of the pollutant concentration.

- Root Mean Square Error (RMSE) - It measures the success of numerical prediction.

RMSE is calculated by the equation [13] [14]

$$\text{RMSE} = \sqrt{\frac{1}{n-1} \sum_{i=1}^{n} (P_i - O_i)^2} \quad \text{.....(4)}$$

where n is the number of sample, O_i is the observed concentrations of the pollutants and P_i is the predicted concentration of the pollutants.

- Prediction accuracy (PA) - The prediction accuracy is computed using by the following equation [15]:

$$\text{PA} = \frac{\sum_{i=1}^{n} (P_i - \bar{P})^2}{\sum_{i=1}^{n} (O_i - \bar{O})^2} \quad \text{.....(5)}$$

where n is the number of sample, O_i is the observed concentrations of the pollutants and P_i is the predicted concentration of the pollutants.

- Index of agreement (IA) - a measure of accuracy, was calculated using Equation (6) [16].

$$\text{IA} = 1 - \left[\frac{\sum_{i=1}^{n} (P_i - O_i)^2}{\sum_{i=1}^{n} (|P_i - \bar{O}| + |O_i - \bar{O}|)^2} \right] \quad \text{.....(6)}$$

where n is the number of sample, O_i is the observed concentrations of the pollutants and P_i is the predicted concentration of the pollutants

- Coefficient of determination (R^2) - The coefficient of determination explains how much the variability in the predicted data can explain by the fact that they are related to the observed values. R^2 is expressed by the following equation [15] :

$$R^2 = \left(\frac{\sum_{i=1}^{n} (P_i - \overline{P})(O_i - \overline{O})}{n\, S_{pred}\, S_{obs}} \right)^2 \quad \text{..... (7)}$$

where n is the number of sample, O_i is the observed concentrations of the pollutants and P_i is the predicted concentration of the pollutants, $\overline{P}$ is the average of predicted value , $\overline{O}$ is the average of observed values, S_{pred} is a standard deviation of the predicted pollutant concentration, S_{obs} is a standard deviation of the observed pollutant concentration.

3. Result and discussion

3.1 MLR model development

MLR modeling (stepwise method) has been performed for finding the predictive equations of the criteria pollutants with the regression assumptions approximately satisfied. During this statistical analysis, the distribution of residuals was approximately with zero mean and constant variance. Variance Inflation Factor (VIF) was mostly below 10 except on very few occasions when the VIF value exceeded 10. Therefore, the MLR predictor variables have negligible collinearity problem.

3.2 PCR model development

PCA was applied for variable reduction and for providing most relevant variable for understanding the pollutant variation. Varimax rotation was applied in PCA for maximizing the loading of a predictor variable on one component. The adequacy of input data for the PCA was assessed using the Kaiser–Meyer–Olkin (KMO) test. The results obtained from application of KMO test on the input data set were more than 0.5 which indicated that the input data set were sufficient for PCA.

Before extraction using PCA, 16 linear components (twelve criteria pollutants, temperature, humidity, cloud cover and wind speed) were used. After performing PCA, three linear factors were considered as principal components (PCs) on the basis of their eigen values. In PCA, the eigen value provides the amount of variation explained by each PC. As the normalized variables each carry one unit of variance, the factors with eigen value more than 1 were chosen for the study. The factors with eigen values less than one are discarded as they provide less information [10]. Occasionally, eigen values smaller than unity are considered as they are very close to one **[17].**The variability of PCs obtained after varimax rotation are summarized in Table 1.The obtained PCs are used as the independent variables (explanatory variables) and the original criteria pollutant as the dependent variables in stepwise multiple linear regression analysis in PCR model. The use of PCs as input in MLR is intended to reduce the complexity and multicollinearity problems of the models.

Table 1. Total variance for different criteria pollutants after varimax rotation

Sl.No.	Parameter	Components	Eigen value	Variability (%)	Cumulative %
1	PM_{10}	PC1	8.880	35.416	35.416
		PC2	1.938	15.687	51.103
		PC3	1.064	23.157	74.260
2	$PM_{2.5}$	PC1	8.962	34.053	34.053
		PC2	1.933	15.327	49.379
		PC3	1.094	25.556	74.935
3	Sulphur dioxide	PC1	8.960	29.705	29.705
		PC2	1.959	14.692	44.397
		PC3	1.091	30.666	75.063
4	Nitrogen dioxide	PC1	8.922	32.114	32.114
		PC2	1.903	15.527	47.641
		PC3	1.087	26.808	74.449
5	Carbon monoxide	PC1	9.596	33.732	33.732
		PC2	1.550	31.157	64.889
		PC3	1.031	11.214	76.104
6	Ammonia	PC1	9.054	32.690	32.690
		PC2	1.897	15.845	48.535
		PC3	1.067	26.579	75.114
7	Ozone	PC1	9.115	30.682	30.682
		PC2	1.848	15.140	45.822
		PC3	1.110	29.638	75.460
8	Benzene	PC1	9.128	40.932	40.932
		PC2	1.960	22.655	63.588
		PC3	0.941	11.591	75.178
9	Benz(a)Pyrene	PC1	9.162	39.499	39.499
		PC2	1.951	23.602	63.101
		PC3	0.999	12.597	75.699
10	Arsenic	PC1	8.962	33.076	33.076
		PC2	1.952	15.137	48.213
		PC3	1.090	26.814	75.027
11	Lead	PC1	9.434	31.759	31.759
		PC2	1.620	12.760	44.520
		PC3	1.103	31.461	75.980
12	Nickel	PC1	9.018	33.232	33.232
		PC2	1.959	15.636	48.868
		PC3	1.097	26.593	75.461

