Sources, Pattern, and Health Impacts of PM2.5 in the Central Region of Bangladesh using PMF, SOM, and Machine Learning Techniques

Md Shareful Hassan (shareful@gmx.com)
Jahangirnagar University https://orcid.org/0000-0003-4668-2951

Mohammad Amir Hossain Bhuiyan
Jahangirnagar University

S M Saify Iqbal
Dhaka University: University of Dhaka

Muhammad Tauhidur Rahman
King Fahd University of Petroleum & Minerals

Research Article

Keywords: PM2.5, ARI, PMF, SOM, Machine learning, Bangladesh

Posted Date: July 1st, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1810022/v1

Sources, Pattern, and Health Impacts of $PM_{2.5}$ in the Central Region of Bangladesh using PMF, SOM, and Machine Learning Techniques

Md. Shareful Hassan[1*], Mohammad Amir Hossain Bhuiyan[1], S M Saify Iqbal[2], Muhammad Tauhidur Rahman[3]

[1]Department of Environmental Sciences, Jahangirnagar University, Dhaka, 1342, Bangladesh
[2]International Centre for Climate Change and Development (ICCCAD), Bangladesh
[3]Department of City and Regional Planning, King Fahd University of Petroleum and Minerals
KFUPM Box 5053, Dhahran – 31261, Kingdom of Saudi Arabia

***Corresponding author:** Md. Shareful Hassan
Mailing address: Ph.D. Student, Environmental Science, Jahangirnagar University, Dhaka, Bangladesh
Email: shareful@gmx.com https://orcid.org/0000-0003-4668-2951

Abstract

The Particulate Matter 2.5 ($PM_{2.5}$) is one of the major environmental and public health threats in Bangladesh. It is important to explore the relationship between $PM_{2.5}$, and other variables to mitigate its adverse health impacts. This study aims to understand the sources, patterns, and health impacts of $PM_{2.5}$ in five central districts of Bangladesh using fourteen variables. These variables have been analyzed by PMF, SOM, Machine Learning, and Multi-regression analysis. This paper has found that $PM_{2.5}$ is correlated positively with NO (0.55), BC (0.45), CH_4 (0.38), and NOx (0.22), while correlated negatively with Rainfall (-0.10), CO (-0.33), and SO_2 (-0.24). In PMF modeling, the R^2 values of settlement density (1.00), SO_2 (0.99), DEM (0.94), Rainfall (0.77), NO (0.74) and Brickfield (0.66) have found as the most correlated variables. In this study, the dominant variables NO, CO, Rainfall, O3, AOT, CH4, and BC are found in Factor 1; SO2, settlement density, and DEM are found in Factor 2; and population density and brickfield are found in Factor 3. In SOM mapping, most of the variables are concentrated in the north-eastern, central, and south-eastern parts of the study area. The prediction of $PM_{2.5}$ using machine learning is significant, showing reasonable R^2 for Random Forest (0.85), Extreme gradient boosting (0.81), and Stepwise Linear (0.76). The impact of $PM_{2.5}$ on child ARI is significant (*$p=0.002$, $R^2=0.75$*); while child mortality is not significant (*$p=0.268$; $R^2=0.55$*). These results will be useful for creating and implementing local and regional $PM_{2.5}$ mitigation plans. Concern institutions and academia may also use these outputs for reducing health impacts, particularly child mortality and acute respiratory infections.

Keywords: $PM_{2.5;}$ ARI, PMF, SOM, Machine learning; Bangladesh.

1. Introduction

The Particulate Matter 2.5 ($PM_{2.5}$) is a heterogeneous combination of suspended particles of various chemical contents and sizes (Liang et al. 2013). The negative effects of $PM_{2.5}$ are determined by its concentrations in the atmosphere, which are influenced by a wide range of anthropogenic and natural sources (e.g., traffic emissions, industrial processes, residential combustions, biogenic emissions), related factors (e.g., climate, meteorological conditions, urbanization levels), and other events such as transportation and deposition of dust particles (Ni et al. 2018; Adães and Pires 2019). Even at concentrations below ambient air quality standards, long-term exposures to $PM_{2.5}$ particles has been linked to cardiovascular diseases, lung cancer, and both chronic and acute respiratory diseases which ultimately lead to untimely death among children and adult population (Andersen et al. 2012; Hoek et al. 2013; Raaschou-Nielsen et al. 2013; Beelen et al. 2014; Cesaroni et al. 2014). According to many researchers, over 400,000 premature children die each year in EU countries as a result of $PM_{2.5}$ (Badyda et al. 2017; Cho and Song 2017). The $PM_{2.5}$ particulates can penetrate deep into the human respiratory system due to its small sizes, especially when exposed for lengthy periods of time (Eeftens et al. 2012; Tallon et al. 2017).

In Bangladesh, the concentration of $PM_{2.5}$ particles in air is currently 15.4 times above the World health Organization (WHO) annual air quality guideline value (IQAIR 2022). Between 2002 and 2019, the average annual $PM_{2.5}$ concentration has increased by 42% in the urban areas of the country due to excessive emissions from various types of poorly maintained automobiles (Begum 2016; Hassan 2022). Furthermore, (Begum and Hopke 2018) highlighted that Dhaka and other major cities of Bangladesh had some of the highest $PM_{2.5}$ concentrations among the global cities for many years. $PM_{2.5}$ causes roughly 234,000 premature deaths each year, accounting for 3.5 percent of global data. Due to the rising trends of the $PM_{2.5}$, it has been identified as a major public health hazard for the people of Bangladesh, especially in the urban and semi-urban areas (Rahman et al. 2019). The high $PM_{2.5}$ standard threshold also has significant impacts on vulnerable demographic groups, especially for pregnant women, children, and elderly (over the age of 60) residents (Miller and Xu 2018).

Due to the excessive environmental threat and public health hazards, the sources, patterns, and possible health impacts should be investigated for mitigating and implementing policies to reduce $PM_{2.5}$ concentrations at Bangladesh's local, regional, and national levels. Some researchers have tried to identify the possible sources and patterns of $PM_{2.5}$ using Positive Matrix Factorization (PMF) model, its spatial concentration using Self-organizing Map (SOM), machine learning for prediction of $PM_{2.5}$, and regression analysis for possible health impacts (Chueinta et al. 2000; Naz et al. 2015; Kim et al. 2018; Joharestani et al. 2019; Ulavi and Shiva Nagendra 2019; Doreswamy et al. 2020; Liu et al. 2020).

The PMF model has recently gained popularity among scientists as an appropriate factorization receptor model for calculating the contributions and sources of pollutants in the environment (Tao et al. 2017; Chen et al. 2019b). When sources are not formally identified, the PMF model is highly recommended, although it necessitates post-treatment source identification. Using PMF analysis, (Nava et al. 2020) identified traffic congestions, biomass burnings, secondary sulfates, secondary nitrates, urban dust storms, Saharan dust particles, and marine aerosols as the seven main sources of $PM_{2.5}$ in the city of Florence, Italy. Secondary sulfate was found to be a significant $PM_{2.5}$ source on a regional scale. (Sharma et al. 2016) also used the PMF model to find secondary aerosols (21.3%) as the leading source of $PM_{2.5}$, followed by soil dusts (20.5%), vehicle emissions (19.7%), biomass burnings (14.3%), fossil fuel combustions (13.7%), industrial emissions (6.2%), and sea salts (4.3%) in the city of Delhi, India. (Kim et al. 2018) assessed the sources of several pollutants that contribute to ambient fine particles ($PM_{2.5}$) in Daebu Island, Korea using PMF model. Chemical speciation data was used in this work to estimate and identify possible $PM_{2.5}$ sources using the PMF model. (Srivastava et al. 2021) used PMF modelling in urban and rural areas of Beijing, China. One of the major limitations of these studies using PMF models is that they have used chemical-based analysis of $PM_{2.5}$ with minimum sample points in small geographical areas.

For mapping the hotspots of the distribution of $PM_{2.5}$ pollutions, SOM became quite common recently. A few researchers have used this to map the spatial distribution and concentration of each pollutant. However, the integration of SOM and PMF could be a new technique for allocating different pollution sources in Bangladesh (Hossain Bhuiyan et al. 2021). (Susanna et al. 2017) described the source characterization of PM_{10} and $PM_{2.5}$ mass concentrations using SOM by taking samples from Sardar Patel Road, Chennai, India during the winter months of January and February of 2008. Their findings revealed that $PM_{2.5}$ mass concentrations were high in their study area due to contributions from six different sources: earth crust/soils, fugitive dusts, marine aerosol/sea, secondary aerosols, traffic pollutions, and industries. (Srivastava et al. 2021) conducted a study on exploring the spatiotemporal interrelation of $PM_{2.5}$ concentration in Northern Taiwan by SOM using temporal datasets.(Lin et al. 2022) evaluated the link between $PM_{2.5}$ concentration, Weather Information System (WIS), precursors, and meteorological factors to investigate the secondary aerosols generation mechanism and trace the likely sources of $PM_{2.5}$ during severe pollution episodes using SOM. However, SOM has not been used intensively in $PM_{2.5}$ source mapping for large geographical areas.

Machine learning is one of the cutting-edge tools for predicting the latent relationship between dependent and different independent variables in air pollution studies. (Tian et al. 2016) estimated $PM_{2.5}$ from multi-source data where they employed different machine learning models in the Pearl River Delta (PRD) in China, using Random Forest (RF) and Gradient Boosting Regression Tree (GBRT). (Deters et al. 2017) employed a machine learning approach to estimate $PM_{2.5}$ concentrations from wind (speed and direction) and precipitation levels, based on six years of meteorological and pollutant data. A machine learning approach was developed by (Chen et al. 2018) to estimate the $PM_{2.5}$ concentrations across China using remote sensing, meteorological, and land use data. (Zhang et al. 2015) assessed the effects of various factors on $PM_{2.5}$ pollution by merging the Random Forest model, Shapley Additive exPlanations (RF-SHAP), Partial Dependence Plot (RF-PDP), and Positive Matrix Factorization (PMF). They have found that anthropogenic emissions and climatic conditions both contributed roughly 67% (40.5 $\mu g/m^3$) and 33% (19.7 $\mu g/m^3$) of the fluctuation in $PM_{2.5}$ concentrations. In all of above-mentioned literature on machine learning, most of the works have used a very few sets of data and information as well as very small sample data. Studies utilizing a large combination of air pollutants, climatic, environmental, and social data for predicting $PM_{2.5}$ is lacking and should be further investigated.

Measuring the impact of public health due to $PM_{2.5}$ is a complex issue, as many hidden determinants are involved. Exposure to ambient $PM_{2.5}$ is associated with child mortality and acute respiratory infection (ARI), which are found in Nairobi, Kenya (Egondi et al. 2018). To predict the statistical relationship of $PM_{2.5}$ with child mortality and ARI, many researchers have used multiple-regression modeling using an array of different datasets (Dominici et al. 2002; Naz et al. 2015; Sultana and Uddin 2019). However, integration and analyses of local hospital-based data with diverse air pollutants, and other variables were not examined.

Based on the literature review, this paper found that the combined use of different sets of data such as air pollutants, environmental, climatic, and social, to identify the sources, patterns, and health impacts of $PM_{2.5}$ is lacking, especially for urban areas of Bangladesh. Furthermore, using PMF for source identification, GIS analysis for factor mapping, SOM for concentration mapping/clustering, and machine learning for prediction of $PM_{2.5}$ concentrations using factorized data and health impact using multi-regression modeling will be unique for the country as well. Therefore, considering this knowledge gap, the goal of this study is to (1) determine the key sources of $PM_{2.5}$, (2) identify the core concentrated areas of the sources, (3) predict the $PM_{2.5}$ using factorized data, and (4) investigate the impacts of child mortality and ARI due to $PM_{2.5}$ and other air pollutants in five central districts of Bangladesh. The results and ultimate outcomes of the study could be used by the government, concerned ministries, UN bodies, domestic and international NGOs, civil societies, and environmental activists for local and regional level mitigation planning and implementations as well as achieve the goal of Sustainable Development Goal (SDG)-11 with indicators 11.6 and 11.6.2.

