PREDICTION OF PARTICULATE MATTER IN AMBIENT ENVIRONMENT IN THE SURFACE COAL MINING AREAS: AN ANALYTICAL APPROACH INTEGRATING ARTIFICIAL NEURAL NETWORK AND FACTOR ANALYSIS

Deboleena Mukhopadhyay *

Dr. Suranjan Sinha**

Abstract

Air borne suspended particulate matter (SPM) and respirable particulate matter (PM_{10}) are emitted from different mining operations in a surface mine. Several modeling techniques are used to predict dust concentration within mining areas. Dust propagation, in mining areas, is a complex phenomenon and relations between different meteorological and mine variables is non linear. Moreover, precise knowledge about generation of dust is not available, which is a prerequisite for any mathematical and statistical model. Modeling of atmospheric data where imprecise knowledgebase and data is available can be done using Artificial Neural Network (ANN). In this paper ANN modeling technique is used to build a model to predict suspended particulate matter and respirable dust particle concentration. To reduce dimensionality of the input space, due to presence of several variables in the model, Factor Analysis (FA) is used. The results improved after integration of ANN with FA.

Key Words: Suspended Particulate Matter (SPM), Particulate Matter of size below 10µ (PM_{10}), Artificial Neural Network (ANN), Back Propagation Neural Network (BPNN), Factor Analysis (FA).

* Junior Research Fellow, BESU, Shibpur, DST sponsored project.

** Professor, Department of Mining Engineering, Bengal Engineering and Science University (BESU), Shibpur, Howrah 711103.

Introduction

Mining is an extractive industry that brings irreversible changes in the ambient environment. Different size fractions of dust are emitted from various point and non point sources within a surface mine. Mine excavation areas, waste dumps, haul roads are few examples of non point while emission from excavators is a point source of propagation of particulate matter in the mine environment. Dispersion and emission of dust depends on the prevailing meteorological conditions and site specific mine variables. Mine variables are dependent on the scale of mining operation and the size of different mine facilities like waste dumps and mine excavation areas. Propagation of both SPM and PM_{10}, within a surface mine, is complicated as relation between mining variables and meteorological variables is often non linear, dynamic and complex. Mathematical models are based mostly on Gaussian model of dispersal dust emitted from point sources (EPA 2006). The fundamental inputs to these models are precisely measured field data that may be expensive and time consuming to collect from a mine site.

ANN mimics the functioning of a human brain. ANN is a non-linear self adaptive approach (Girish .K., 2007). It is a learning algorithm that can identify correlated patterns between input data set and output data set. The network is trained using a set of input data. After training, the network model is used to predict with another data set (testing dataset). The predicted value is compared with the known output data (target). Error is then calculated and checked with the user defined tolerance limit. Weighs are readjusted by back propagation until the error reaches the prescribed tolerance value. . Several researchers have used Artificial Neural Network technique to model pollutant concentration. Hornik et al., 1989 stated that ANN can act as universal approximations of non-linear functions. ANN is a constructive tool either where no precise theoretical model is available, or when uncertainty in input parameters complicates deterministic modeling as, for example, in ecological or environmental systems (Huang and Foo, 2002; Lee et al., 2002; Scardi, 2001). Empirical air pollution forecasting systems can be developed using ANN approach (Gardner and Dorling, 1998; Jorquera et al., 1998). SO2 and PM10 concentrations can be predicted using ANN. (Boznar et al., 1993; Mok and Tam, 1998; Saral and Ertürk, 2003; Chelani et al., 2002; Onat et al., 2004; Sahin et al., 2005, Yildirim and Bayramoğlu, 2006). Gardner and Dorling (1998) have published a comprehensive review of studies using an ANN approach for environmental air pollution modeling. Kukkonen et al. (2003) have studied five neural network (NN) models, a linear statistical model and a deterministic modeling system for

the prediction of urban NO2 and PM10 concentrations. Sahin et al. (2004) used a multi-layer neural network model to predict daily CO concentrations, using meteorological variables, in the European side of Istanbul, Turkey. Kurt et al. (2008) also developed an online air pollution forecasting system in Istanbul using NN. Another NN model developed by Saral and Ertürk (2003) was also used to predict regional SO2 concentrations. Junninen et al. (2004) applied regression based imputation, nearest neighbor interpolation, a self organizing map, a multi-layer perceptron model and hybrid methods to simulate missing air quality data. Nagendra and Khare (2006) studied the usefulness of NN in understanding the relationship between traffic parameters and NO2 concentrations. Recently, several researchers used NN techniques to predict airborne PM concentrations: e.g. Ordieres et al. (2005) Hooyberghs et al. (2005), Perez and Reyes (2006) and Slini et al. (2006). All of these studies reported that ANN could be used to develop efficient air-quality analysis models. In case of complex phenomenon like dust dispersal in atmosphere there is problem of large dimensionality in the input space. Due to the complexity of the environmental system the number of weight coefficients of the ANN model rise into the millions (Lary, D.J et al.2009). Principal Factor Analysis, can be applied to reduce the number of variables. In this analysis, factors are structured according to the proportion of the variance in the input dataset that can be explained by these factors. These factors are rotated for elucidation. Further discussion on both ANN and FA is done in the proceeding subsections.

The paper is structured as follows. In the beginning a brief description of the study area where from data is collected is highlighted. Next, theoretical foundation of ANN and FA is built to develop a theoretical foundation for analytical framework for modeling of SPM and PM_{10} concentration in the mining areas. Neural network is trained with data from three mines under a mining company and network is developed for prediction of dust concentration. Due to high dimensionality in input space FA is used to improve the performance of the model.

Brief description of the study area

The study area is covered with agricultural land and waste land. Three mega surface mines are located within the study area. Also there are other mines within 10 – 15 km from these mines. Mining operations are mechanized with production over five million tonnes from each mining units. The major sources of dust generation are due to truck movement on mine roads and coal handling plants. Water sprinklers are regularly used to suppress dust in the mine roads.