3.3 Comparison of MLR and PCR models

MLR and PCR models provide an estimate of 24 hour average concentration of all the criteria pollutants (Table 2).

Table 2. Summary of models of all the criteria pollutants using Multiple Linear Regression (MLR) and Principal Component Regression (PCR)

Sl.No.	Parameter	Method	R^2	Model
1	PM_{10}	MLR	0.962	PM_{10} = 0.135+4.138*As + 17.633*BAP + 1.788*Ni + 1.156*$PM_{2.5}$
		PCR	0.918	PM_{10} = 102.688 + 23.684*PC1 + 17.671PC2 + 35.397*PC3
2	$PM_{2.5}$	MLR	0.945	$PM_{2.5}$ = 6.22 - 4.01*BAP -0.13*O_3 + 0.529*PM_{10} + 1.745*SO2 - 2.455*WS
		PCR	0.852	$PM_{2.5}$ = 59.280 + 11.535*PC1 + 12.138*PC2 + 18.645*PC3
3	Sulphur dioxide (SO_2)	MLR	0.898	SO_2 = 1.184 + 0.093*NH_3 - 0.283*As +3.553*Pb + 0.108*NO_2 - 0.042*O_3+ 0.018*$PM_{2.5}$
		PCR	0.779	SO_2 = 8.22 + 1.01*PC1 + 0.746*PC2 + 1.231*PC3
4	Nitrogen dioxide (NO_2)	MLR	0.937	NO_2 = 28.993 - 20.994*CO + 0.298*O_3 + 3.837*SO_2 - 0.723*RH
		PCR	0.888	NO_2 = 53.871 + 9.939*PC1 +1.281*PC2 + 11.993*PC3
5	Carbon monoxide (CO)	MLR	0.603	CO = 0.743 + 0.498*Pb - 0.005*Ni - 0.004*T
		PCR	0.439	CO = 0.665 + 0.017*PC2 + 0.046*PC3
6	Ammonia (NH_3)	MLR	0.874	NH_3 = 3.957 + 1.039*As - 2.153*C_6H_6 + 7.971*CO - 17.578*Pb + 0.410*Ni +0.142*O3 + 1.085*SO_2
		PCR	0.799	NH_3 = 25.773 + 2.498*PC1 + 4.244*PC3
7	Ozone (O_3)	MLR	0.871	O_3 = 14.02 + 1.255*NH_3 + 4.316*As + 0.659*NO_2 -4.083*SO_2 - 0.066*CC
		PCR	0.77	O_3 = 53.433 + 7.695*PC1 - 1.579*PC2 + 12.798*PC3
8	Benzene (C_6H_6)	MLR	0.837	C_6H_6 = 1.093 + 0.536*BAP - 0.447*CO + 0.002*PM_{10}
		PCR	0.698	C_6H_6 = 1.352 + 0.229*PC1 + 0.21*PC2
9	Benzo(a)pyrene (BAP)	MLR	0.858	BAP = -0.643 + 0.067*As + 0.755*C_6H_6 + 0.003*PM10 - 0.038*SO_2
		PCR	0.71	BAP = 0.579 + 0.265*PC1 +0.318*PC2
10	Arsenic (As)	MLR	0.868	As = -0.306 + 0.05*NH_3 + 0.538*BAP - 2.130*CO + 3.853*Pb - 0.086*Ni + 0.026*O_3 + 0.006*PM_{10} + 0.106*WS
		PCR	0.796	As = 2.000 + 0.674*PC1 +0.114*PC2 + 0.695*PC3
11	Lead (Pb)	MLR	0.842	Pb = -0.457 - 0.009*NH_3 + 0.05*As - 0.054*BAP + 0.646*CO +0.015*Ni + 0.001*PM_{10} + 0.01*SO_2 +0.007*RH - 0.01*WS
		PCR	0.59	Pb = 0.163 + 0.076*PC2 + 0.056*PC3
12	Nickel (Ni)	MLR	0.825	Ni = 2.925 + 0.205*NH_3 - 0.949*As - 9.310*CO + 14.681*Pb + 0.049*O_3 + 0.037*PM_{10}
		PCR	0.749	Ni = 8.814 + 1.769*PC1 + 1.158*PC2 + 2.197*PC3

* Temperature (T), relative humidity (RH), wind speed (WS) and cloud cover (CC)