2. Methods and Materials

2.1 Study area

The study area is located in the Dhaka division of Bangladesh, covering its five major industrial districts (Dhaka, Gazipur, Narayanganj, Narsingdi, and Munshiganj). With an area of 6,043 $km^{2,}$ the study area is home to over 20 million residents (Fig. 1). The area has a tropical wet and dry climate, with annual average rainfall ranging from 694 mm to 2,376 mm (highest in Narsingdi district and lowest in Munshiganj district). According to (Deters et al. 2017), the air and water in the districts are becoming increasingly polluted because of increasing population, decreasing wetland and green spaces, rising multi-storied buildings, and growing commercial real estate developments. The areas also experience higher levels of traffic congestion, unplanned migration, and unplanned urban activities (Begum and Hopke 2019; Rahman et al. 2019; Iqbal et al. 2020). In addition, the study area also houses industries of ready-made garments, textiles, pharmaceuticals, cements, brickfields, fertilizers, and processing of raw material. All of these are the primary triggering factors for massive emissions of $PM_{2.5}$.

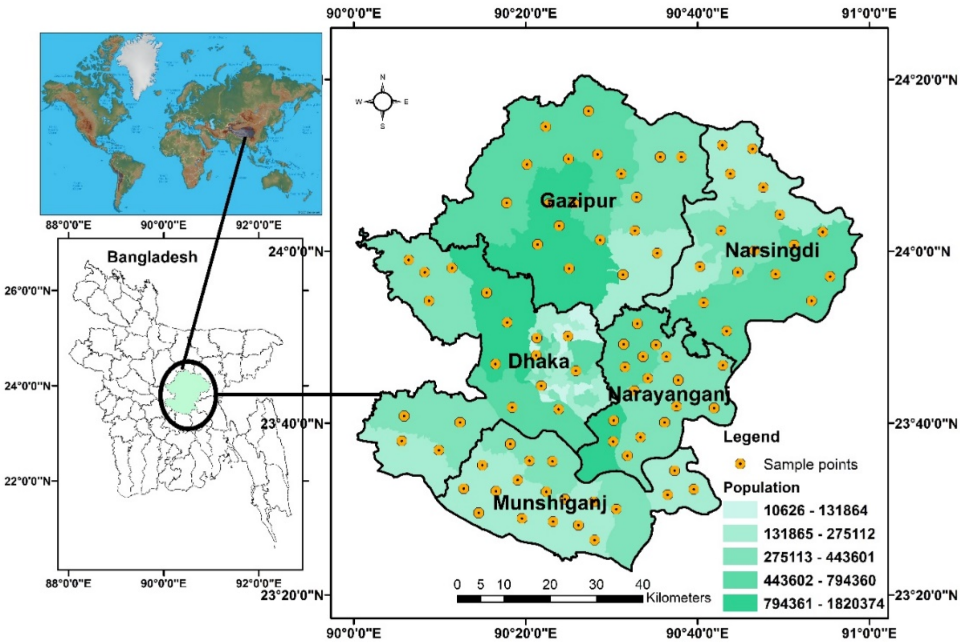

Fig. 1 Location of the study with the distribution of total sample points and the total population of each district.

2.2 Data sources and materials

For this study, numerous sources of data were gathered, processed, and analyzed. Some of these data were obtained from the Bangladesh Bureau of Statistics (BBS), Bangladesh Department of Health (BDH), and the Humanitarian Data Exchange Website. Data pertaining to air pollutants and their relevant parameters were obtained from earth observing satellite sensors and were downloaded from various online sources. High-resolution Google map was also used to digitize locations of brickfields. Table 1 summarizes the various parameters and their data sources used in the study.

Table 1 Characteristics of multisource databases used in the study

Parameters	**Unit**	**Data Sources**	**Temporal/data year**
$PM_{2.5}$	µg/m^3	https://sedac.ciesin.columbia.edu/ https://www.ecmwf.int/en/forecasts/accessing-forecasts	2010-2020
NO	nm^2	https://aura.gsfc.nasa.gov/	2010-2020
AOT	-	https://neo.sci.gsfc.nasa.gov/	2010-2020
SO_2	kg m^3	https://search.earthdata.nasa.gov/search	2019-2020
CO	Ppm	https://search.earthdata.nasa.gov/search	2010-2020
BC	kg m^3	https://giovanni.gsfc.nasa.gov/giovanni/	2010-2020
O_3	Dobson	https://search.earthdata.nasa.gov/search	2010-2020
DEM	Meter	https://search.earthdata.nasa.gov/search	2019
CH_4	ppbv	https://giovanni.gsfc.nasa.gov/giovanni/	2010-2020
Rainfall	Mm	https://gpm.nasa.gov/trmm	2010-2020
Wind speed	Mph	https://www.worldclim.org/data/worldclim21.html	2017
Settlement density	Per dot	Bangladesh Bureau of Statistics	2011
Brickfield	Meter	Direct digitization using Google map	2019
Poverty	%	https://data.humdata.org/	2017
Health Data	person	Bangladesh Department of Health (Upazila Health Complex)	2021-2022

Since the data from BBS and BDH were collected in table format, they were converted into GIS shapefiles for further analysis. All the data preprocessing, post-processing, analysis, mapping, and model performance assessments were conducted using Excel (Microsoft 356), ArcGIS v. 10.08, PMF v. 5.0, Matlab v. 2021a, STATA v. 16, and SPSS v. 26 software.

2.3 Sample point collection and processing

In addition to the remote sensing and statistical data, this study also collected field data from 212 sampled locations across the study area. The point locations were chosen random sampling method (Lin and Kuwayama 2016; Maduekwe and de Vries 2019; Howell et al. 2020) considering the different geographical characteristics and 85% of were collected from urban areas while the remaining 15% were collected from semi-urban areas (Fig. 1). Each district had a minimum of at least 40 sampled locations. The sampling density was a bit lower in the north, north-eastern, and western parts of the study area due to thick forests and rivers (Haque et al. 2018; Iqbal et al. 2020; Pavel et al. 2021). The data collected from the sampled locations were recorded first in Microsoft Excel and was later imported and converted into ArcGIS v. 10.8 shapefile for further analysis.

For predicting the health impacts of child ARI and mortality, data from 60 *Upazila* (sub-districts) health complexes were collected from the MIS system of Ministry of Health. All the 60-health data were then converted into a GIS point shapefile, extracting the same locational values of all air pollutants and other variables. These 60 data points were used to run multi-regression analysis.

2.4 Positive Matrix Factorization (PMF)

Receptor models are mathematical methods for calculating the contribution of sources to samples according to their composition or properties (USEPA 2014). Many studies have used receptor models recently, and they have demonstrated their capacity to reliably identify potential ambient PM emission sources at a receptor site (Waked et al. 2014). The PMF model is a multivariate receptor model that uses a weighted least square approach to estimate the source profiles and their contributions (Paatero and Tapper 1994; Paatero 1997) and it is extensively used in determining the air quality (Tan et al. 2014; Zhang et al. 2015). In this study, PMF (v. 5.0) was used to quantify the contribution of various emission sources to $PM_{2.5}$ (USEPA 2014). The model requires two input files: one for the 'species' measured concentrations and another for the 'estimated uncertainty' of the concentrations (Sharma et al. 2016). Based on the PMF user guide, the equation below is used to identify the number of factors *p*, species profile *f*, and factor to contribution *g* (USEPA 2014):

$$x_{ij} = \sum_{k=1}^{p} g_{ik} f_{kj} + e_{ij} \qquad \text{Eq 01}$$

Where *i* and *j* are the number of samples and chemical species, and e_j is the residual of individual sample/species. The equation below is used for factor contribution and profiles.

$$Q = \sum_{i=1}^{n} \sum_{j=1}^{m} \left[\frac{x_{ij} - \sum_{k=1}^{p} g_{ik} f_{kj}}{u_{ij}} \right]^2 \qquad \text{Eq 02}$$

Where *Q* is a critical factor, showing *Q(true)* and *Q(robust)*.

Two input files, species concentration, and sample uncertainty are needed to run a PMF model. This study used the following uncertainty equations (USEPA 2014):

$$Unc = \frac{5}{6} \times MDL \qquad \text{Eq 03}$$

$$Unc = \sqrt{(\text{ Error Fraction } \times \text{ concentration })^2 + (0.5 \times MDL)^2} \qquad \text{Eq 04}$$

Where *Unc* is the concentration of each sample, *MDL* is the sample-specific method limitation, and error fraction is the percentage of measurement uncertainty.

2.5 Interpolation of point data

This paper used the point interpolation method to perform GIS mapping for all factors and visualize their spatial concentration in the study area. The Inverse distance weighted (IDW) interpolation method was selected, due to the near distances of each data point (Yu et al. 2019), to estimate the unknown values of new points surrounding the nearest known points in the study area. This is a very crucial point interpolation method used in many point source identification and public health data analyses (Feng et al. 2015; Hu et al. 2017; Huang et al. 2017; Iqbal et al. 2020). In this paper, factor 1 to 5 data was interpolated using ArcGIS v. 10.08 software. The calculation of IDW is in the equation below:

Eq 05

$$z_j = \frac{\sum_i \frac{z_i}{d_{ij}}}{\sum_i \frac{1}{d^{n}_{ij}}}$$

Where z_i is the value of a known data point, d_{ij} is the distance of a known point, z_j is the value at the unknown point.

2.6 Self-Organizing Map (SOM)

The SOM is thought to be a suitable artificial intelligence tool for extracting features because the input data is considered as a continuum rather than depending on correlation and cluster analysis (Liu et al. 2006). The SOM has been widely utilized for data downscaling and visualization in different areas (Kohonen 1982). Combining SOM with cluster analysis can help characterize various groupings of items that can logically be classified as characteristics (Nakagawa et al. 2020). Because SOM is stronger at classifying and recognizing patterns in elements, combining it with PMF could help SOM's findings to more accurately allocate contributions from various sources (Pearce et al. 2014; Dyson 2015; Katurji et al. 2015; Stauffer et al. 2016; Jiang et al. 2017). To imply such a concept, this paper used SOM to identify the pattern and concentration of each factorized data by investigating all neurons and their neighborhood relations using Matlab v. 2021a. In SOM, a similar color pattern shows a positive relationship while heterogeneous color shows a negative relationship, by clustering all data. In addition to this, a SOM shows the spatial distribution of each variable in 2-dimensional space by following the equation below:

$$\|x - m_c\| = min\{\|x - m_i\|\} \quad \text{Eq 06}$$

Where x is the input vector, m is the weight vector, and || || is the distance measure.

2.7 Machine learning in predicting $PM_{2.5}$

Machine learning is the technique for creating computer algorithms that can emulate human intelligence. It incorporates principles from several fields, including artificial intelligence, probability and statistics, computer science, information theory, psychology, control theory, and philosophy (Michel 1997; Bishop 2017). Generally, machine learning models are performed using numerous alternative algorithms to evaluate their effectiveness and select the best prediction. Algorithms of machine learning include Extremely Randomized Trees Regression (Extra Trees Regression), Random Tree, Stepwise linear, Linear Regression, Extreme Gradient Boosting (XGBoost), Least Absolute Shrinkage, and Selection Operator (Minh et al. 2021). In this paper, three supervised classifiers of machine learning (1) Random Tree (Joharestani et al. 2019) (2) Extreme gradient boosting (Ma et al. 2020a), and (3) Stepwise Linear (Chen et al. 2019a) were used to predict the relationship between $PM_{2.5}$ and fourteen variables using Matlab v. 2021a. To run these three classifiers, this paper followed several steps (Minh et al. 2021). The first step was to divide the pre-processing data into training and test data sets. Second, using the training dataset, the machine learning model was trained using each algorithm. The test dataset was then placed to evaluate the training efficiency of each method. The final stage was to assess each model's performance using the assessment parameters. Lastly, the Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Mean Square Error (MSE), and Coefficient of Determination (R^2) were used to evaluate the model performance of these three classifiers. The key equations of these performance indexes are given below:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |P_i - O_i| \quad \text{Eq 07}$$

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (P_i - O_i)^2} \quad \text{Eq 08}$$

$$\text{MSE} = \frac{1}{N}\sum_{i=1}^{N} (y_i - \hat{y}_i)^2 \quad \text{Eq 09}$$

$$R^2 = \frac{\sum_{i=1}^{n} (P_i - \overline{O})^2}{\sum_{i=1}^{n} (O_i - \overline{O})^2} \quad \text{Eq 10}$$

Where *n* is the number of sample data points (212), P_i is the predicted data point, and O is the mean of the observed data. $PM_{2.5}$ was used as a dependent variable while 14 variables were used as the independent variable (Doreswamy et al. 2020).