Artificial Neural Network (ANN)

ANN is represented by a set of nodes and arrows (Fig1) which is a fundamental concept in graph theory. Training of the network is the process of learning when the error is calculated, as the difference between the predicted output and actual output (target). If average error reaches user defined error tolerance limit, the training is stopped; otherwise weights are readjusted by back propagation. In feed forward neural network, each node of a layer is connected to the output of all nodes of the previous layer. All inputs to a node are weighted independently, summed with bias and fed into logistic or other non-linear functions. The output is then connected to all neurons of the next layer. In general, the preparation of a neural network requires a forward model for computing a training set and test data set and a neural network training procedure. A successful pattern classification methodology depends heavily on the particular choice of the features used by the classifier (Sahin et al, 2011). The Back-Propagation Neural Network (BPNN) is the best known and widely used learning algorithm in training multilayer feed forward neural networks. BPNN is a multi-layer feed forward, supervised learning network based on gradient descent learning rule. It provides a computationally efficient method for changing the weights in feed forward network, with different activation function units. Being a gradient descent method it minimizes the total squared error of the output computed by the net.

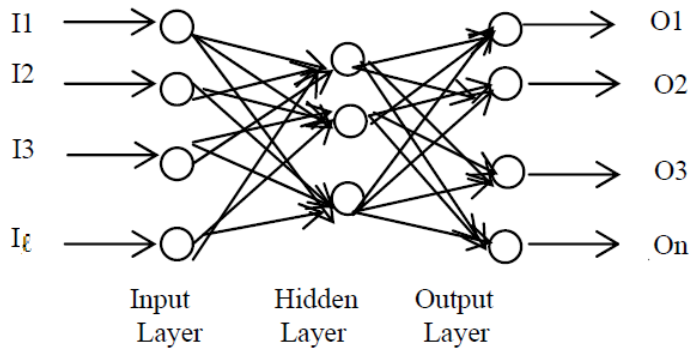


Fig. 1 **Feed forward back-propagation neural network**

The steps in the BPN algorithm are:

Step 1:

Normalize the input set $\{ I \}_I$ and output set $\{ O \}_o$ in the range [0.1 0.9] with respect to their minimum and maximum values. In normalized form assume that there are ℓ inputs given by $\{ I_T \}_I$ and n outputs namely $\{ O_T \}_I$ .Also assume that the number of neurons in the hidden layer lie in between $\ell \leq m \leq 2\ell$.

Step 2:

Add biases to each neuron in the hidden layer and output layer. Set the input of biases as +1.

Step 3:

Let [V] be the weights of synapses connecting input neuron and hidden neuron.

Let [b_V] be the weights of synapses connecting bias and hidden neuron.

Let [W] be the weights of synapses connecting hidden neuron and output neuron.

Let [b_W] represents the weights of synapses connecting bias and output neuron.

Initialize the weights to small random values usually lying between -1 to +1;

Step 4:

For training data, we need to present one set of inputs and outputs. Present the pattern as inputs to the input layer $\{ I \}_I$. Then by using linear activation function, the output of the input layer may be evaluated as $\{ O \}_I = \{ I \}_I$

Step 5:

Compute $\{ I \}_H$,the inputs to the hidden layer by multiplying corresponding weights of synapses and adding bias as

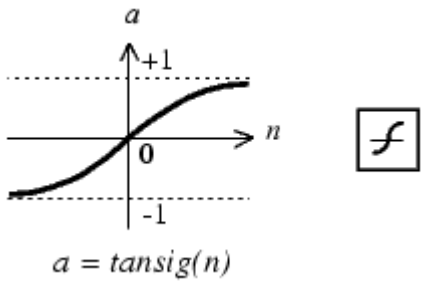
$$\underset{m \times 1}{\{ I \}_H} = \underset{m \times \ell}{[V]^T} \cdot \underset{\ell \times 1}{\{O\}_I} + \underset{m \times 1}{[b_V]^T}$$

Step 6:

Let the hidden layer units, evaluate the output $\{ O \}_H$ using the *tansig function* (Hyperbolic tangent sigmoid transfer function) as

$$\{O\}_H = \underset{m \times 1}{\left\{ \begin{array}{c} \cdots\cdots \\ \cdots\cdots \\ \dfrac{2}{(1+e^{(-2*I_{Hi})}) - 1} \\ \cdots\cdots \\ \cdots\cdots \end{array} \right\}}$$

The graph and symbol of *tansig* transfer function:



$a = tansig(n)$

Step 7:

Compute $\{ I \}_O$, the inputs to the output layer by multiplying weights of synapses and adding bias as

$$\underset{n \times 1}{\{ I \}_O} = \underset{n \times m}{[W]^T} . \underset{m \times 1}{\{O\}_I} + \underset{n \times 1}{[b_W]^T}$$

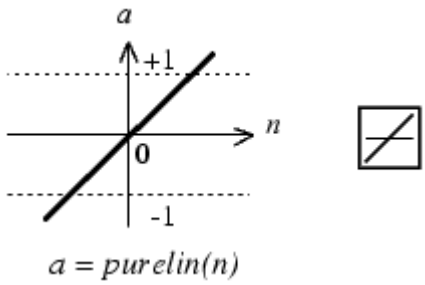
Step 8:

Let the output layer units, evaluate $\{ O \}_O$ (output) using *purelin function* as

$$\{O\}_O = \begin{Bmatrix} \\ \\ \dfrac{1}{I_{Oj}} \\ \\ \\ \end{Bmatrix}_{n \times 1}$$

The above is the network output.

The graph and symbol of *purelin* function:



$a = purelin(n)$

Step 9:

To update weight and bias values according to Levenberg-Marquardt optimization algorithm, *trainlm* function is used.

Step 10:

Calculate the instantaneous error and the difference between the network output $\{ O \}_O$ and the desired output $\{ T \}_O$ as for the i^{th} training set as

$$E^P = \frac{1}{2} \sum_{j \in C} (T_j - O_{Oj})$$

Where the set C includes all the neurons in the output layer of the network. That is the number of elements in C is n. Store E^P, p indicates the iteration.

Step 11:

Steps 5 – 10 are repeated as an epoch that for each data pattern.

Step 12:

Calculate Mean absolute error

$$E = \frac{1}{N} \sum_{n=1}^{N} E^P$$

Where N is the total number of data patterns.