The MLR models were found to perform better than the corresponding PCR models as the R^2 values of the MLR models are higher than those of PCR models (Table 2). The predictive equations suggested by the statistical models suggest that meteorological factors (temperature, relative humidity, cloud cover and wind speed) play an important role in the prediction of concentration of the criteria pollutants. For example, cloud cover is negatively associated with ozone concentration in the predictive equation proposed by the MLR model which is in agreement with the mechanism of photochemical formation of tropospheric ozone. In general, high wind speed flushes out the air pollutants near the earth's surface thereby leading to lower concentration of ground level air pollutants. Such a result is reflected in the predictive equations of the MLR model. The PCR has more degrees of freedom and offers variable combinations for the principal components in choosing multiple components but the use of PCs as the inputs in the MLR could not improve the performance of the model. Actually, the PCA is an unsupervised dimension reduction methodology which does not consider the correlation among the dependent and independent variables. This might be a reason for the

failure of the PCR model. Elbayoumi *et al*. [5] also concluded that the use of PCR could not improve the accuracy in predicting indoor PM_{10} and $PM_{2.5}$ in the Gaza Strip (Palestine) over MLR. Awang *et al*. [4] also reported the optimal performance of MLR model for daytime ground level ozone in terms of normalized absolute error, index of agreement, prediction accuracy, and coefficient of determination (R^2). The R^2 for the correlation between the observed and the predicted concentration of the criteria pollutants for MLR and PCR models are shown in Figures 1 to 4. The performances of the two models are further compared on the basis of the performance indicators namely normalized absolute error (NAE), root mean square error (RMSE), prediction accuracy (PA), index of agreement (IA) and coefficient of determination (R^2) (Table 3). Good prediction models should have minimal errors (closer to 0 for NAE and RMSE) and high accuracy (closer to 1 for IA, PA, and R2). On the basis of this principle, MLR models for prediction of air pollutants are found to give better performance than the corresponding PCR model.

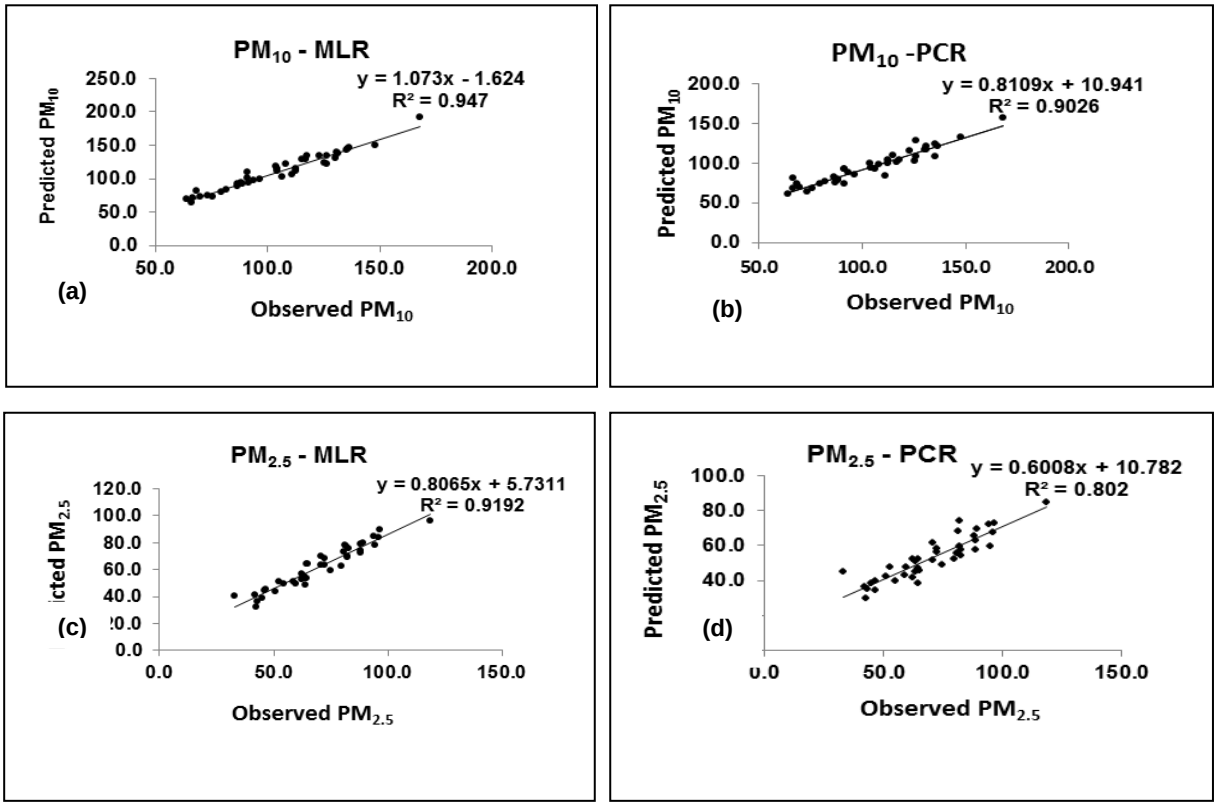
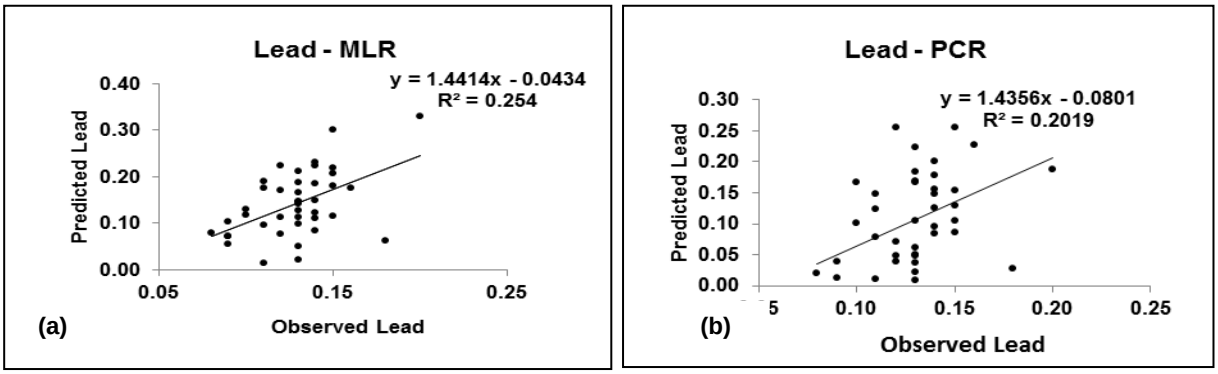