2.8 Health impact assessment

The multiple-regression analysis is a widely used statistical tool to predict the relationship between one dependent and numerous independent variables. An array of researchers has used multi-regression analysis for predicting the relationship between air pollutants, particularly $PM_{2.5}$, and various types of health data (Azad 2008; Naz et al. 2015; Egondi et al. 2018; Ulavi and Shiva Nagendra 2019; Hassan et al. 2021). The basic calculation of a multi-regression analysis is:

$$Y = a + b_1X_1 + b_2X_3 + \cdots + b_nX_n + e \quad \text{Eq 11}$$

Where, Y is the dependent variable, $b_1 \ldots b_n$ is the beta-coefficient, $X_1 \ldots\ldots X_n$ is the independent variable and e is the residual error.

The equation below was used to estimate the relationship between child ARI and other variables:

$$\begin{aligned}\text{ARI\_U5} = {} & -168 + 27.3\ \text{NOx} - 10.0\ \text{BC} + 12.6\ CH_4 + 5.11\ \text{AOT} - 0.042\ SO_2 \\ & - 0.775\ O_3 + 0.00008\ \text{Sellte\_Den} - 0.1112\ \text{DEM} + 0.00621\ \text{Brick\_Den} \\ & - 2.827\ \text{Windspeed} + 0.00784\ \text{Rainfall} + 3.56\ \text{CO} - 0.00783\ NO_2 \\ & - 0.0206\ PM_{2.5}\end{aligned} \quad \text{Eq 12}$$

Likewise, the below equation was used to estimate the relationship between child mortality and other variables:

$$\begin{aligned}\text{Mortality} = {} & -2384 + 139\ \text{NOx} - 238.3\ \text{BC} + 321.1\ CH_4 + 29.3\ \text{AOT} - 2.66\ SO_2 \\ & - 3.54\ O_3 + 0.01969\ \text{Sellte\_Den} + 0.492\ \text{DEM} + 0.0400\ \text{Brick\_Den} \\ & - 2.22\ \text{Windspeed} - 0.0510\ \text{Rainfall} + 16.4\ \text{CO} - 0.0265\ NO_2 \\ & + 0.651\ PM_{2.5}\end{aligned} \quad \text{Eq 13}$$

Finally, this paper used R^2 (Mukta et al. 2020), beta-coefficients (Thurston et al. 2011), and *p*-values (Wang et al. 2021) to understand the relationship between these two models and the internal robustness of each variable against the dependent variable. Both STATA v. 16 and SPSS v. 26 were used to complete multiple-regression analysis.

3. Results and Discussion

3.1 Descriptive statistics of all sample points

The descriptive statistics of each variable are shown in Table 2. Data analysis highlighted that the mean concentration of all the variables ranges from 0.55 to 82.11. Among all the variables, the mean concentration was the highest for CO (82.11) followed by Poverty (73.59), $PM_{2.5}$ (65.19), NO (56.63), Rainfall (27.46), and Settlement density (23.52). The lowest mean concentration was for AOT (0.55). In addition, in comparison with other variables, Standard Deviation was higher for Brickfield (51.99), Rainfall (44.11), No (12.91), and Settlement (10.88), which indicated that these variables have a higher level of spatial variation compared to other variables. Also, the Coefficient of variation is high in the brick field (2.17).

Table 2 descriptive statistics of all parameters used in this paper.

	Mean	Std. Deviation	Coefficient of Variation
NOx	1.19	0.00	0.00
BC	0.67	0.01	0.02
CH_4	6.33	0.03	0.00
AOT	0.55	0.04	0.08
SO_2	1.02	0.37	0.36
O_3	23.28	0.77	0.00
Settlement	23.52	10.88	0.46
DEM	10.02	3.56	0.35
Brickfield	23.67	51.99	2.17
Poverty	73.59	2.86	0.03
Wind Speed	2.10	0.20	0.09
Rainfall	27.46	44.11	0.16
CO	82.11	0.09	0.00
NO	56.63	12.91	0.22
$PM_{2.5}$	65.19	1.26	1.58

3.2 Correlation analysis

Spearman's Correlation method was applied among the fifteen variables (including $PM_{2.5}$) to know which variables were correlated positively and negatively with each other (Fig. 2). From the data analysis, it was observed that NO and CH_4 (0.83), CO and DEM (0.58), and CH_4 and BC (0.56) had strong positive relationships (p>0.05). On the other hand, NO and CO (-0.80), CO and CH_4 (-0.69), CO and Rainfall (-0.68), and Rainfall and DEM (-0.82) showed a moderate to very strong negative relationships (p>0.05). Moreover, Brickfield was one of the weaker variables with all other variables except for $PM_{2.5}$ (0.25). $PM_{2.5}$ was positively correlated with NO (0.55), BC (0.45), CH_4 (0.38) and NOx (0.22), while negatively correlated with Rainfall (-0.10), CO (-0.33), and SO_2 (-0.24).

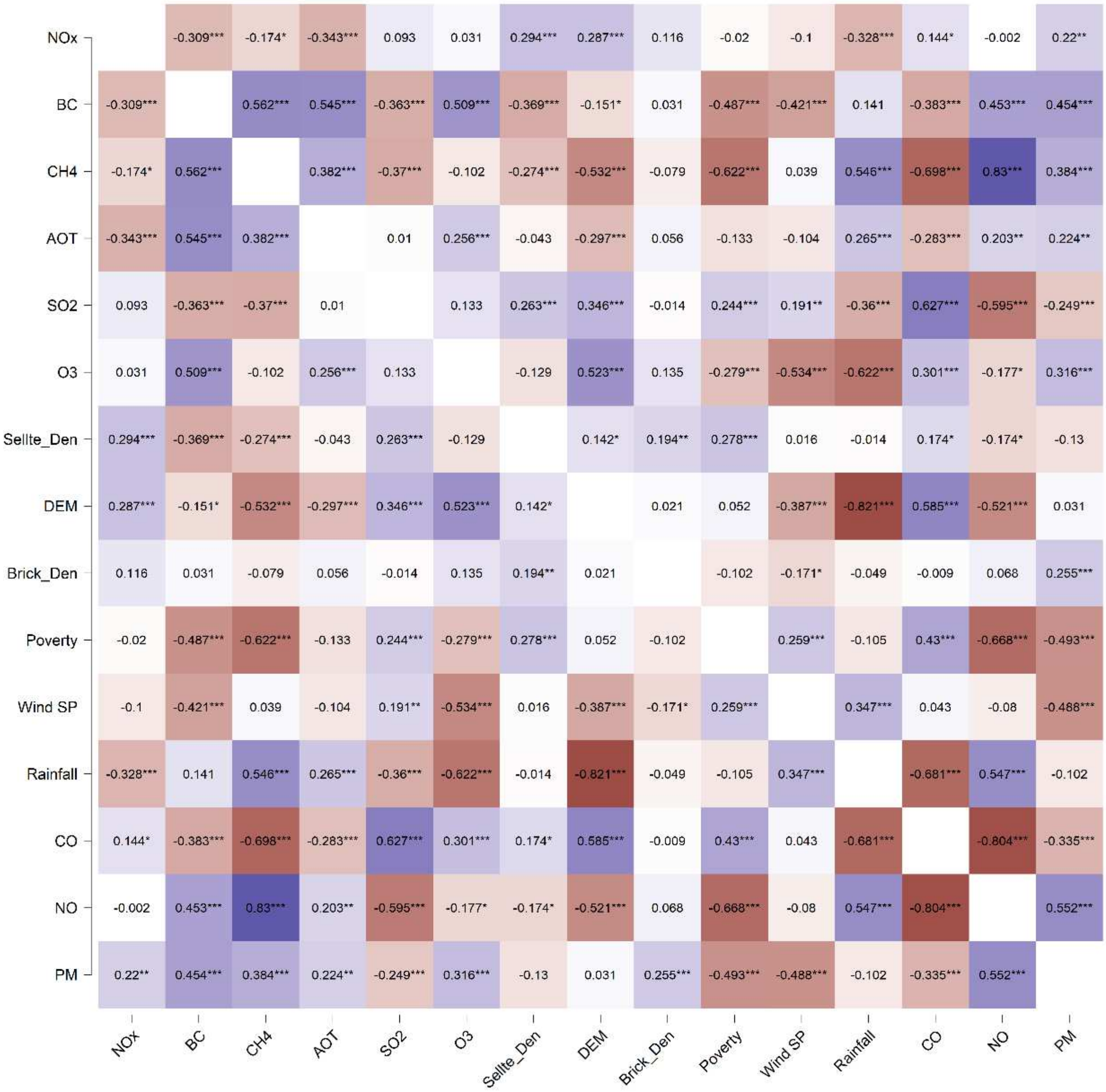

Fig. 2 Results of Spearman's Correlation of all variables, where *p < .05, **p < .01, ***p < .001

3.3 PMF process for source appointment

As state previously, to classify and quantity the key sources of $PM_{2.5}$ in the study area and to determine the major fingerprint of each pollutant, this study used PMF Model (version 5.0) using 19 variables (Fig. S1). First, about 212 sample points were added to the model along with 212 uncertainty data (Table S1). Then, the model was executed 20 times until the following conditions were met: having 1-5 factors, considering the base random seed of 98, 0% extra modeling uncertainty, bootstraps value of 100, and minimum R^2 value of 0.2. All the input data were defined as "Strong" variables, except for APR, Pop density, LST, HCHO, and poverty. Five variables (APR, population density, HCHO, LST, and water vapor) were excluded from this analysis as they were not statistically significant. Finally, the standard model was selected when the Q values reached close to +1. As an outcome of the PMF model, the coefficient of determination (R^2) between the observed and predicted value of each pollutant is shown in Fig. 3.

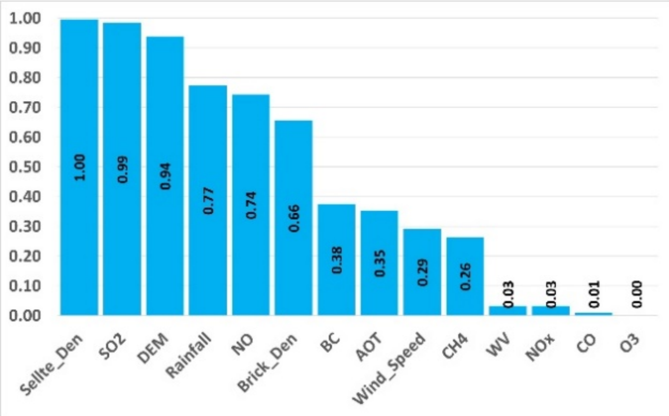

Fig. 3 Relationship between the observed and predicted of each pollutant generated from the PMF modeling.

The R^2 values of settlement density (1.00), SO_2 (0.99), DEM (0.94), Rainfall (0.77), NO (0.74) and Brickfield density (0.66) were found to be the most correlated and significant variables in this model. Black Carbon, AOT, Wind Speed, and CH_4 had lower R^2 values ranging from 0.26 to 0.38.

3.3.1 Factor profile analysis

In this study, the first factor was highly dominated by NO (70%) and Rainfall (60%). It suggests that both traffic, diesel engine, and climatic factors are the responsible sources for $PM_{2.5}$. Different emissions from traffic and vehicles are being significantly contributed to about 2/3rd of air pollution in Dhaka and its surrounding areas (Hassan et al. 2019). A study by (Pavel et al. 2021) showed that NO contributes ~74% to factor 1. (Begum and Hopke 2019) suggested that rainfall and other metrological factors have a critical influence on $PM_{2.5}$ in Bangladesh. Moreover, about 50% of loading of CO, water vapor, O_3, BC, and NOx were found as second contributed sources in this factor 1 (Fig. 4). These gaseous substances have a substantial role as the major source of $PM_{2.5}$ (Zhang et al. 2013; Rahman et al. 2017; Samek et al. 2017). Like other particulate matter, BC emissions mainly from the incomplete combustion of fossil fuels, biofuels, and biomass have long-term negative implications for both public health and global climatic changes (Haque et al. 2018). In addition to factor 1, unplanned urban development, uncovered construction materials, huge population, emissions from different industries, less vegetation coverage, narrow road transportation systems, brick fields, and transboundary air flow from, neighboring countries have all played major catalysts to increase $PM_{2.5}$ in Dhaka and the surrounding districts in the study area (Begum et al. 2009; Hasan et al. 2013; Rahman et al. 2017; Iqbal et al. 2020).