As stated before several researchers have pointed out that BPNN is an efficient tool to model propagation, dispersal and prediction of atmospheric pollutants. However, atmosphere being a complex non linear system there is high dimensionality in the input space. As the dimensionality of input variables increases, the amount of training data required by the model increases rapidly, hence, size of the network, rapidly increases as well. Reduction of dimensionality of the input data can be achieved through the use of Factor Analysis (FA).

Factor Analysis (FA)

The general objectives of a FA (Tiwary et al, March 2011) are data reduction and data interpretation (Johnson and Wichern, 1992). Factor Analysis is normally conducted in a sequence of steps (Pao-Wen Grace Liu, 2009).

Step 1: Initial Extraction of the Components:

Correlation matrixes of the involved variables are the input to compute eigenvalues and eigenvectors. The principal component with a large eigenvalue implies the capability to explain relative high variance. The first component can be expected to account for a fairly large amount of the total variance. Each succeeding component will account for progressively smaller amounts of variance. Although a large number of components may be extracted in this way, only the first few components will be important enough to be retained for interpretation.

Step 2: Determining the Number of Meaningful Components to Retain:

The first few components will account for meaningful amounts of variance, and that the latter components will tend to account for only trivial variance. The next step of the analysis, therefore, is to determine how many meaningful components should be retained for interpretation. Two criteria that are used in making this decision: the eigenvalue-one criterion and the scree test. By using Kaiser criterion (Kaiser, 1960) any component with an eigenvalue greater than 1.00 will be retained. Any component that displays an eigenvalue greater than 1.00 is accounting for a greater amount of variance than had been contributed by one variable. Such a component is therefore accounting for a meaningful amount of variance, and is worthy of being retained. On the other hand, a component with an eigenvalue less than 1.00 is accounting for less variance than had been contributed by one variable. With the scree test, we plot the eigenvalues associated with each component and look for a break between the components with relatively large eigenvalues and those with small eigenvalues **(**Cattell, 1966). The components that appear before the break are assumed to be meaningful can be retained for rotation. Those appearing after the break are unimportant and hence not retained.

Step 3: Rotation to a Final Solution using Factor Analysis:

Factor patterns and factor loadings:

After extracting the initial components from the correlation matrix, an unrotated factor pattern matrix is created. The rows of this matrix represent the variables being analyzed, and the columns represent the retained components; these components are referred to as Factor 1, Factor 2 etc. A factor loading is a general term for a coefficient that appears in a factor pattern matrix.

Rotations:

When more than one factor has been retained in an analysis, the interpretation of an unrotated factor pattern is usually quite difficult. To make interpretation easier, an operation

called a rotation. A rotation is a linear transformation that is performed on the factor solution for the purpose of making the solution easier to interpret. In this study 'varimax rotation' is done. The varimax rotation is developed by Kaiser (1958), probably the most commonly used orthogonal rotation, compared to some other types of rotations (Herv´e Abdi 2003). Varimax rotation tends to maximize the variance of a column of the factor pattern matrix. This simplifies the interpretation because, after a varimax rotation each factor represents only a particular set variables. In addition, the factors can often be interpreted from the opposition of few variables with positive loadings to few variables with negative loadings. The components are the linear combination of the involved variables explained by each factor. In this study, the derived components are further used as the input of the Artificial Neural Network Model. Since the number of input variables used in the prediction process can be reduced, so that the computational cost can be reduced as well.

Results & Discussions:

Descriptive data statistics for the meteorological parameters , mean wind speed, relative humidity and average temperature, and mining parameters , distance from mines facilities, dump, haul-road length , coal handling plant (CHP), siding; coal production, overburden removed(OBR) are shown in Table1 . Data summary of airborne pollutants SPM and PM10 are included in Table 2**.**

Variables	Observation	Mean	Std. Dev.	Min	Max
OBR	98	13617.77	5401.64	932	32644
Coal	98	25083.73	10483.25	8102	47624
Dump	98	677.6122	246.9432	350	1760
Haul Road Area	98	168.3673	58.09154	100	240
Excavation Area	98	617.6939	218.4314	310	960
Siding	98	593.2041	207.6752	50	816
CHP	98	156.5306	35.27863	110	240
Mean Wind Speed	98	0.9630612	0.797768	0	3.13
Relative Humidity	98	75.80429	18.31664	44.53	100
Avg. Temp	98	30.00388	4.519643	19.72	38.25
Distance	98	557.6531	132.0637	300	900

Table 1: Data description for independent variables

Variables	Observation	Mean	Std. Dev.	Min	Max
SPM	98	179.3265	34.38466	82	277
PM10	98	88.03061	15.73194	51	153

Table 2: Data description for output variables

Next, eleven independent variables are used as input to the BPNN model, discussed earlier, for the prediction of SPM and PM10.

Comparative Analysis	Without Factor Analysis	
	11 Variables	
	SPM	PM10
No of observations	70	78
Correlation Coefficient	0.67	0.66
Mean Absolute Error	0.2472	0.2125

Table 3: Predictions of SPM & PM10.

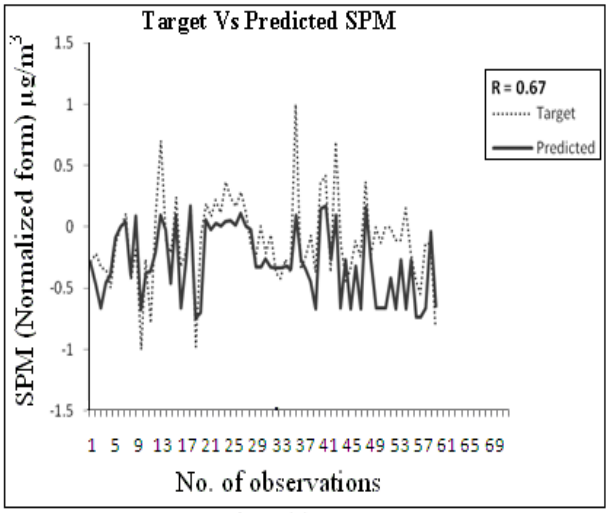


Fig 2 (a)