Figure 1. Scatter plots of observed and predicted values of (a) PM_{10} by MLR method, (b) PM_{10} by PCR method, (c) $PM_{2.5}$ by MLR method, (d) $PM_{2.5}$ by PCR method



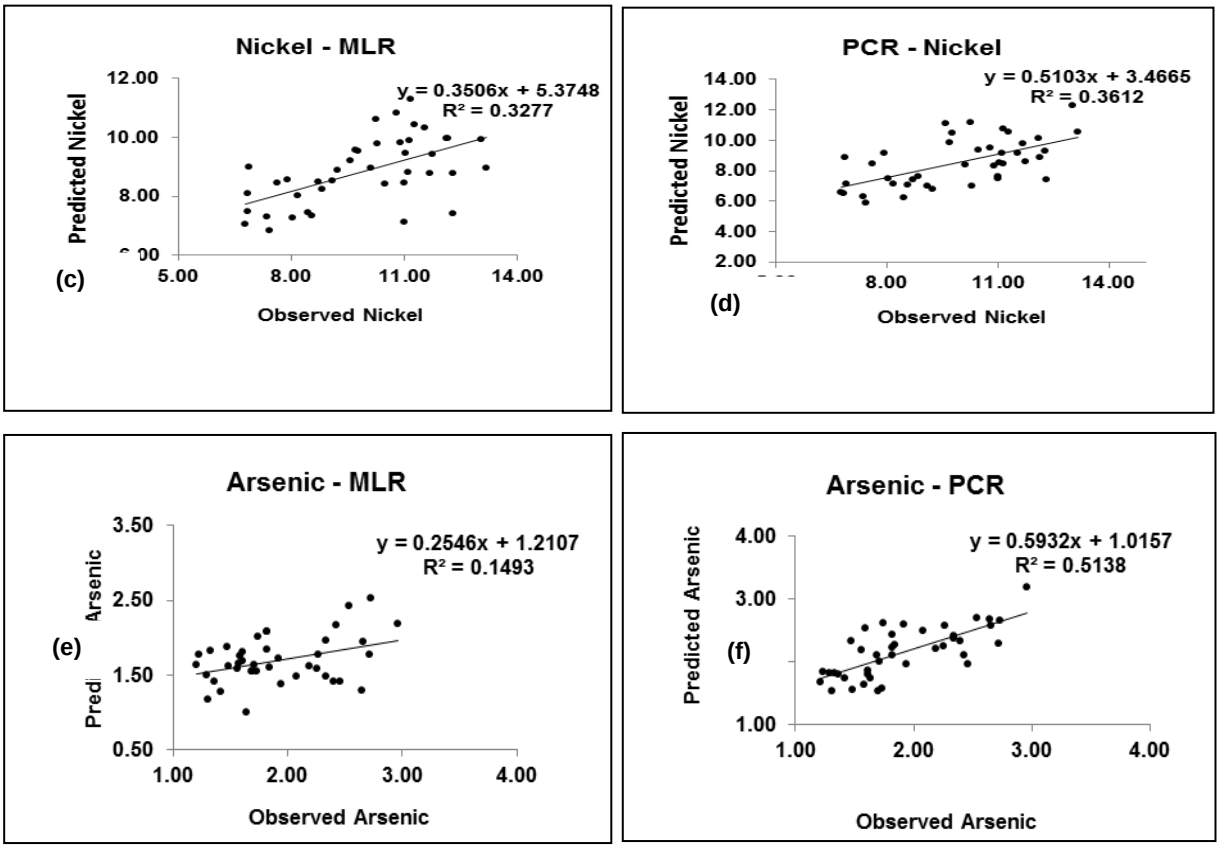
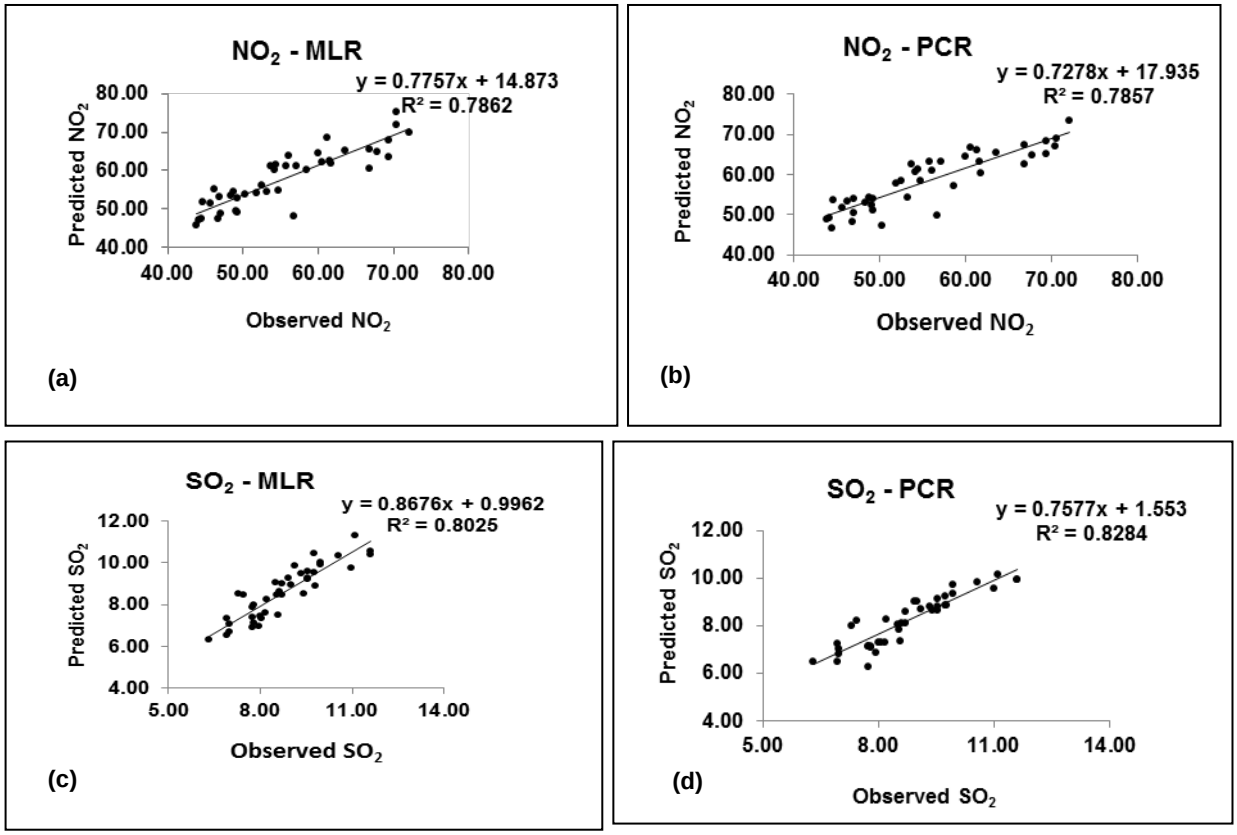


Figure 2. Scatter plots of observed and predicted values of (a) Lead (Pb) by MLR method, (b) Lead (Pb) by PCR method, (c) Nickel (Ni) by MLR method, (d) Nickel (Ni) by PCR method, (e) Arsenic (As) by MLR method and (f) Arsenic (As) by PCR method