The second factor was depicted by SO_2 (55%) and settlement density (54%). The main sources of generating SO_2 are mainly anthropogenic processes including motor vehicles, emissions from brick fields, industries, and urbanization (Zhang et al. 2013; Mukta et al. 2020). (Huang et al. 2017) used a land use regression model in China using three independent variables to predict the $PM_{2.5,}$ where SO_2 explained the highest variance (83%). (M.M. et al. 2018) mentioned that ready-made garment (RMG) factories in Bangladesh release huge amounts of wastes, liquid particulates, and gaseous substances, which are the key components for increasing SO_2 in major urban areas. The consumption level of liquefied petroleum gas (LPG) and electricity at the urban household level is very common in Bangladesh. It is expected that about 36.5% of the total country has been speculated to be urbanized (BBS 2020). The emissions from LPG and electricity have a significant outcome on

$PM_{2.5}$ as well as overall air pollutions (Muindi et al. 2016; Fotheringham et al. 2019). In addition, the digital elevation model (DEM) was loaded moderately by 50%, while settlement density was more than 50% (Fig. 4).

The third factor was dominated by population density with a loading of 60%, which was a high maker in this group. Many researchers concluded that a higher density of urban population increases emissions from cars, traffic patterns, slow travel speeds, compact roads, and unplanned urbanization, which are mainly the triggering factors for $PM_{2.5}$ (Chueinta et al. 2000; Han and Sun 2019; Nouri et al. 2021). As a result of migration from climate changes across Bangladesh, more than 4.1 million people have been displaced and relocated to urban areas (Khan et al. 2021). These populations are being created into huge urban slums and formed informal economic activities for their daily lives and livelihoods. Slums and low-income settings have a significant role to enhance $PM_{2.5}$ due to their household fuel consumption, use of stoves, and poor ventilation systems (Gaita et al. 2014; Muindi et al. 2016). In addition to factor 3, brickfield contributed about 30% of loading (Fig. 4). There are more than 4,500 brick fields in and around the Dhaka division, which are mainly run by traditional coal and biomass fuels (Haque et al. 2018). In turn, these brick fields are producing various gaseous and particulate matters in the study area.

The fourth factor was characterized by settlement density (49%). This is one of the crucial urban problems in Bangladesh. Increasing population, settlement, and manufacturing agglomeration for overall local and regional economic development have fetched a signature of emitting pollution from diverse sources (Zhao et al. 2021). Different urban landscape patterns like edge density (ED) and patch density (PD) are influenced by $PM_{2.5}$ in urban settings (Wu et al. 2015). As a second major source of this factor, brick field (Fig. 4) was dominated by 30% loading in the north and central parts of the study area. (Al Nayeem et al. 2019) mentioned that about 92% of brick fields use fixed chimney kilns (FCKs), which emit dusts, fine coal, organic matters, and some gaseous particles. Even the location of roadside brick fields increases the concentration of $PM_{2.5}$ along with rapid urbanization activities.

The fifth factor was dominated by brick field and SO_2 (Fig. 4) with the loading of 60% and 10%. As a major by-product of brickfield, SO_2 pollutes the quality of air. About 7,500 brick fields across Bangladesh are being operated illegally, violating the Environmental Conservation Rules (DoE 2022). Each year, more than 20,00,000 metric tons of raw fire wood and low-quality coals are burned in these brick fields by the traditional brick-making processes, which are emitting SO_2, CO_2, and TSP (Ahmed and Hossain 2008). It is noted that traditional, non-eco-friendly, and non-compliant brick kilns enhance $PM_{2.5}$ and are susceptible to unhealthy atmospheric conditions (Saha and Hosain 2016). (Thygerson et al. 2019) concluded that brick fields along with population density and urban traffic have far exceeded the standard levels in Bhaktapur, Nepal. In turn, the emitted gaseous and particulate matters have impacted public health, crops, vegetation, and land uses (Rahman 2022).

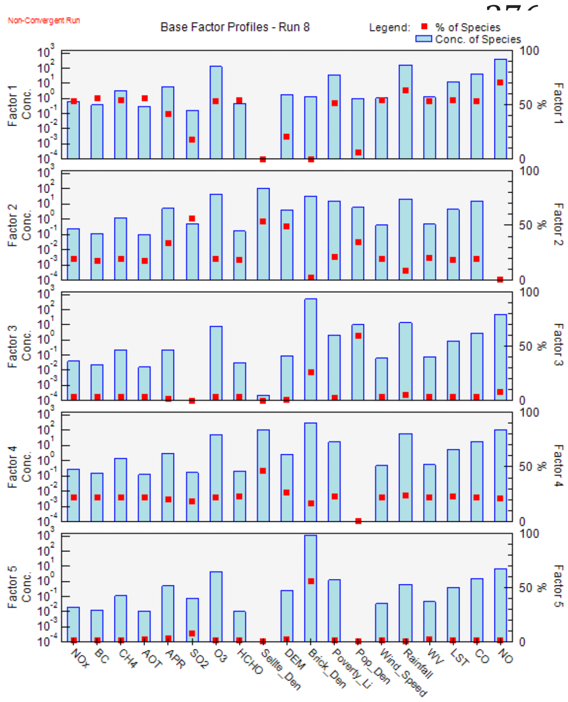

Fig. 4 Factor profile and source contribution from the PMF modeling.

3.3.2 GIS mapping of all factors and their spatial concentration

To understand the spatial distribution of each factor (Table S1), the inverse distance weighted (IDW) interpolation method was used. The highest spatial concentration of factor 1 was found in the central to the northern parts of the study area, covering central Dhaka and Gazipur districts (Fig. 5a). The mean concentration of this factor was 1.00, the standard deviation was 0.49 with a 0.06 confidence interval (95%). Both areas have been exposed to a high concentration level of emission due to huge traffic, construction works, the RMG sector, and massive industrial activities. Moreover, unplanned urbanization activities, reduced vegetation areas, and inadequate transportation systems have all created atmospheric pollutions, particularly $PM_{2.5}$ (Hasan et al. 2020). In Dhaka, the recent urbanization and various city development projects like elevated express, flyovers, new buildings, an extension of the airport, shrinkage of water bodies, uncontrolled land development and landfills have all caused heavily deteriorating air quality by releasing different gaseous and particulate matters (Nahar et al. 2021). On the other hand, Factor 2 was loaded in the central, south, and eastern parts of the study area, including Dhaka, Munshiganj, and Narshingdi districts (Fig. 5b). The mean concentration of this factor was 0.99, the standard deviation was 0.38 with a 0.52 confidence interval (95%). It is noted that most of the ready-made garment factories, dyeing industries, and informal machine factories are located in both Munshiganj and Narshingdi districts, which are producing vast amount of emissions (M.M. et al. 2018). Even more, enormous uncontrolled brick fields have been installed and are operating in these areas (Haque et al. 2018; Rahman 2022). In Factor 3 (Fig. 5c), a pollution hotspot was found in central Dhaka and in the southern part of the Narayanganj districts, with a mean concentration value of 1.00 and standard deviation of 3.28 with a 0.44 confidence interval (95%). Several waterbodies have been flowing in these areas. Due to lower transportation costs for firewood, supply bricks, cheaper labor costs, and the availability of quality clay, brick kilns were established in these areas (Hassan et al. 2019). Factor 4 (Fig. 5d) was dominated in the northern part of Narshingdi and the western part of Dhaka districts. The mean concentration of this factor was 0.99 and the standard deviation was 1.51 with a 0.20 confidence interval (95%). Many studies have found that both of these cities are releasing different gaseous and particulate matters and are affecting the local public health, ecology, and ecosystems (Hasan et al. 2013; Al Nayeem et al. 2019; Iqbal et al. 2020; Pavel et al. 2021) . In Factor 5, sporadic loading of a few variables was found in few areas of Dhaka, Munshiganj, and Narayanganj districts (Fig. 5e). The mean concentration of this factor was 0.99 with a standard deviation of 3.17 with a 0.43 confidence interval (95%). Factor 5 was not found to be a major concern in identifying the key sources of $PM_{2.5}$ in this study.

From this analysis, it is observed that Dhaka, Gazipur, Narayanganj, and Narshingdi are the prominent areas for generating different sources of $PM_{2.5}$ due to similar anthropogenic, natural, and development patterns. While the results of this analysis are satisfactory considering the discussion and published literature, a major limitation is not considering seasonal concentration levels. This study used average annual data from earth observing satellite sensors. Perhaps, seasonal data may reveal a better understanding of possible sources. Further investigation is needed to explore this variation.

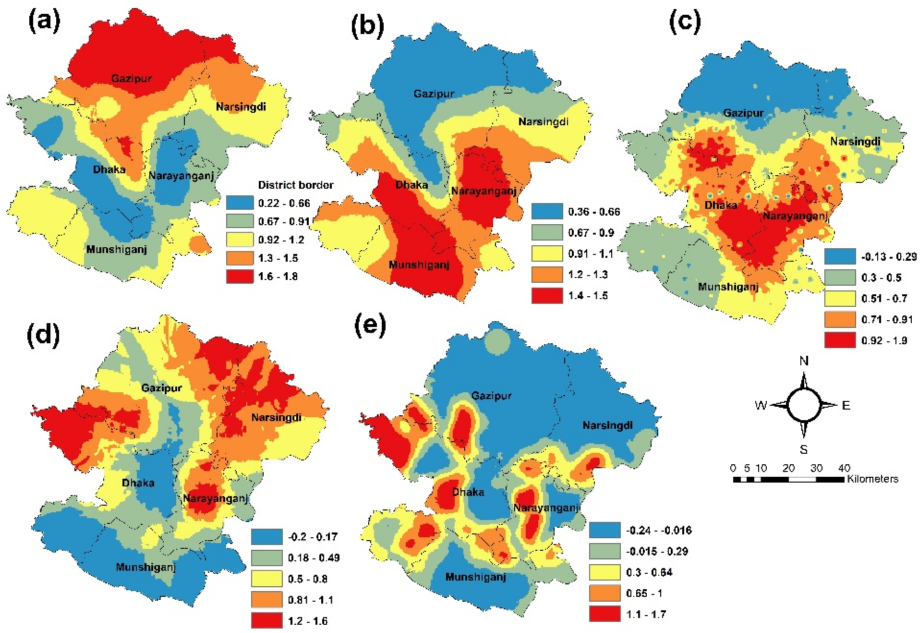

Fig. 5 The spatial concentration mapping of all five factors from the PMF modeling, where factor 1 (5a), factor 2 (5b), factor 3 (5c), factor 4 (5d), and factor 5 (5e).

3.4 Concentration mapping using SOM

In Figure 06, the output of the component planes of the SOM analysis is shown in detail. Color-coded SOM planes were created to highlight the concentration of the given variables for each SOM unit. The more similar the properties of samples are, the smaller the hexagonal space is. A similar color indicates a positive correlation between variables in a component plane, while dissimilar colors indicate negative correlations.

In most of the samples, the upper right corner of the map showed a higher concentration. No variable was found to be highly concentrated in the left part. The variables O_3 (Fig. 6g), Poverty (Fig. 6l), NO (Fig. 6j), Settlement Density (Fig. 6g), CH_4 (Fig. 6c), and DEM (Fig. 6i) showed a similar pattern of concentration which indicated a common association and a positive correlation among these variables. For BC (Fig. 6b), most of the portion of the component plane had high concentrations. In addition, the brick field had lower concentrations at the lower-left corner which was analogous to wind speed. The concentration of $PM_{2.5}$ (Fig. 6o) was similar to NO (Fig. 6j), CO (Fig. 6e), and AOT (Fig. 6d). The concentration of $PM_{2.5}$ was higher in the north-eastern part.