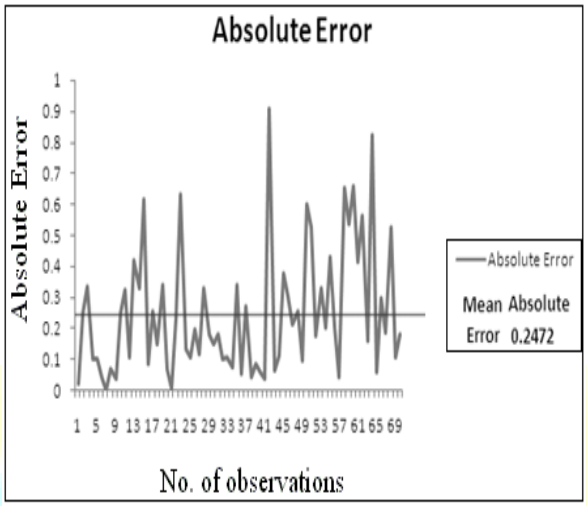


Fig 2 (b)

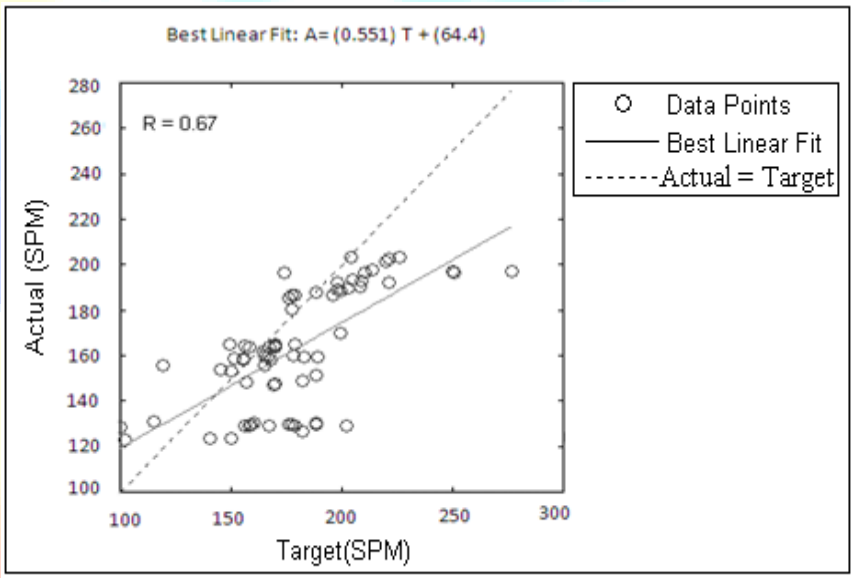


Fig 2 (c)

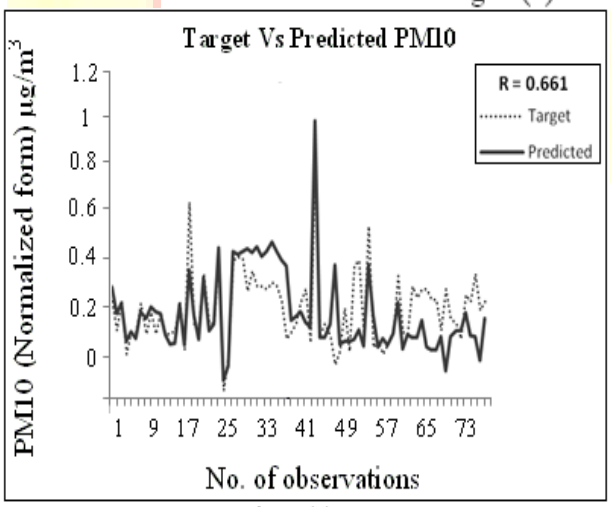


Fig 3 (a)

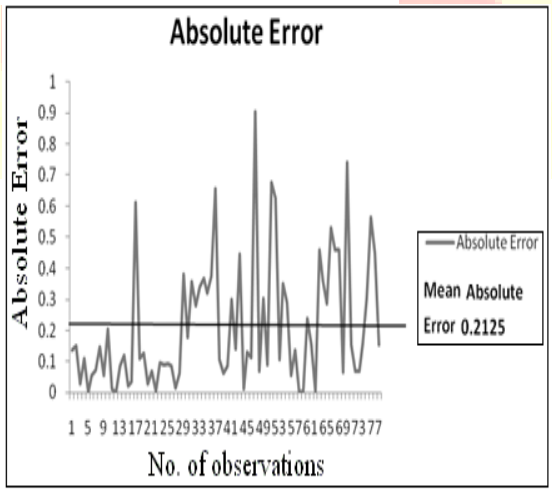


Fig 3 (b)

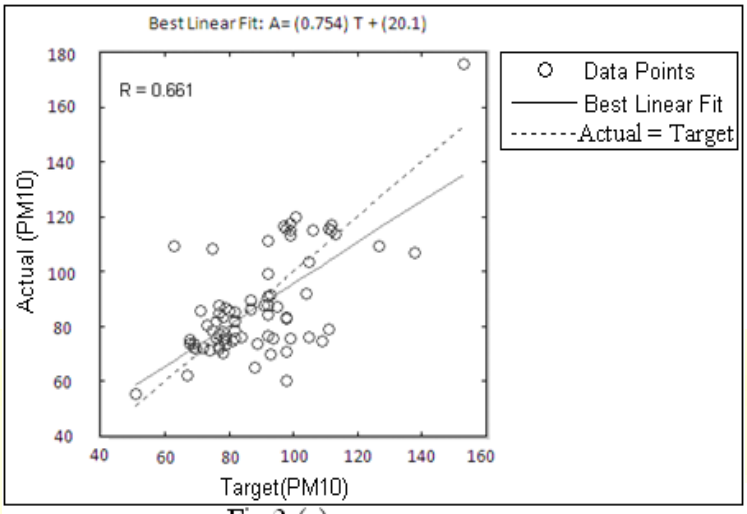


Fig 3 (c)

Fig 2(a) and Fig 3(a) illustrate the graph plot between actual (target) versus predicted SPM and PM_{10} respectively. Fig 2(b) & 3(b) show plot of the absolute error and mean absolute error lines for both the outputs . In Figs 2(c),3(c) the regression line fitted between target and predicted values of SPM and PM_{10} are shown.