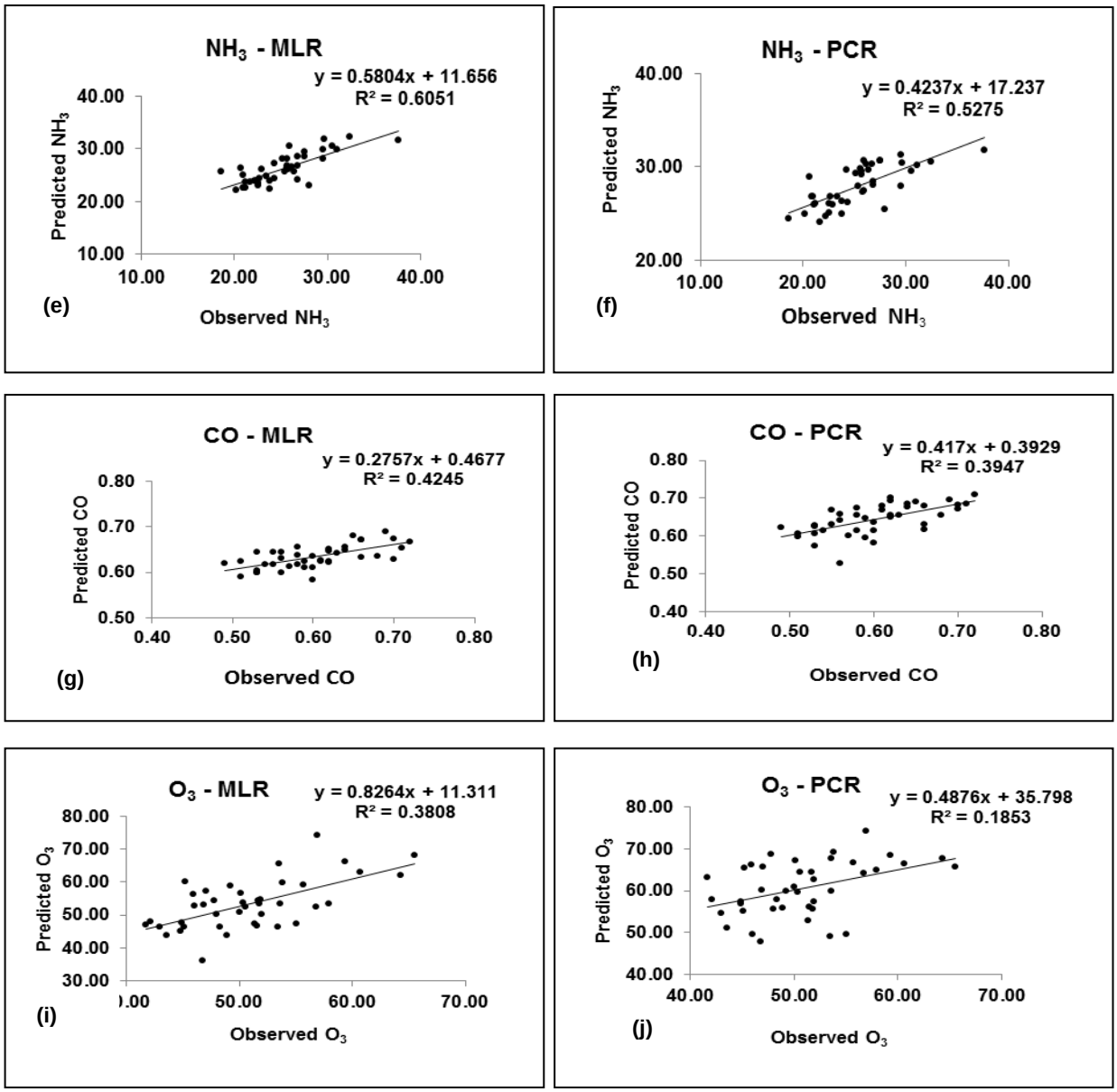


Figure 3. Scatter plots of observed and predicted values of (a) NO_2 by MLR method, (b)NO_2 by PCR method, (c) SO_2 by MLR method, (d) SO_2 by PCR method, (e) NH_3 by MLR method , (f) NH_3 by PCR method, (g) CO by MLR method, (h) CO by PCR method, (i) O_3 by MLR method and (j) O_3 by PCR method