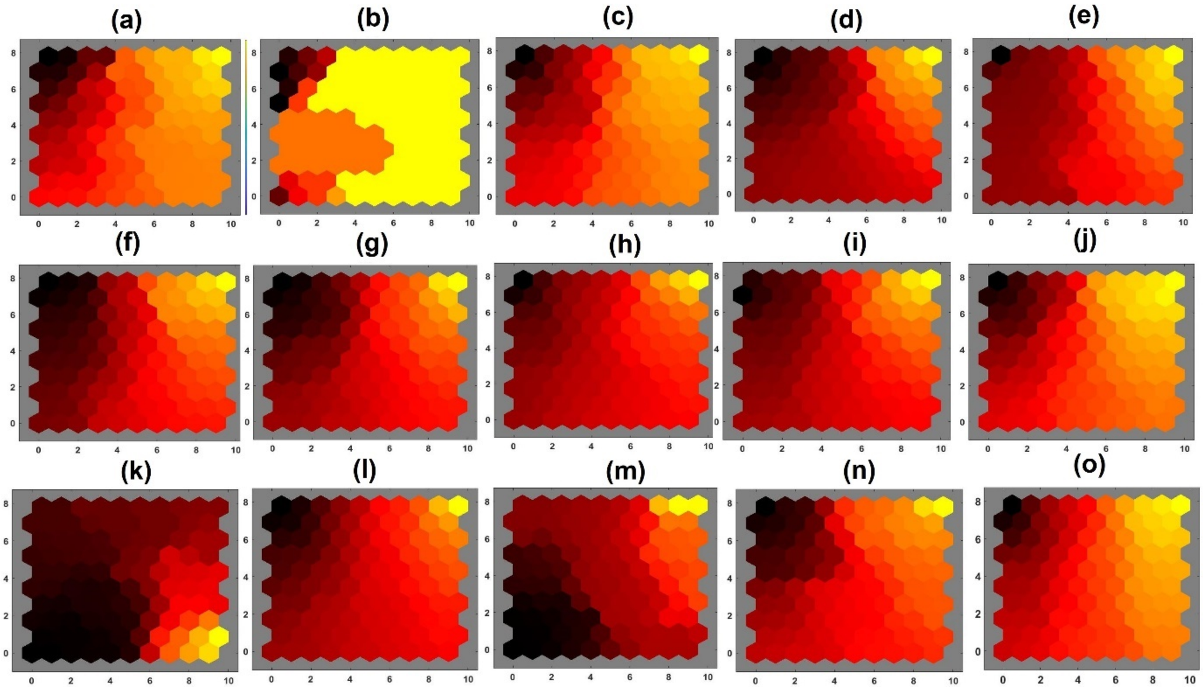

Fig. 6 The spatial concentration of SOM map of all variables derived from the PMF modeling, tagging NOx (a), BC (b), CH_4 (c), AOT (d), CO (e), SO_2 (f), O_3 (g), Settlement density (h), DEM (i), NO (j), Brickfield (k), Poverty (l), Windspeed (m), Rainfall (n) and $PM_{2.5}$ (o).

3.5 Machine leaning

This study used random forest, extreme gradient boosting, and stepwise linear supervised classifiers of a machine learning tool to predict $PM_{2.5}$ using fourteen factorized independent variables derived from the PMF modeling. About 70% of training data used (to discover the relationship between dependent and independent variables) and 30% of test data (to measure the accuracy of the hypothesis) of the 212 sample points were selected to run the whole machine learning process using SPSS v. 26 and Matlab v. 2021. For the three classifiers, a 10-fold cross-validation method was followed to fetch better results (Feng et al. 2015; Hu et al. 2017). Finally, the MSE (mean squared error), RMSE (root-mean-square error), MAE (mean absolute error), and R^2 (coefficient of determination) were used to check the good fit of the results and compare the three supervised classifiers that were found to be suitable for this study (Deters et al. 2017; Doreswamy et al. 2020; Godoy et al. 2021). The key results of these three classifiers are discussed in the following sections.

3.5.1 Random Forest

The results of the observed and predicted values of $PM_{2.5}$ in the Random Forest classifier (Fig. 7a) was distributed around the reference line of X and Y, showing a very close relationship. This prediction revealed a strong relationship between $PM_{2.5}$ and other fourteen variables (the prediction matrix was R^2= 0.85, MAE= 0.37, RMSE= 0.54, and MSE= 0.29) (Table 2). (Yazdi et al. 2020) found a similar R^2 (0.85) result while they were working to predict Fine Particulate Matter ($PM_{2.5}$) in the Greater London Area using a similar machine learning technique where they used several climate variables. (Deters et al. 2017) suggested that a better prediction of

$PM_{2.5}$ derives when the climatic conditions are used (like strong wind or high levels of precipitation). This current paper followed the same notion by adding some climatic variables (rainfall and wind speed). (Dai et al. 2021) estimated the R^2 (0.85) for the annual prediction of $PM_{2.5}$ in the US, which is similar to this paper's result too.

3.5.2 Extreme gradient boosting

The relationships between observed and predicted values of $PM_{2.5}$ using an extreme gradient boosting classifier is shown in Fig. 7b. The relationships between $PM_{2.5}$ and the independent variables were reasonably strong (R^2= 0.81, MAE= 0.32, RMSE= 0.51, and MSE= 0.26) (Table 3). Pan 2018 researched to predict the hourly PM2.5 concentration in Tianjin city, China and found a strong R^2 (0.95) between observed and predicted values. The result found in this study was a bit lower since Pan (2018) used hourly predictions of $PM_{2.5}$. (Dai et al. 2021) used spatio-temporal feature selection to predict $PM_{2.5}$ in the Fenwei plain, China. Their estimated R^2 was about 0.87 with a higher RMSE (11.07). Interestingly, an R-value of –0.65 was found in the eastern Chinese city of Shanghai. (Ma et al. 2020b) concluded this negative value was estimated because predicted $PM_{2.5}$ and meteorological factors were smaller than other pollutions used as independent variables. On the other hand, (Pan 2018) found a robust R^2 (0.95) using the extreme gradient boosting classifier to $PM_{2.5}$ concentration in Tianjin city, China.

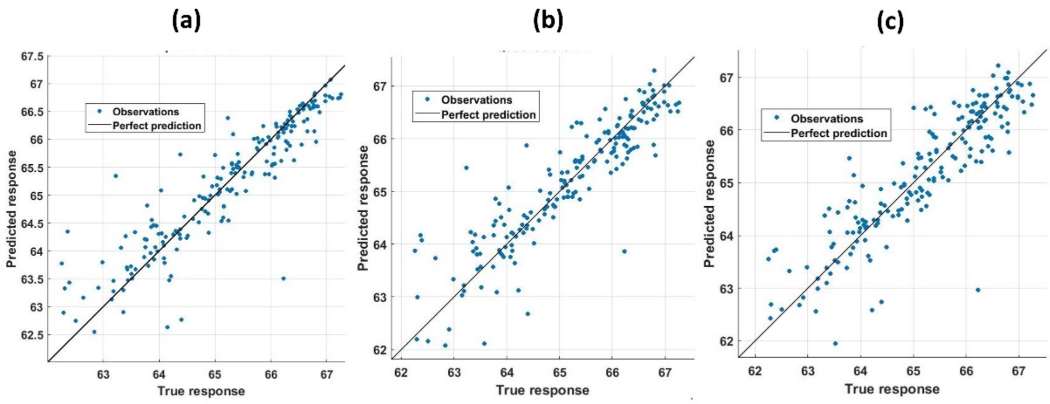

Fig. 7 Scatterplot of predicted and observed values of $PM_{2.5}$. Figure 7a is the Random Forest, 7b is the Extreme gradient boosting while 7c is the Stepwise linear.

Table 3 the key parameters derived from three classifiers to assess the prediction accuracy.

Classifier	MSE	RMSE	MAE	R^2
Random forest	0.29	0.54	0.37	0.85
Extreme gradient boosting	0.26	0.51	0.32	0.81
Stepwise Linear	0.38	0.62	0.43	0.76

3.5.3 Stepwise Linear Classifier

The observed and predicted values of $PM_{2.5}$ using a Stepwise linear classifier is shown in Fig. 7c. The prediction matrix was for R^2= 0.76, MAE= 0.43, RMSE= 0.62 and MSE= 0.38 (Table 3). This classifier showed a slightly less performance accuracy than the other two classifiers used in this study. (Wu et al. 2015) studied how urban landscape patterns affected $PM_{2.5}$ in Beijing, China using stepwise linear modeling. They found a lower R^2

(0.65) compared to the current study due to different urban landscapes associated with other variables like air follow, traffic congestions, population density, etc. (Chen et al. 2019b) predicted the annual average of $PM_{2.5}$ in ESCAPE sites in Europe using 16 algorithms of machine learning with even lower R^2 (0.61) due to the inclusion of NO_2. The R^2 (0.70) was calculated by (Ulavi and Shiva Nagendra 2019) in Chennai, India. While they used several meteorological variables, their coefficient of determination was the closest to the results obtained in this study.

Long-term prediction of $PM_{2.5}$ considering different variables with temporal sessional variations can be a better management tool for local and regional level air pollution mitigation. This study has not used any temporal sessional data. The inclusion of sessional and other data is suggested in its future study.

3.6 Health Impact analysis

The possible health impacts of mortality and ARI of children under the age of five due to different air pollutants were estimated separately using multiple-regression analysis.

Table 4 highlights the relationship between child mortality and fourteen air pollutants. The results show that the relationship was moderately statistically significant, with an R^2 of 0.55. The significant *P*-values were in BC, CH_4, and settlement density at 0.01, 0.00, and 0.02, respectively with a 95% confidence level. The beta-coefficient results suggests that if 1 unit of NOx, CH_4, and AOT increases in the air, then it will heavily affect 139 children, 321 children, and 29 children. Currently, 29 children die per 1,000 live births in Bangladesh (UNICEF 2022) due to various factors. In this analysis, $PM_{2.5}$ did not affect child mortality. However, (Naz et al. 2015) found that the amount of outdoor air pollution exposed is a critical factor in child mortality in the country. (Dominici et al. 2002) studied the relationship between $PM_{2.5}$ and mortality in 88 largest cities in the US and found a strong R^2 with positive coefficients for O_3 and NO_2. There is a significant difference between higher and lower air polluted areas. (Egondi et al. 2018) found that in Nairobi (Kenya), child mortality is higher in regions with poor economic conditions and high air pollutions areas irrelevant of gender. The national panel child mortality and $PM_{2.5}$ data and their statistical relationship from 16 Asian countries revealed that R^2 values were 0.75 and 0.87 in WHO and World Bank datasets, respectively (Anwar et al. 2021). Daily mortality increases when the concentration of $PM_{2.5}$ increases which was found by a pooled concentration-response analysis conducted in 652 cities in the world (Liu et al. 2020). The fixed-effect model and spatial econometric modeling can be a good way to measure the relationship between $PM_{2.5}$ and infant mortality. (Li et al. 2021) used these models in China and found that R^2 was 0.70 while urbanization (*p=0.00*), hospital beds per ten thousand persons (*p=0.01*), and hospital agencies per ten thousand persons (0.04) were significant predictors at $p < 0.05$ for child mortality.