To improve the prediction capability of the model the dimensions of input variables (11) are reduced using Factor Analysis**.** Factor Analysis is applied in the data set . Correlation matrix is calculated correlations between the independent variables are measured, which are in the range of -0.7482 to 0.8036 (shown in Table 4). Another important test for Factor Analysis is the Kaiser Meyer Olkin (KMO) (Perry R. Hinton, 2004). The KMO test examines the data for sampling adequacy; this gives a measure of the common variance amongst the variables that the factors will be able to account for. The KMO statistic ranges from 0 to 1.The observed result of KMO in our study is 0.6523 (revealed in Table 5).

Variables	OBR	Dump	Haul Road Area	Excavation Area	Mean Wind Speed	Relative Humidity	Coal	Siding	CHP	Avg. Temp	Distance
OBR	1.0000	0.4439	0.3359	0.2985	-0.3315	0.0606	**0.8039**	-0.2020	0.2558	-0.0977	0.0280
Dump	0.4439	1.0000	**0.6386**	**0.7425**	-0.1484	-0.0195	0.3252	-0.1685	**0.7280**	0.0488	-0.0125
Haul Road Area	0.3359	**0.6386**	1.0000	**0.7425**	-0.3060	0.1816	0.3540	0.1147	0.5336	-0.1274	0.2402
Excavation Area	0.2985	**0.7425**	**0.7425**	1.0000	-0.2633	0.1099	0.3068	0.0886	**0.6442**	0.0140	0.3980
Mean Wind Speed	-0.3315	-.1484	-0.3060	-0.2633	1.0000	**-0.7482**	-0.3650	-0.1832	-0.1996	0.5516	-0.2269
Relative Humidity	0.0606	-.0195	0.1816	0.1099	**-0.7482**	1.0000	0.1339	0.3153	0.0997	**-0.6249**	0.2111
Coal	**0.8039**	0.3252	0.3540	0.3068	-0.3650	0.1339	1.0000	0.0337	0.2322	-0.0640	0.0591
Siding	-0.2020	-.1685	0.1147	0.0886	-0.1832	0.3153	0.0337	1.0000	0.1992	-0.3026	0.2497
CHP	0.2558	**0.7280**	0.5336	**0.6442**	-0.1996	0.0997	0.2322	0.1992	1.0000	-0.0922	-0.0808
Avg. Temp	-0.0977	0.0488	-0.1274	0.0140	0.5516	**-0.6249**	-0.0640	-0.3026	-0.0922	1.0000	0.0118
Distance	0.0280	-.0125	0.2402	0.3980	-0.2269	0.2111	0.0591	0.2497	-0.0808	0.0118	1.0000

Table 4 : correlation matrix (11 X 11) of the involved variables

Variable	*KMO*
OBR	0.5623
Dump	0.7414
Haul Road Area	0.8816
Excavation Area	0.7027
Mean Wind Speed	0.7420
Relative Humidity	0.6798
Coal	0.5611
Siding	0.3651
CHP	0.6784
Avg. Temp	0.6668
Distance	0.3333
Overall	0.6523

Table 5 :KMO statistics

Any value over 0.6 (KMO) is regarded as acceptable for a factor analysis. Lesser values would mean that the factor analysis will not be able to account for much of the variability in the data and so is not worth undertaking.

Next, eigen values and corresponding eigen vector matrix are calculated (Table 6 & Table 7)from the 11 X 11 square matrix (Table 4). Four eigen values, exceeded 1and are plotted in the Scree plot (Fig 4). The amount of proportion explained by theses four eigen values are realively high. The comparative analysis of the eigen values are shown in Table 8.

0.105907	0	0	0	0	0	0	0	0	0	0
0	0.148993	0	0	0	0	0	0	0	0	0
0	0	0.167071	0	0	0	0	0	0	0	0
0	0	0	0.205439	0	0	0	0	0	0	0
0	0	0	0	0.355233	0	0	0	0	0	0
0	0	0	0	0	0.431412	0	0	0	0	0
0	0	0	0	0	0	0.814217	0	0	0	0
0	0	0	0	0	0	0	1.137964	0	0	0
0	0	0	0	0	0	0	0	1.495417	0	0
0	0	0	0	0	0	0	0	0	2.303231	0
0	0	0	0	0	0	0	0	0	0	3.835116

Table 6: Eigen Value Matrix

0.504122	0.160803	-0.41119	0.068815	-0.27941	-0.09296	-0.13785	-0.17248	-0.54345	-0.12761	0.314805
-0.36874	0.65305	0.095414	0.239772	-0.17723	0.067264	0.169956	0.216736	0.076051	-0.32496	0.38443
-0.05596	-0.07172	-0.33558	-0.04472	0.613976	-0.52623	0.082689	-0.0272	0.19732	-0.1023	0.40773
0.530926	-0.14592	0.59726	-0.06203	-0.06841	0.006714	0.133005	-0.12219	0.31005	-0.17688	0.414116
0.050378	-0.20803	0.01668	0.673259	-0.15444	-0.35593	-0.24338	0.006474	0.1759	-0.40738	-0.30647
0.130709	-0.08805	-0.05727	0.675087	0.212032	0.306801	0.230987	0.085511	-0.02525	0.521974	0.20807
-0.42239	-0.25886	0.395717	0.099964	0.17501	0.047931	-0.40508	-0.24703	-0.47743	-0.05201	0.315149
0.127704	0.36707	-0.00759	-0.00825	0.044898	0.016156	-0.76383	-0.02082	0.375922	0.337951	0.084566
-0.12797	-0.50199	-0.34167	-0.09464	-0.27089	0.335764	-0.19988	0.413121	0.228194	-0.16661	0.359672
0.122702	0.100631	-0.17639	0.044483	0.438101	0.61129	-0.0635	-0.3297	0.097554	-0.47997	-0.15326
-0.28125	-0.07962	-0.21075	0.043713	-0.37696	-0.00933	0.153246	-0.74794	0.318424	0.151355	0.137242