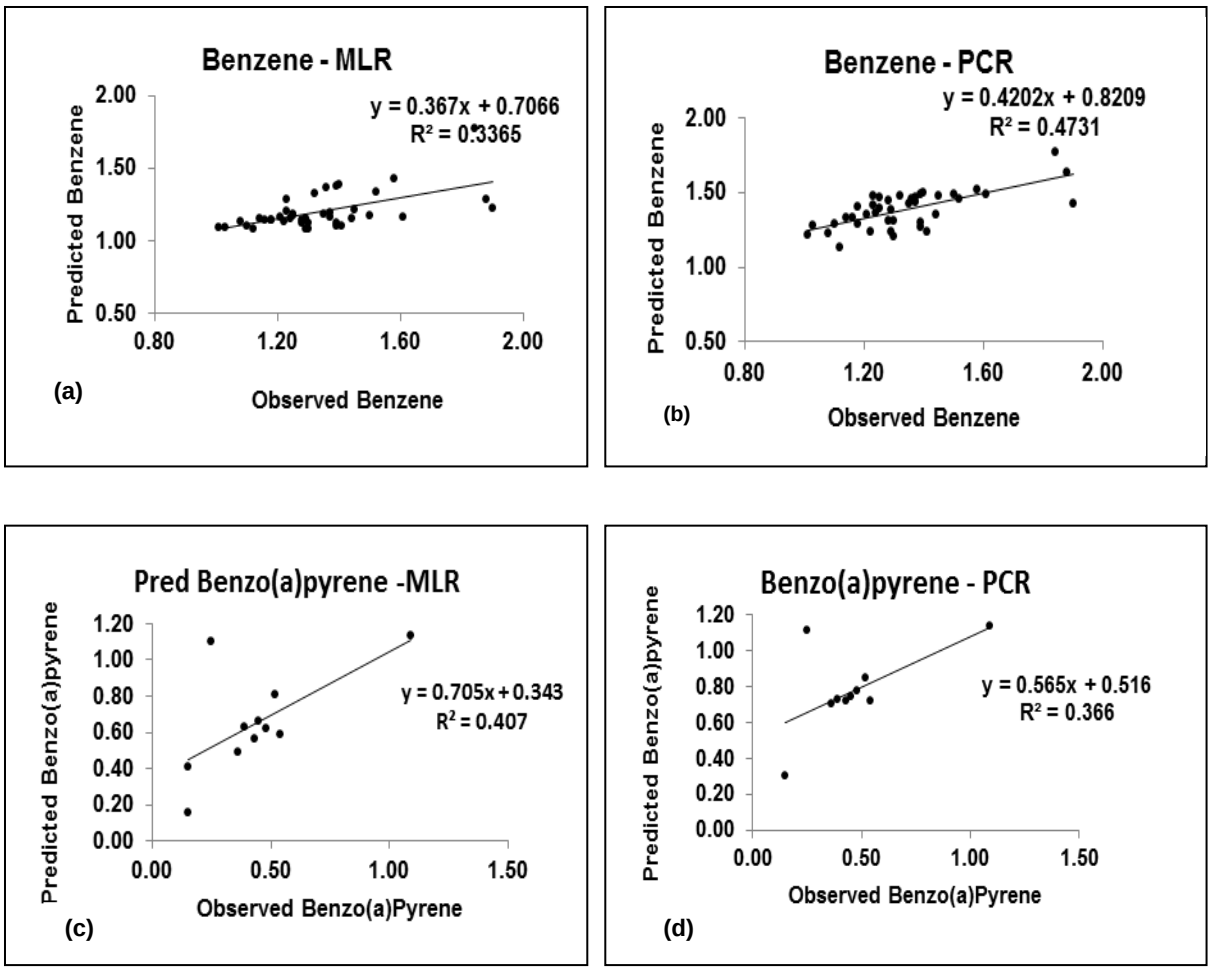


Figure 4. Scatter plots of observed and predicted values of (a) Benzene by MLR method, (b)Benzene by PCR method, (c) Benzo(a)pyrene by MLR method and (d) Benzo(a)pyrene by PCR method

It appears from Table 3 that the error indicators (NAE and RMSE) are minimum and accuracy indicators (IA, PA and R^2) are maximum in case of each criteria pollutant by using MLR model (except Benzene and Arsenic). This observation suggests that the physico-chemical characteristics and the interaction of Benzene and Arsenic with other substances in the atmosphere should be explored for understanding these outcomes of these statistical models.

Table 3. Summary of performance indicators (PIs) of the models

Sl.No.	Parameters	Method	NAE	RMSE	IA	PA	R^2
1	PM_{10}	MLR	0.068	8.981	0.971	0.823	0.902
		PCR	0.097	12.126	0.934	0.728	0.859
2	$PM_{2.5}$	MLR	0.117	9.8	0.924	0.708	0.875
		PCR	0.254	19.718	0.725	0.451	0.763
3	Sulphur dioxide	MLR	0.054	0.613	0.941	0.938	0.764
		PCR	0.076	0.799	0.891	0.693	0.789
4	Nitrogen dioxide	MLR	0.068	4.634	0.915	0.765	0.748
		PCR	0.076	4.889	0.902	0.674	0.748
5	Carbon monoxide	MLR	0.076	0.057	0.635	0.181	0.423
		PCR	0.088	0.063	0.643	0.438	0.732
6	Ammonia	MLR	0.075	2.59	0.839	0.557	0.576
		PCR	0.130	3.752	0.678	0.34	0.504
7	Ozone	MLR	0.102	6.593	0.744	0.558	0.362
		PCR	0.202	11.842	0.488	0.779	0.176
8	Benzene	MLR	0.115	0.219	0.645	0.400	0.32
		PCR	0.034	0.157	0.762	0.373	0.453
9	Benz(a)Pyrene	MLR	1.326	0.361	0.529	0.702	0.278
		PCR	1.899	0.469	0.428	0.955	0.283
10	Arsenic	MLR	0.205	0.511	0.605	0.433	0.141
		PCR	0.170	0.415	0.784	0.685	0.488
11	Lead	MLR	0.379	0.062	0.52	0.124	0.251
		PCR	0.484	0.074	0.448	0.100	0.196
12	Nickel	MLR	0.138	1.863	0.670	0.375	0.312
		PCR	0.178	2.092	0.690	0.721	0.344

4. Conclusion

In this study, multiple linear regression analyses have been performed using the original variables and principal components (PCs) as the inputs. MLR can encounter the complexity of multicollinearity as the environmental variables are correlated to each other. MLR using the PCs as the inputs is known as principal component regression (PCR) and the use of this technique is expected to reduce the problem of multicollinearity. Both models provide an estimate of 24 hour average concentration of all the criteria pollutants. On the basis of the performance indicators, the MLR model was found to perform better than the PCR in most cases (except Benzene and Arsenic). Analysis of the physico - chemical properties and mode of interaction of Benzene and Arsenic with other substances present in the ambient

environment may further clarify the characteristics of these two criteria pollutants. Meteorological parameters, particularly temperature, relative humidity and cloud cover are found to influence the concentration of the air pollutants over that region. The use of characteristics of boundary layer processes and traffic may further improve the accuracy of prediction of the criteria pollutants over urban areas.