Table 4 model parameters derived from the multi-regression analysis between mortality and 14 air pollutants

Variables	Coef	SE Coef	95% CI	T-Value	P-Value	VIEW	R^2
Constant	-2384	2645	(-771, 294)	-0.90	0.372		
NOx	139	258	(-380, 657)	0.54	0.593	1.50	
BC	-238.3	94.6	(-428.9, -47.6)	-2.52	0.015	7.70	
CH_4	321.1	70.9	(178.2, 464.0)	4.53	0.000	12.96	
AOT	29.3	25.0	(-21.1, 79.8)	1.17	0.247	3.66	
SO_2	-2.66	3.46	(-9.63, 4.30)	-0.77	0.446	4.67	
O_3	-3.54	2.02	(-7.60, 0.53)	-1.75	0.087	3.93	
Sellte_Den	0.019	0.008	(0.002, 0.037)	2.27	0.028	2.61	
DEM	0.492	0.301	(-0.115, 1.100)	1.63	0.109	2.38	0.55

Brick Den	0.040	0.028	(-0.017, 0.097)	1.41	0.164	1.55
Windspeed	-2.22	4.03	(-10.34, 5.89)	-0.55	0.584	2.11
Rainfall	-0.051	0.036	(-0.125, 0.022)	-1.40	0.169	4.72
CO	16.4	34.7	(-53.5, 86.3)	0.47	0.639	17.60
NO_2	-0.026	0.024	(-0.075, 0.022)	-1.09	0.280	24.48
$PM_{2.5}$	0.651	0.581	(-0.51, 1.82)	1.12	0.268	1.51

Table 5 revealed that ARI and 14 air pollutants had a strong statistically significant relationship (R^2 = 0.75. The significant *P*-values were in O_3, DEM, wind speed, NO_2 and $PM_{2.5}$ at 0.01, 0.02, 0.00, 0.04 and 0.02, respectively with 95% confidence level. The beta-coefficient results also suggested that if 1 unit of NOx, CH_4, and AOT increases in the air, then 27, 12, 3, and 5 children, respectively will be affected heavily by ARI-related diseases. In this analysis, $PM_{2.5}$ had a strong effect on ARI. ARI is one of the leading causes of death in children under 5 years old with reported annual total deaths of more than 50,000 children in Bangladesh (Hassan et al. 2021). The main reasons for ARI in Bangladesh are mainly industrial pollutions, emissions from motor vehicles, lack of mother's knowledge, vitamin A deficiency, etc. (Azad 2008). (Wang et al. 2021) concluded that a 1 μg/m^3 increase in $PM_{2.5}$ concentration may cause a 1.316 of hospital visits due to ARI in Taiwan. Traffic pollutions, $PM_{2.5}$, and O_3 are known to increase the lower and upper respiratory infections in early life in Altana, USA (Darrow et al. 2014). (Samet et al. 2000) found that O_3 did not affect ARI-related deaths in 20 cities in the USA. This is contradicting the results found in this study. However, they found that other air pollutants like $PM_{2.5}$, NO, SO and CO have influences on the ARI. On the other hand, exposure to $PM_{2.5}$, NOx and NO_2 have a significant role in increasing the risk of respiratory infection (Kirwa et al. 2021).

Table 5 model parameters derived from the multi-regression analysis between ARI and 14 air pollutants

Variables	Coef	SE Coef	95% CI	T-Value	P-Value	VIF	R^2
Constant	-168	410	(-993, 658)	-0.41	0.684		
NOx	27.3	39.9	(-53.1, 107.6)	0.68	0.498	1.50	
BC	-10.0	14.7	(-39.5, 19.5)	-0.68	0.498	7.70	
CH_4	12.6	11.0	(-9.5, 34.7)	1.15	0.257	12.96	
AOT	5.11	3.88	(-2.70, 12.92)	1.32	0.194	3.66	
SO_2	-0.042	0.536	(-1.121, 1.037)	-0.08	0.938	4.67	
O_3	-0.775	0.313	(-1.405, -0.144)	-2.48	0.017	3.93	
Sellte_Den	0.008	0.0135	(-0.026, 0.029)	0.06	0.953	2.61	
DEM	-0.112	0.467	(-0.205, -0.017)	-2.38	0.022	2.38	0.75
Brick Den	0.621	0.438	(-0.262, 0.0504	1.42	0.164	1.55	
Windspeed	-2.827	0.624	(-4.084, -1.570)	-4.53	0.000	2.11	
Rainfall	0.0784	0.056	(-0.003, 0.019)	1.39	0.172	4.72	
CO	3.56	5.37	(-7.26, 14.38)	0.66	0.511	17.60	
NO_2	-0.078	0.0037	(-0.539, -0.027)	-2.09	0.043	24.48	
$PM_{2.5}$	0.020	0.089	(0.201, 0.160)	0.23	0.002	1.51	

4. Summary and Conclusion

This study has assessed the different air pollutants, environmental, climatic, social, and health variables ($PM_{2.5}$, NO, AOT, SO_2, CO, BC, O_3, CH_4, DEM, rainfall, wind speed, settlement density, brickfield, poverty, health) for identification of possible sources, patterns, and health impacts of $PM_{2.5}$ in the central region of Bangladesh. To fill up the main knowledge gap, this study had four major objectives: (1) to determine the key sources of $PM_{2.5}$, (2) to identify the core concentrated areas of the sources, (3) to predict the $PM_{2.5}$ using factorized data, and (4) to investigate the impacts of child mortality and ARI due to $PM_{2.5}$ and other air pollutants. The main outcomes of the study are as follows:

- GIS, PMF, SOM, machine learning, and multi-regression analysis derived reliable outcomes for the study.
- $PM_{2.5}$ was correlated positively with NO (0.55), BC (0.45), CH_4 (0.38) and NOx (0.22), while correlated negatively with rainfall (-0.10), CO (-0.33), and SO_2 (-0.24).
- In PMF modeling, the R^2 values of settlement density (1.00), SO_2 (0.99), DEM (0.94), Rainfall (0.77), NO (0.74) and Brickfield density (0.66) were found to be the most correlated and signified variables.
- Factor 1 (NO, CO, Rainfall, O_3, AOT, CH_4, and BC) and Factors 2 (SO_2, settlement density, and DEM) were dominant in identifying the key sources of $PM_{2.5}$, while Factor 3 was dominated by only population density and brickfield.
- The central parts of Dhaka, the northern parts of Munshiganj, the western parts of Narshingdi and Narayanganj, and the southern parts of Gazipur districts were the highly concentrated areas due to diverse pollutant sources.
- In SOM mapping, most of the variables were concentrated in the north-eastern, central, and south-eastern parts of the study area, where NOx, CH_4, AOT, CO, settlement density, DEM, NO, Poverty, and $PM_{2.5}$ have similar concentration patterns.
- The prediction of $PM_{2.5}$ using machine learning was significant, showing reasonable R^2 for random forest (0.85), extreme gradient boosting (0.81) and stepwise linear (0.76).
- The impact of $PM_{2.5}$ on child ARI was significant ($p=0.002$) while the R^2 was 0.75. However, the impacts of $PM_{2.5}$ on child mortality was not significant ($p=0.268$) while the R^2 was 0.55. However, other variables like BC, CH_4, settlement density, O_3, DEM, wind speed, and NO_2 were critical for both child mortality and ARI.

In a country like Bangladesh, where air pollution and $PM_{2.5}$ data are limited and sparse, the results found in this study will be useful for local and regional level $PM_{2.5}$ mitigation and implementation plans. The government of Bangladesh, concerned ministries, UN bodies, and local and international NGOs may use these outputs for reducing health impacts (particularly child mortality and ARI), and for enhancing the environmental health in the study area as well as in the region. In addition to this, the overall methodology used can be replicated in similar urban and semi-urban settings with additional data. Future studies should also utilize additional data and parameters on ecological, seasonal air pollutants, and economic factors.

Supplementary Information a supplementary data sheet is attached to the main paper.

Acknowledgements The authors also deeply acknowledge to NASA and EPA for their freely available datasets.

Author contribution- Shareful: model conceptualization, methodology, data collection, analysis, writing the original draft. Rahman: review and editing. Amir: methodology and editing, Saify: analysis and writing.

Funding The authors declare that no funds, grants, or other support were received during the preparation of this manuscript.

Data availability- All data generated or analysed during the current study are presented in this article. However, the raw data will be also accessible from the corresponding author.

Code Availability Not applicable.

Declarations

Conflict of interest The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Ethical Approval We certify that this manuscript is original and has not been published and will not be submitted elsewhere for publication. This study follows all ethical practices during its writing.

Consent to Participate All authors duly participated.

Consent for Publication This is confirmed that the publication of this manuscript has been approved by all co-authors

Reference:

1. Adães J, Pires JCM (2019) Analysis and modelling of PM2.5 temporal and spatial behaviors in European cities. Sustain 11:. https://doi.org/10.3390/su11216019
2. Ahmed S, Hossain I (2008) Applicability of Air pollution Modeling in a Cluster of Brickfields in Bangladesh. Chem Eng Res Bull 12:28–34. https://doi.org/10.3329/cerb.v12i0.1495
3. Al Nayeem A, Sahadat Hossain M, Kamruzzaman Majumder A, Carter WS (2019) Spatiotemporal Variation of Brick Kilns and it's relation to Ground-level PM2.5 through MODIS Image at Dhaka District, Bangladesh. Int J Environ Pollut Environ Model 2:277–284
4. Andersen ZJ, Bønnelykke K, Hvidberg M, et al (2012) Long-term exposure to air pollution and asthma hospitalisations in older adults: A cohort study. Thorax 67:6–11. https://doi.org/10.1136/thoraxjnl-2011-200711
5. Anwar A, Ullah I, Younis M, Flahault A (2021) Impact of air pollution (PM2.5) on child mortality: Evidence from sixteen asian countries. Int J Environ Res Public Health 18:. https://doi.org/10.3390/ijerph18126375
6. Azad KMAK (2008) Risk Factors for Acute Respiratory Infections (ARI) Among Under-five Children in Bangladesh. J Sci Res 1:72–81. https://doi.org/10.3329/jsr.v1i1.1055
7. Badyda AJ, Grellier J, Dąbrowiecki P (2017) Ambient PM2.5 exposure and mortality due to lung cancer and cardiopulmonary diseases in polish cities. Adv Exp Med Biol 944:9–17. https://doi.org/10.1007/5584_2016_55
8. BBS (2020) Upazila specific population data. In: Bangladesh Bur. Stat. http://www.bbs.gov.bd/. Accessed 25 Jul 2020

9. Beelen R, Raaschou-Nielsen O, Stafoggia M, et al (2014) Effects of long-term exposure to air pollution on natural-cause mortality: An analysis of 22 European cohorts within the multicentre ESCAPE project. Lancet 383:785–795. https://doi.org/10.1016/S0140-6736(13)62158-3
10. Begum BA (2016) Dust Particle (PM10 and PM2.5) Monitoring for Air Quality Assessment in Naryanganj and Munshiganj , Bangladesh. Nucl Sci Appl 25:45–47
11. Begum BA, Hopke PK (2018) Ambient air quality in dhaka bangladesh over two decades: Impacts of policy on air quality. Aerosol Air Qual Res 18:1910–1920. https://doi.org/10.4209/aaqr.2017.11.0465
12. Begum BA, Hopke PK (2019) Identification of sources from chemical characterization of fine particulate matter and assessment of ambient air quality in Dhaka, Bangladesh. Aerosol Air Qual Res 19:118–128. https://doi.org/10.4209/aaqr.2017.12.0604
13. Begum BA, Tazmin A, Rabbani K, et al (2009) Investigation of Sources of Particulate Matter from the Tajgaon Industrial Area, Dhaka. J Bangladesh Acad Sci 33:71–85. https://doi.org/10.3329/jbas.v33i1.2952
14. Bishop CM (2017) Sparse Additive Gaussian Process with Soft Interactions. Open J Stat 7:
15. Cesaroni G, Forastiere F, Stafoggia M, et al (2014) Long term exposure to ambient air pollution and incidence of acute coronary events: Prospective cohort study and meta-analysis in 11 european cohorts from the escape project. BMJ 348:. https://doi.org/10.1136/bmj.f7412
16. Chen C, Ye W, Zuo Y, et al (2019a) Graph Networks as a Universal Machine Learning Framework for Molecules and Crystals. Chem Mater 31:3564–3572. https://doi.org/10.1021/acs.chemmater.9b01294
17. Chen G, Li S, Knibbs LD, et al (2018) A machine learning method to estimate PM2.5 concentrations across China with remote sensing, meteorological and land use information. Sci Total Environ 636:52–60. https://doi.org/10.1016/j.scitotenv.2018.04.251
18. Chen XC, Ward TJ, Cao JJ, et al (2019b) Source identification of personal exposure to fine particulate matter (PM2.5) among adult residents of Hong Kong. Atmos Environ 218:116999. https://doi.org/10.1016/j.atmosenv.2019.116999
19. Cho B, Song M (2017) Distributions and Origins of PM10 in Jeollabuk-do from 2010 to 2015. J Korean Soc Atmos Environ 33:251–264. https://doi.org/10.5572/kosae.2017.33.3.251
20. Chueinta W, Hopke PK, Paatero P (2000) Investigation of sources of atmospheric aerosol at urban and suburban residential areas in Thailand by positive matrix factorization. Atmos Environ 34:3319–3329. https://doi.org/10.1016/S1352-2310(99)00433-1
21. Dai H, Huang G, Zeng H, Yang F (2021) PM2.5 concentration prediction based on spatiotemporal feature selection using XGBoost-MSCNN-GA-LSTM. Sustain 13:. https://doi.org/10.3390/su132112071
22. Darrow LA, Klein M, Dana Flanders W, et al (2014) Air pollution and acute respiratory infections among children 0-4 years of age: An 18-year time-series study. Am J Epidemiol 180:968–977. https://doi.org/10.1093/aje/kwu234
23. Deters JK, Zalakeviciute R, Gonzalez M, Rybarczyk Y (2017) Modeling PM2.5 Urban Pollution Using Machine Learning and Selected Meteorological Parameters. 2017:
24. DoE (2022) 700 illegal brickfields to be demolished to check air pollution. In: Bangladesh Dep. Environ. https://www.thefinancialexpress.com.bd/national/700-illegal-brickfields-to-be-demolished-to-