Table 7: Eigen Vector Matrix

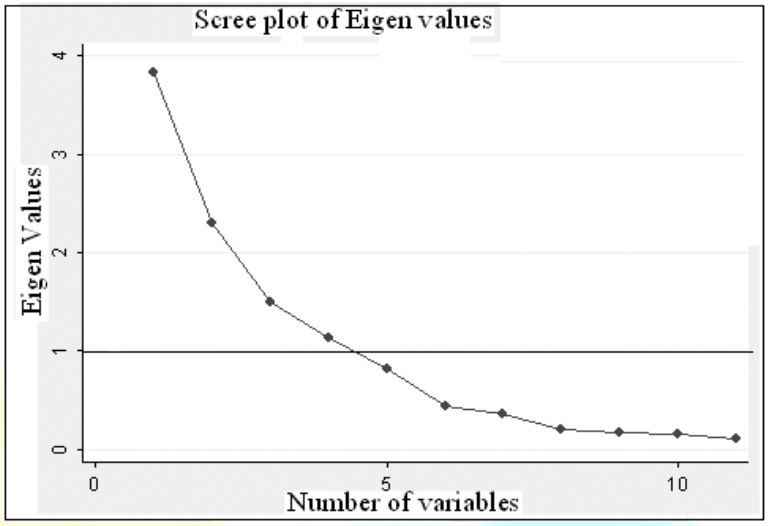


Fig 4: The Scree plot

Factors	Eigen Value	Difference	Proportion	Cumulative
Factor1	3.83513	1.53193	0.3486	0.3486
Factor2	2.30320	0.80777	0.2094	0.5580
Factor3	1.49543	0.35756	0.1359	0.6940
Factor4	1.13788	0.32364	0.1034	0.7974
Factor5	0.81424	0.38279	0.0740	0.8714
Factor6	0.43145	0.07619	0.0392	0.9107
Factor7	0.35526	0.14982	0.0323	0.9430
Factor8	0.20543	0.03836	0.0187	0.9616
Factor9	0.16707	0.01805	0.0152	0.9768
Factor10	0.14902	0.04314	0.0135	0.9904
Factor11	0.10589	_	0.0096	1.0000

Table 8: Comparative analysis of Eigen values

Using *STATA 10* Software the Factor Pattern Matrix corresponds to the above Eigen values (3.835116, 2.303231, 1.495417 and 1.137964) is evaluated (Table 9). After performing varimax rotation on the Factor Pattern Matrix four orthogonal factors are produced. The rotated pattern matrix here offers a clear picture of the relevance of each variable in the factor. Here Factor1 is highly correlated with dump, haul road area, excavation area and CHP. In similar manner Factor 2 has high correlation with relative humidity, mean wind speed and avg. temperature; Factor 3 by coal and OBR and siding; Factor 4 is significantly correlated with distance (Table 10).

Variables	Factor 1	Factor 2	Factor 3	Factor 4	Uniqueness
OBR	0.6165	0.1937	-0.6646	0.1840	0.1069
COAL	0.6171	0.0790	-0.5838	0.2636	0.2026
DUMP	0.7529	0.4932	0.0930	-0.2312	0.1279
HAUL ROAD AREA	0.7985	0.1552	0.2413	0.0290	0.2793
EXCAVATION AREA	0.8110	0.2684	0.3791	0.1303	0.1095
SIDING	0.1656	-0.5129	0.4597	0.0222	0.4977
CHP	0.7044	0.2529	0.2790	-0.4407	0.1679
MEAN WIND SPEED	-0.6002	0.6182	0.2151	-0.0069	0.2112
RELATIVE HUMIDITY	0.4075	-0.7922	-0.0309	-0.0912	0.1971
AVG TEMPARATURE	-0.3002	0.7284	0.1193	0.3517	0.2414
DISTANCE	0.2688	-0.2298	0.3895	0.7978	0.0868

Table 9: Unrotated Factor Pattern Matrix

After rotate the Factor Loading Matrix using Varimax method Table 10 is obtained.

Variable	Factor1	Factor2	Factor3	Factor4
OBR	-0.07319	-0.02163	**0.49989**	-0.02902
COAL	-0.09041	0.00029	**0.46999**	0.06382
DUMP	**0.31914**	-0.06501	0.01318	-0.15258
HAUL ROAD AREA	**0.24595**	0.00253	-0.01397	0.11785
EXCAVATION AREA	**0.27976**	-0.06937	-0.04978	0.22392
SIDING	0.06778	0.18565	**-0.26111**	0.19773
CHP	**0.37338**	0.06080	-0.17035	-0.24173
MEAN WIND SPEED	0.04458	**-0.30617**	-0.14354	-0.03076
RELATIVE HUMIDITY	-0.03699	**0.36677**	-0.02006	0.01229
AVG TEMPARATURE	0.00587	**-0.39520**	0.05371	0.22128
DISTANCE	-0.08044	-0.08567	0.02901	**0.74808**

Table 10: Factor Pattern Matrix after varimax rotation

Using the above 4 factors four new components can be deduced by the linear combination of involved variables.

Component1 = (0.31914 * DUMP + 0.24595 * HAUL ROAD AREA + 0.27976 * EXCAVATION AREA + 0.37338 * CHP)

Component2 = (0.36677 * RELATIVE HUMIDITY – 0.30617 * MEAN WIND SPEED – 0.3920 * AVG. TEMPERATURE)

Component3 = (0.46999 * COAL + 0.49989 * OBR – 0.26111 * SIDING)

Component 4 = DISTANCE.

The derived components contain the impact of all the independent variables. There exists lesser correlation among these derived components (Table 11). Thus we find out 4 uncorrelated components by factor analysis.