References

[1] Chaloulakou, A. and Mavroidis, I. (2002) Comparison of indoor and outdoor concentrations of CO at a public school. Evaluation of an indoor air quality model. *Atmospheric Environment* **36**, 1769 – 1781.

[2] Özbay, B. (2012) Modeling the effects of meteorological factors on SO_2 and PM_{10} concentrations with statistical approaches, *CLEAN- Soil, Air, Water* **40**, 571 – 577.

[3] Hrust, L., Klai´c, Z.B., Križan, J., Antoni´c, O. and Hercog, P. (2009) Neural network forecasting of air pollutants hourly concentrations using optimised temporal averages of meteorological variables and pollutant concentrations, *Atmospheric Environment* **43**, 5588–5596.

[4] Awang, N. R., Ramli, N. A., Yahaya, A. S. and Elbayoumi, M. (2015) Multivariate methods to predict ground level ozone during daytime, nighttime, and critical conversion time in urban areas, *Atmospheric Pollution Research* **6**, 726 - 734.

[5] Elbayoumi, M., Ramli, N.A., Yusof, N.F.F.M., Yahaya, A.S.B., Madhoun, W.A. and Ul-Saufie, A.Z. (2014). Multivariate methods for indoor PM_{10} and $PM_{2.5}$ modelling in naturally ventilated schools buildings, *Atmospheric Environment* **94**, 11 – 21.

[6] Luvsan, M.E., Shie, R.H., Purevdori, T., Badarch, L., Baldorj, B. and Chan, C.C. (2012) The influence of emission sources and meteorological conditions on SO_2 pollution in Mongolia. *Atmospheric Environment* **61**, 542 – 549.

[7] Sayegh, A. S., Munir, S. and Habeebullah, T. M (2014) Comparing the Performance of Statistical Models for Predicting PM10 Concentrations, *Aerosol and Air Quality Research* **14**, 653–665.

[8] Bruno, P., Caselli, M., Gennaro, G. and Traini, A. (2001) Source apportionment of gaseous atmospheric pollutants by means of an absolute principal component scores (APCS) receptor model, *Fresenius Journal* of *Analytical Chemistry* **371**, 1119 – 1123.

[9] Guo, H., Wang, T. and Louie P.K.K. (2004) Source apportionment of ambient non-methane hydrocarbons in Hong Kong: Application of a principal component analysis/absolute principal component scores (PCA/APCS) receptor model, *Environmental Pollution* **129**, 489 – 498.

[10] Maenhaut, W. and Cafmeyer, J. (1987) Particle induced X-ray emission analysis and multivariate techniques: An application to the study of the sources of respirable atmospheric particles in Gent, Belgium, *Journal of* Trace *and* Microprobe *Techniques* **5**, 135 – 158.

[11] Kaiser, H. F. (1958) The varimax criterion for analytic rotation in factor analysis, *Psychometrika* **23**, 187 – 200.

[12] Gvozdic, V., Kovac – Andric, E. and Brana, J.(2011) Influence of meteorological factors NO_2, SO_2, CO and PM_{10} on the concentration of O_3 in the urban atmosphere of Eastern Croatia, *Environmental Modeling & Assessment* **16**, 491 – 501.

[13] Alshitawi, M., Awbi, H. and Mahyuddin, N. (2009) Particulate matter mass concentration (PM_{10}) under different ventilation methods in classrooms. *International Journal of Ventilation* **8**, 93 – 108.

[14] Karppinen, A., Kukkonen, J., Elolähde, T., Konttinen, M., Koskentalo, T. and Rantakrans, E., (2000) A modelling system for predicting urban air pollution: model description and applications in the Helsinki metropolitan area, *Atmospheric Environment* **34**, 3723 – 3733.

[15] Gervasi, O. (2008) Computational Science and its Applications - ICCSA 2008, Springer, Italy.

[16] Yusof, N.F.F.M., Ramli, N.A., Yahaya, A.S., Sansuddin, N., Ghazali, N.A. and al Madhoun, W. (2010) Monsoonal differences and probability distribution of PM_{10} concentration, *Environmental Monitoring and Assessment* **163**, 655 – 667.

[17] Ul–Saufie, A.Z., Yahaya, A.S., Ramli, N.A., Rosaida, N. and Hamid, H.A. (2013) Future daily PM_{10} concentrations prediction by combining regression models and feed forward back propagation models with principle component analysis (PCA), *Atmospheric Environment* **77**, 621 – 630.