check-air-pollution-1610427426
25. Dominici F, Daniels M, Zeger SL, Samet JM (2002) Air pollution and mortality: Estimating regional and national dose-response relationships. J Am Stat Assoc 97:100–111. https://doi.org/10.1198/016214502753479266
26. Doreswamy, Harishkumar KS, Km Y, Gad I (2020) Forecasting Air Pollution Particulate Matter (PM2.5) Using Machine Learning Regression Models. Procedia Comput Sci 171:2057–2066. https://doi.org/10.1016/j.procs.2020.04.221
27. Dyson LL (2015) A heavy rainfall sounding climatology over Gauteng, South Africa, using self-organising maps. Clim Dyn 45:3051–3065. https://doi.org/10.1007/s00382-015-2523-3
28. Eeftens M, Tsai MY, Ampe C, et al (2012) Spatial variation of PM2.5, PM10, PM2.5 absorbance and PMcoarse concentrations between and within 20 European study areas and the relationship with NO2 - Results of the ESCAPE project. Atmos Environ 62:303–317. https://doi.org/10.1016/j.atmosenv.2012.08.038
29. Egondi T, Ettarh R, Kyobutungi C, et al (2018) Exposure to outdoor particles (PM2.5) and associated child morbidity and mortality in socially deprived neighborhoods of Nairobi, Kenya. Atmosphere (Basel) 9:1–12. https://doi.org/10.3390/atmos9090351
30. Feng X, Li Q, Zhu Y, et al (2015) Artificial neural networks forecasting of PM2.5 pollution using air mass trajectory based geographic model and wavelet transformation. Atmos Environ 107:118–128. https://doi.org/10.1016/j.atmosenv.2015.02.030
31. Fotheringham AS, Charlton ME, Brunsdon C (2019) Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. Trans GIS 1:. https://doi.org/DOI: 10.1111/tgis.12580
32. Gaita SM, Boman J, Gatari MJ, et al (2014) Source apportionment and seasonal variation of PM2.5 in a sub-Saharan African city: Nairobi, Kenya. Atmos Chem Phys 14:9977–9991. https://doi.org/10.5194/acp-14-9977-2014
33. Godoy ARL, Silva AEA da, Bueno MC, et al (2021) Application of machine learning algorithms to PM2.5 concentration analysis in the state of São Paulo, Brazil. Brazilian J Environ Sci 56:152–165. https://doi.org/10.5327/z21769478782
34. Han S, Sun B (2019) Impact of population density on PM2.5 concentrations: A case study in Shanghai, China. Sustain 11:1–16. https://doi.org/10.3390/su11071968
35. Haque MI, Nahar K, Kabir MH, Salam A (2018) Particulate black carbon and gaseous emission from brick kilns in Greater Dhaka region, Bangladesh. Air Qual Atmos Heal 11:925–935. https://doi.org/10.1007/s11869-018-0596-y
36. Hasan M, Rahman S, Paul N, et al (2013) Analysis of Exhaust Emission of Vehicles in Dhaka City of Bangladesh. Glob J Sci Forntier Res 13:1–6
37. Hasan R, Hiya HJ, Marzia S (2020) Atmospheric Content of Particulate Matter PM2.5 in Gazipur and Mymensingh City Corporation Area of Bangladesh. Int J Res Environ Sci 6:. https://doi.org/10.20431/2454-9444.0602003
38. Hassan MM, Juhász L, Southworth J (2019) Mapping time-space brickfield development dynamics in Peri-Urban Area of Dhaka, Bangladesh. ISPRS Int J Geo-Information 8:.

https://doi.org/10.3390/ijgi8100447

39. Hassan MZ, Monjur MR, Biswas MAAJ, et al (2021) Antibiotic use for acute respiratory infections among under-5 children in Bangladesh: A population-based survey. BMJ Glob Heal 6:1–13. https://doi.org/10.1136/bmjgh-2020-004010
40. Hassan S (2022) Assessment of Health Impacts of Particulate Matter (PM2.5) on the Vulnerable Groups in the Central part of Bangladesh. 1–24
41. Hoek G, Krishnan RM, Beelen R, et al (2013) Long-term air pollution exposure and cardio-respiratory mortality: A review. Environ Heal A Glob Access Sci Source 12:. https://doi.org/10.1186/1476-069X-12-43
42. Hossain Bhuiyan MA, Chandra Karmaker S, Bodrud-Doza M, et al (2021) Enrichment, sources and ecological risk mapping of heavy metals in agricultural soils of dhaka district employing SOM, PMF and GIS methods. Chemosphere 263:128339. https://doi.org/10.1016/j.chemosphere.2020.128339
43. Howell CR, Su W, Nassel AF, et al (2020) Area based stratified random sampling using geospatial technology in a community-based survey. BMC Public Health 20:1–9. https://doi.org/10.1186/s12889-020-09793-0
44. Hu X, Belle JH, Meng X, et al (2017) Estimating PM2.5 Concentrations in the Conterminous United States Using the Random Forest Approach Department of Environmental Health , Rollins School of Public Health , Emory University , Department of Biostatistics & Bioinformatics , Rollins School of. Environ Sci Technol 1–29
45. Huang L, Zhang C, Bi J (2017) Development of land use regression models for PM2.5, SO2, NO2 and O3 in Nanjing, China. Environ Res 158:542–552. https://doi.org/10.1016/j.envres.2017.07.010
46. IQAIR (2022) Air quality in Bangladesh. https://www.iqair.com/bangladesh
47. Iqbal A, Afroze S, Rahman MM (2020) Vehicular PM emissions and urban public health sustainability: A probabilistic analysis for Dhaka City. Sustainability 12:1–18. https://doi.org/10.3390/SU12156284
48. Jiang N, Scorgie Y, Hart M, et al (2017) Visualising the relationships between synoptic circulation type and air quality in Sydney, a subtropical coastal-basin environment. Int J Climatol 37:1211–1228. https://doi.org/10.1002/joc.4770
49. Joharestani MZ, Cao C, Ni X, et al (2019) PM2.5 prediction based on random forest, XGBoost, and deep learning using multisource remote sensing data. Atmosphere (Basel) 10:12–18. https://doi.org/10.3390/atmos10070373
50. Katurji M, Noonan B, Zawar-Reza P, et al (2015) Characteristics of the springtime alpine valley atmospheric boundary layer using self-organizing maps. J Appl Meteorol Climatol 54:2077–2085. https://doi.org/10.1175/JAMC-D-14-0317.1
51. Khan MR, Huq S, Risha AN, Alam SS (2021) High-density population and displacement in Bangladesh. Science (80-.). 372:1–5
52. Kim S, Kim TY, Yi SM, Heo J (2018) Source apportionment of PM2.5 using positive matrix factorization (PMF) at a rural site in Korea. J Environ Manage 214:325–334. https://doi.org/10.1016/j.jenvman.2018.03.027
53. Kirwa K, Eckert CM, Vedal S, et al (2021) Ambient air pollution and risk of respiratory infection among adults: Evidence from the multiethnic study of atherosclerosis (MESA). BMJ Open Respir Res

8:1–10. https://doi.org/10.1136/bmjresp-2020-000866
54. Kohonen T (1982) Self-Organized Formation of Topologically Correct Feature Maps. Biol Cybern 43:59–69
55. Li G, Li L, Liu D, et al (2021) Effect of PM2.5 pollution on perinatal mortality in China. Sci Rep 11:1–12. https://doi.org/10.1038/s41598-021-87218-7
56. Liang CS, Yu TY, Chang YY, et al (2013) Source apportionment of PM2.5 particle composition and submicrometer size distribution during an Asian dust storm and non-dust storm in Taipei. Aerosol Air Qual Res 13:545–554. https://doi.org/10.4209/aaqr.2012.06.0161
57. Lin GY, Chen HW, Chen BJ, Yang YC (2022) Characterization of temporal PM2.5, nitrate, and sulfate using deep learning techniques. Atmos Pollut Res 13:101260. https://doi.org/10.1016/j.apr.2021.101260
58. Lin Y, Kuwayama DP (2016) Using satellite imagery and GPS technology to create random sampling frames in high risk environments. Int J Surg 32:123–128. https://doi.org/10.1016/j.ijsu.2016.06.044
59. Liu T, Hu B, Yang Y, et al (2020) Characteristics and source apportionment of PM2.5 on an island in Southeast China: Impact of sea-salt and monsoon. Atmos Res 235:104786. https://doi.org/10.1016/j.atmosres.2019.104786
60. Liu Y, Weisberg RH, Mooers CNK (2006) Performance evaluation of the self-organizing map for feature extraction. J Geophys Res Ocean 111:1–14. https://doi.org/10.1029/2005JC003117
61. M.M. H, M.A. R, B. M, et al (2018) Assessment of nitrogen oxides and sulphur dioxide content in the ambient air near the garments industries of Bangladesh. Environ Soc Sci 5:3–6
62. Ma B, Meng F, Yan G, et al (2020a) Diagnostic classification of cancers using extreme gradient boosting algorithm and multi-omics data. Comput Biol Med 121:103761. https://doi.org/10.1016/j.compbiomed.2020.103761
63. Ma J, Yu Z, Qu Y, et al (2020b) Application of the xgboost machine learning method in PM2.5 prediction: A case study of Shanghai. Aerosol Air Qual Res 20:128–138. https://doi.org/10.4209/aaqr.2019.08.0408
64. Maduekwe E, de Vries WT (2019) Random spatial and systematic random sampling approach to development survey data: Evidence from field application in Malawi. Sustainability 11:. https://doi.org/10.3390/SU11246899
65. Michel TM (1997) Does Machine Learning Really Work? AI Mag 18:71–83
66. Miller L, Xu X (2018) Ambient PM2.5 Human Health Effects—Findings in China and Research Directions. Atmosphere (Basel) 9:424. https://doi.org/10.3390/atmos9110424
67. Minh VTT, Tin TT, Hien TT (2021) PM2.5 Forecast System by Using Machine Learning and WRF Model, A Case Study: Ho Chi Minh City, Vietnam. Aerosol Air Qual Res 21:210108. https://doi.org/10.4209/aaqr.210108
68. Muindi K, Kimani-Murage E, Egondi T, et al (2016) Household air pollution: Sources and exposure levels to fine particulate matter in Nairobi slums. Toxics 4:12–14. https://doi.org/10.3390/toxics4030012
69. Mukta TA, Hoque MMM, Sarker ME, et al (2020) Seasonal variations of gaseous air pollutants (SO2, NO2, O3, CO) and particulates (PM2.5, PM10) in Gazipur: An industrial city in Bangladesh. Adv