	Component 1	Component 2	Component 3	Component 4
Component 1	1.0000	0.0540	0.3849	0.1696
Component 2	0.0540	1.0000	0.1194	0.1784
Component 3	0.3849	0.1194	1.0000	0.0485
Component 4	0.1696	0.1784	0.0485	1.0000

Table 11: Correlation Matrix of four components

Now, these 4 components are applied as the inputs of BPNN model and get better prediction of the SPM and PM10. The significant advantage is that as the number of input variables used in the prediction process can be reduced, so that the structure of the predictor and the computational cost can be reduced as well. In the analysis 3 components are used and 4^{th} component is eliminated. Finally, the best performance by using all the 4 components, instead of 3components, as the input of BPNN prediction model (Table 12). Comparative analysis of prediction of SPM and PM_{10}, using BPNN model, before and after reducing the input variables by Factor Analysis is indicated in Table 13.

Comparative Analysis	After Factor Analysis			
	3 Components		4 Components	
	SPM	PM10	SPM	PM10
No of observations	70	78	70	78
Correlation Coefficient	0.68	0.646	0.849	0.848
Mean Absolute Error	0.2308	0.2023	0.0970	0.0860

Table 12: Comparative result analysis for SPM and PM10 taking principal factor as input of BPNN

Supporting figures for SPM:

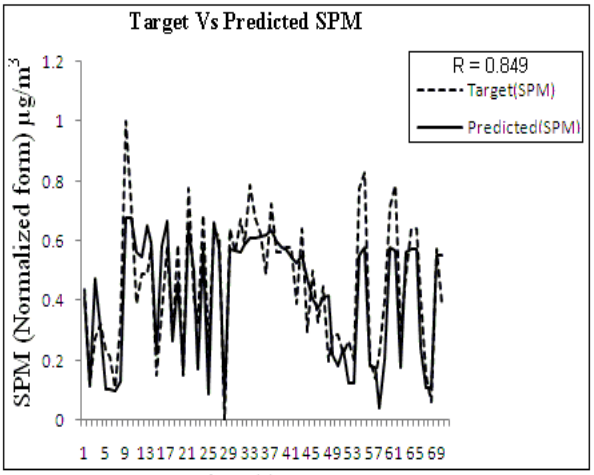


Fig 5 (a)

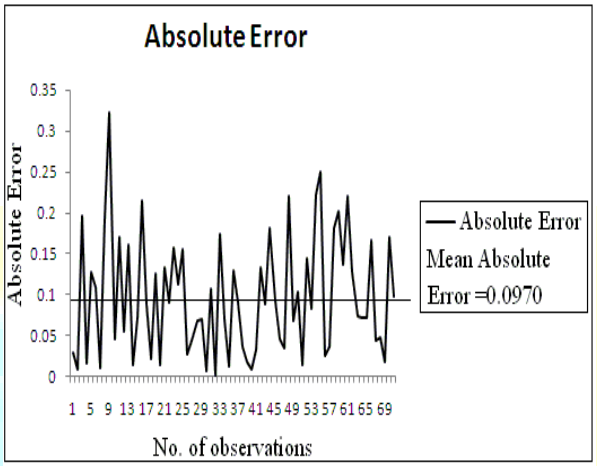


Fig 5 (b)

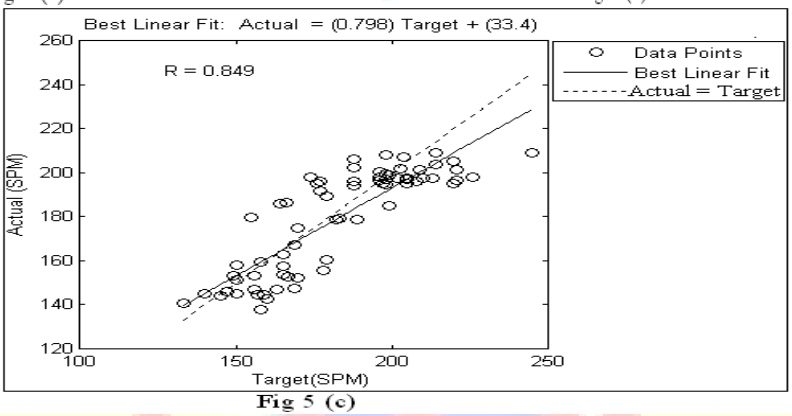


Fig 5 (c)

Fig 5: Prediction performance of SPM using BPNN taking 4 components (after Factor analysis)

Supporting figures for PM10:

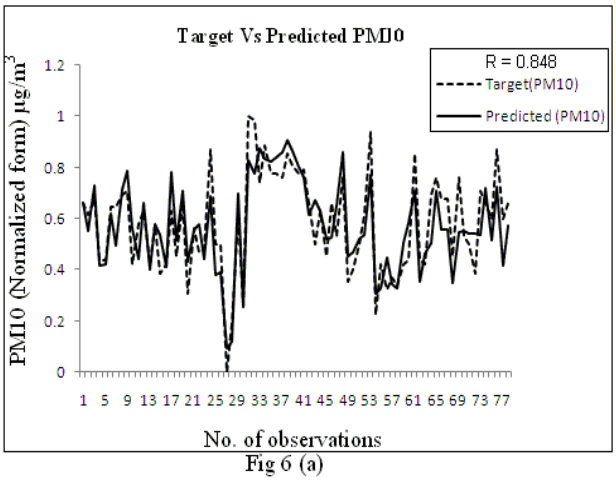


Fig 6 (a)

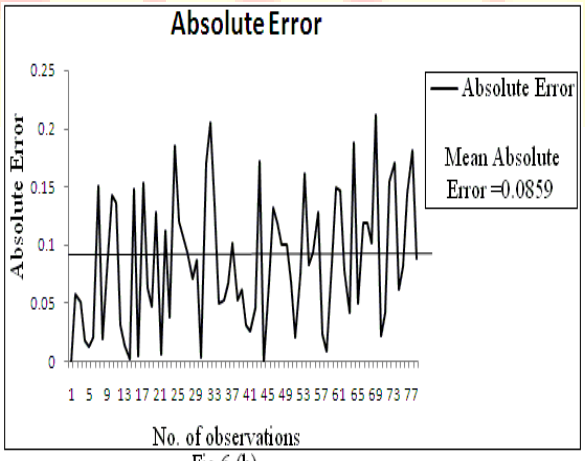


Fig 6 (b)

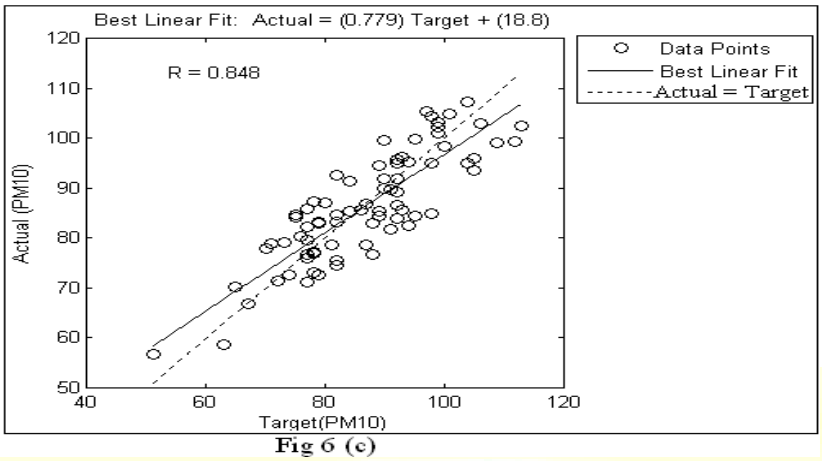


Fig 6 (c)

Fig 6: Prediction performance of PM10 using BPNN taking 4 components (after Factor analysis)

Comparative Analysis	Without Factor Analysis		After Factor Analysis			
	11 Variables		3 Components		4 Components	
	SPM	PM10	SPM	PM10	SPM	PM10
No of observations	70	78	70	78	70	78
Correlation Coefficient	0.67	0.66	0.68	0.646	0.849	0.848
Mean Absolute Error	0.2472	0.2125	0.2308	0.2023	0.0970	0.0860

Table 13: Output table with and without Factor analysis

Conclusion

Modeling of dust concentration prediction in the ambient environment in a surface mining area is a complex dynamic phenomenon as explicit knowledge base is not available. In developed countries regulatory models are used, as per requirement of statute, to predict and mitigate dust concentration in surface mines. Developing countries are yet to build similar models to capture myriad variability involved in propagation and dispersal of atmospheric pollutants. In this paper by integrating ANN and FA an analytical framework is developed to predict dust concentration at surface mines. Further validation is essential using data from several mines across the country. This paper is a small a step towards filling up the huge gap in existing research on modeling of environmental pollutants emitted by different industrial activities.

Acknowledgement

Data is collected under the a project funded by Department of Science and Technology, Government of India

References

- Cattell, R. B. (1966).” The scree test for the number of factors. Multivariate Behavioral Research”, Volume 1, 245-276.
- EPA (2006): http://www.epa.gov/scram001/dispersion_prefrec.htm
- Gardner, M. W., Dorling, S.R., (1998). “Artificial neural networks (the multilayer perception) – a review of application in the greater Seoul area”, Atmospheric Environmental 36, 201-212.
- Herv´e Abdi (2003),”Factor Rotations in Factor Analyses”, The University of Texas at Dallas, Encyclopedia of Social Sciences.
- Hornik, K. Stinchcombe, M., White, H., (1989). ”Multilayer feedforward networks are universal approximators”.Neural Network 2,359 – 366.
- Huang, W.R., Foo, S., (2002). ‘Neural network modeling of salinity variation in Apalachicola River”. Water Research 36, 356–362.
- Junninen, H., Niska, H., Tuppurainen, K., et al., (2004). “Methods for imputation of missing values in air quality data sets”, Atmospheric Environment 38, 2895–2907.
- J.Girish (2007): “Artificial Neural Network and its application”, I.A.R.I, V 41- 49.
- Kukkonen, J., Partanen, L., Karppinen, A., et al., (2003), “Extensive evaluation of neural network models for the prediction of NO2 and PM10 concentrations, compared with a deterministic modeling system and measurements in central Helsinki”, Atmospheric Environment 37, 4539–4550.
- Kurt, A., Gülbağcı, B., Karaca, F., Alagha, O., (2008). “An online air pollution forecasting system using neural networks”, Environmental International 34, 592–598.
- Lary, D.J., Remer, L.A., Macneill, D., Roscoe, B., Paradise, S., (2009). “Machine learning and bias correction of MODIS aerosol optical depth”. IEEE Geoscience and Remote Sensing Letters 6 (4), 694–698.
- Nagendra, S.M.S., Khare, M., (2006). “Artificial neural network approaches for modeling nitrogen dioxide dispersion from vehicular exhaust emissions”, Ecological Modeling 190, 99–115.
- Ordieres, J.B., Vergara, E.P., Capuz, R.S., Salazar, R.E., (2005). Neural network prediction model for fine particulate matter (PM2.5) on the US–Mexico border in El Paso

(Texas) and Ciudad Juarez (Chihuahua). Environmental Modeling & Software 20, 547–559.

- Pao-Wen Grace Liu. (2009) "Simulation of the daily average PM10 concentrations at Ta-Liao with Box–Jenkins time series models and multivariate analysis", Atmospheric Environment 43 (2009) 2104–2113. [Article 3.2 PCA, p-2106]
- Perry R. Hinton (2004), Statistics Explained, 2nd Edition, [Factor Analysis, p-304-308].
- Sahin, U., Ucan, O.N., Soyhan, B., Bayat, C., (2004). "Modeling of CO distribution in Istanbul using artificial neural networks", Fresenius Environmental Bulletin 13 (9), 839–845.
- Sahin, U.A., Bayat, C., Osman, N.U, (2011). " Application of cellular neural network (CNN) to the prediction of missing air pollutant data", Atmospheric Research 101-314-326.
- Saral, A., Ertürk, F., (2003). "Prediction of ground level SO2 concentrations using artificial neural network", Water, Air, and Soil Pollution 3, 297–306.
- Tiwari, S., Singh, A.K., Shukla, V.P., (March 2011) "Statistical Moments based Noise Classification using Feed Forward Back Propagation Neural Network" ,International Journal of Computer Applications (0975 – 8887), Volume 18– No.2, [Article 4, BPN].