Environ Technol 6:195–209. https://doi.org/10.22104/AET.2021.4890.1320
70. Nahar N, Mahiuddin S, Hossain Z (2021) The Severity of Environmental Pollution in the Developing Countries and Its Remedial Measures. Earth 2:124–139. https://doi.org/10.3390/earth2010008
71. Nakagawa K, Yu ZQ, Berndtsson R, Hosono T (2020) Temporal characteristics of groundwater chemistry affected by the 2016 Kumamoto earthquake using self-organizing maps. J Hydrol 582:124519. https://doi.org/10.1016/j.jhydrol.2019.124519
72. Nava S, Calzolai G, Chiari M, et al (2020) Source apportionment of PM2.5 in Florence (Italy) by PMF analysis of aerosol composition records. Atmosphere (Basel) 11:1–16. https://doi.org/10.3390/ATMOS11050484
73. Naz S, Page A, Agho KE (2015) Household air pollution and under-five mortality in Bangladesh (2004–2011). Int J Environ Res Public Health 12:12847–12862. https://doi.org/10.3390/ijerph121012847
74. Ni X, Cao C, Zhou Y, et al (2018) Spatio-temporal pattern estimation of PM2.5 in Beijing-Tianjin-Hebei Region based on MODIS AOD and meteorological data using the back propagation neural network. Atmosphere (Basel) 9:. https://doi.org/10.3390/atmos9030105
75. Nouri A, Ghanbarzadeh Lak M, Valizadeh M (2021) Prediction of PM2.5 Concentrations Using Principal Component Analysis and Artificial Neural Network Techniques: A Case Study: Urmia, Iran. Environ Eng Sci 38:89–98. https://doi.org/10.1089/ees.2020.0089
76. Paatero P (1997) Least square formulation of robust non-negetive factor analysis. Chemom Intell Lab Syst 37:23.35
77. Paatero P, Tapper U (1994) Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. Environmetrics 5:111–126. https://doi.org/10.1002/env.3170050203
78. Pan B (2018) Application of XGBoost algorithm in hourly PM2.5 concentration prediction. In: IOP Conference Series: Earth and Environmental Science
79. Pavel MRS, Zaman SU, Jeba F, et al (2021) Long-Term (2003–2019) Air Quality, Climate Variables, and Human Health Consequences in Dhaka, Bangladesh. Front Sustain Cities 3:. https://doi.org/10.3389/frsc.2021.681759
80. Pearce JL, Waller LA, Chang HH, et al (2014) Using self-organizing maps to develop ambient air quality classifications: a time series example. Environ Heal 13:. https://doi.org/doi:10.1186/1476-069X-13-56
81. Raaschou-Nielsen O, Andersen ZJ, Beelen R, et al (2013) Air pollution and lung cancer incidence in 17 European cohorts: Prospective analyses from the European Study of Cohorts for Air Pollution Effects (ESCAPE). Lancet Oncol 14:813–822. https://doi.org/10.1016/S1470-2045(13)70279-1
82. Rahman AM (2022) CO2 Emission from Brickfields in Bangladesh: Can Ethical Responsibility by Doing Reduce Level of Emission? Athens J Soc Sci 9:255–272. https://doi.org/10.30958/ajss.9-3-3
83. Rahman MM, Mahamud S, Thurston GD (2019) Recent spatial gradients and time trends in Dhaka, Bangladesh air pollution and their human health implications. J Air Waste Manage Assoc 69:478–501. https://doi.org/10.1080/10962247.2018.1548388
84. Rahman MM, Mazaheri M, Clifford S, Morawska L (2017) Estimate of main local sources to ambient

ultrafine particle number concentrations in an urban area. Atmos Res 194:178–189. https://doi.org/10.1016/j.atmosres.2017.04.036

85. Saha CK, Hosain J (2016) Impact of brick kilning industry in peri-urban Bangladesh. Int J Environ Stud 73:1–11. https://doi.org/10.1080/00207233.2016.1179014
86. Samek L, Stegowski Z, Furman L, et al (2017) Quantitative Assessment of PM2.5 Sources and Their Seasonal Variation in Krakow. Water Air Soil Pollut 228:. https://doi.org/10.1007/s11270-017-3483-5
87. Samet JMS, Ominici FD, Urriero FCC, et al (2000) Fine particulate air pollution and mortality in 20 U.S. cities, 1987–1994. N Engl J Med 343:
88. Sharma SK, Mandal TK, Jain S, et al (2016) Source Apportionment of PM2.5 in Delhi, India Using PMF Model. Bull Environ Contam Toxicol 97:286–293. https://doi.org/10.1007/s00128-016-1836-1
89. Srivastava D, Xu J, Vu T V., et al (2021) Insight into PM2.5sources by applying positive matrix factorization (PMF) at urban and rural sites of Beijing. Atmos Chem Phys 21:14703–14724. https://doi.org/10.5194/acp-21-14703-2021
90. Stauffer RM, Thompson AM, Young GS (2016) Tropospheric ozonesonde profiles at long-term U.S. monitoring sites: 1. A climatology based on self-organizing maps. J Geophys Res Atmos 121:. https://doi.org/10.1038/175238c0
91. Sultana M, Uddin MK (2019) Parental rejection, depression, and internet addiction among young adults. Pakistan J Psychol Res 34:419–432. https://doi.org/10.33824/PJPR.2019.34.2.23
92. Susanna NT, Alex M, Jose J, B S (2017) Application of self organising maps (SOM) in source characterization of PM10 and PM2.5 mass concentrations. In: International Conference on Recent Advances in Bioresource Technology (RABT 2017)
93. Tallon LA, Manjourides J, Pun VC, et al (2017) Cognitive impacts of ambient air pollution in the National Social Health and Aging Project (NSHAP) cohort. Environ Int 104:102–109. https://doi.org/10.1016/j.envint.2017.03.019
94. Tan J hua, Duan J chun, Chai F he, et al (2014) Source apportionment of size segregated fine/ultrafine particle by PMF in Beijing. Atmos Res 139:90–100. https://doi.org/10.1016/j.atmosres.2014.01.007
95. Tao S yang, Zhong B qing, Lin Y, et al (2017) Application of a self-organizing map and positive matrix factorization to investigate the spatial distributions and sources of polycyclic aromatic hydrocarbons in soils from Xiangfen County, northern China. Ecotoxicol Environ Saf 141:98–106. https://doi.org/10.1016/j.ecoenv.2017.03.017
96. Thurston GD, Ito K, Lall R (2011) A source apportionment of U.S. fine particulate matter air pollution. Atmos Environ 45:3924–3936. https://doi.org/10.1016/j.atmosenv.2011.04.070
97. Thygerson SM, Beard JD, House MJ, et al (2019) Air-quality assessment of on-site brick-kiln worker housing in bhaktapur, nepal: Chemical speciation of indoor and outdoor PM2.5 pollution. Int J Environ Res Public Health 16:. https://doi.org/10.3390/ijerph16214114
98. Tian YZ, Chen G, Wang HT, et al (2016) Source regional contributions to PM2.5 in a megacity in China using an advanced source regional apportionment method. Chemosphere 147:256–263. https://doi.org/10.1016/j.chemosphere.2015.12.132
99. Ulavi S, Shiva Nagendra SM (2019) Regression modelling for predicting visibility in a tropical urban environment. WIT Trans Ecol Environ 236:149–157. https://doi.org/10.2495/AIR190151

100. UNICEF (2022) Key demographic indicators in Bangladesh. In: Unicef. https://data.unicef.org/country/bgd/. Accessed 18 May 2022
101. USEPA (2014) EPA Positive Matrix Factorization (PMF) 5.0 Fundamentals and User Guide
102. Waked A, Favez O, Alleman LY, et al (2014) Source apportionment of PM10 in a north-western Europe regional urban background site (Lens, France) using positive matrix factorization and including primary biogenic emissions. Atmos Chem Phys 14:3325–3346. https://doi.org/10.5194/acp-14-3325-2014
103. Wang F, Chen T, Chang Q, et al (2021) Respiratory diseases are positively associated with PM2.5 concentrations in different areas of Taiwan. PLoS One 16:1–11. https://doi.org/10.1371/journal.pone.0249694
104. Wu J, Xie W, Li W, Li J (2015) Effects of urban landscape pattern on PM2.5 Pollution-A Beijing Case Study. PLoS One 10:1–20. https://doi.org/10.1371/journal.pone.0142449
105. Yazdi MD, Kuang Z, Dimakopoulou K, et al (2020) Predicting fine particulate matter (PM2.5) in the greater london area: An ensemble approach using machine learning methods. Remote Sens 12:. https://doi.org/10.3390/rs12060914
106. Yu X, Venecek M, Kumar A, et al (2019) Regional sources of airborne ultrafine particle number and mass concentrations in California. Atmos Chem Phys 19:14677–14702. https://doi.org/10.5194/acp-19-14677-2019
107. Zhang L, Wang S, Wang L, et al (2015) Updated emission inventories for speciated atmospheric mercury from anthropogenic sources in China. Environ Sci Technol 49:3185–3194. https://doi.org/10.1021/es504840m
108. Zhang R, Jing J, Tao J, et al (2013) Chemical characterization and source apportionment of PM2.5 in Beijing: Seasonal perspective. Atmos Chem Phys 13:7053–7074. https://doi.org/10.5194/acp-13-7053-2013
109. Zhao Z, Lao X, Gu H, et al (2021) How does air pollution affect urban settlement of the floating population in China? New evidence from a push-pull migration analysis. BMC Public Health 21:1–15. https://doi.org/10.1186/s12889-021-11711-x

Figures

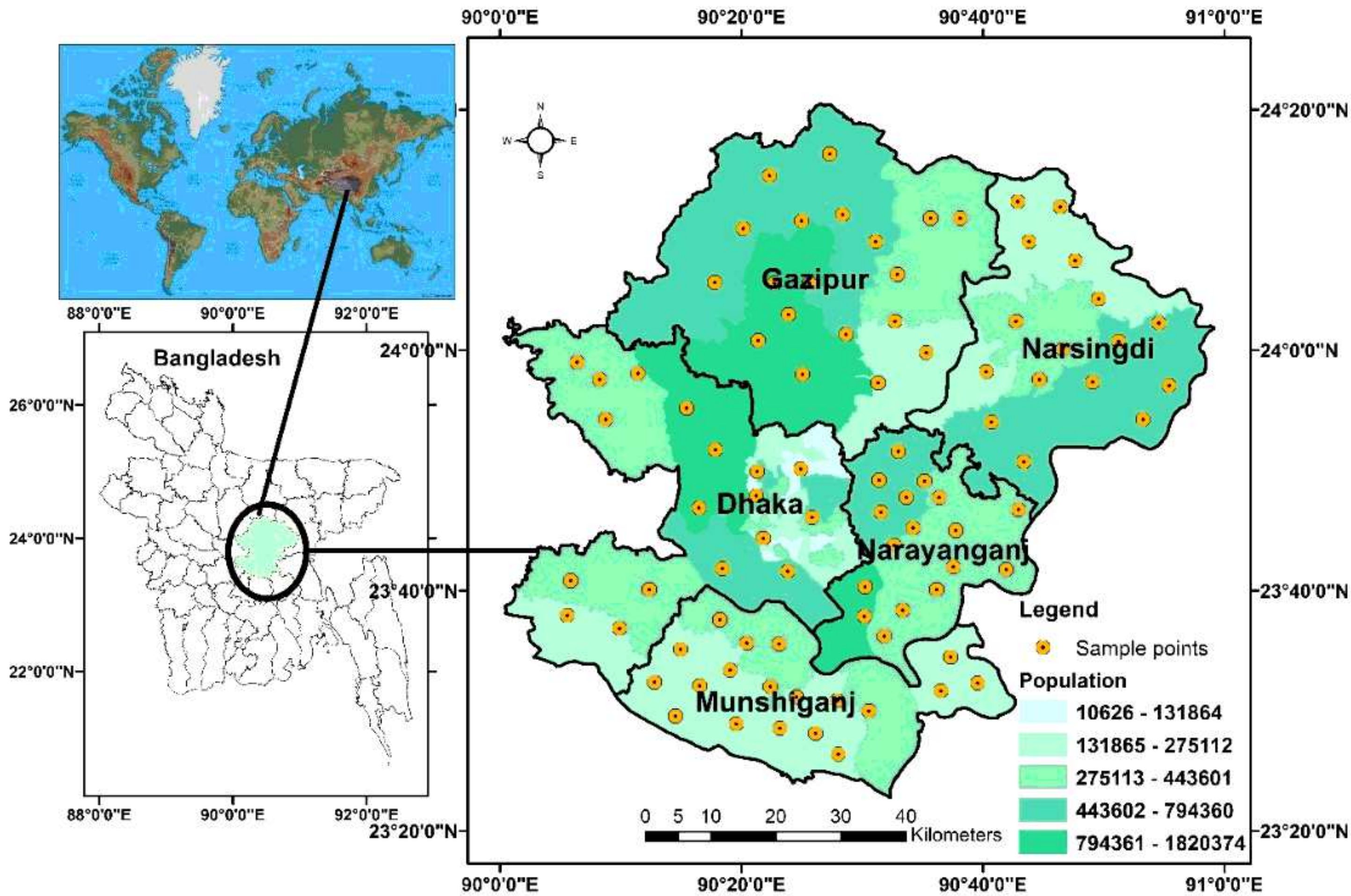

Figure 1

Location of the study with the distribution of total sample points and the total population of each district.

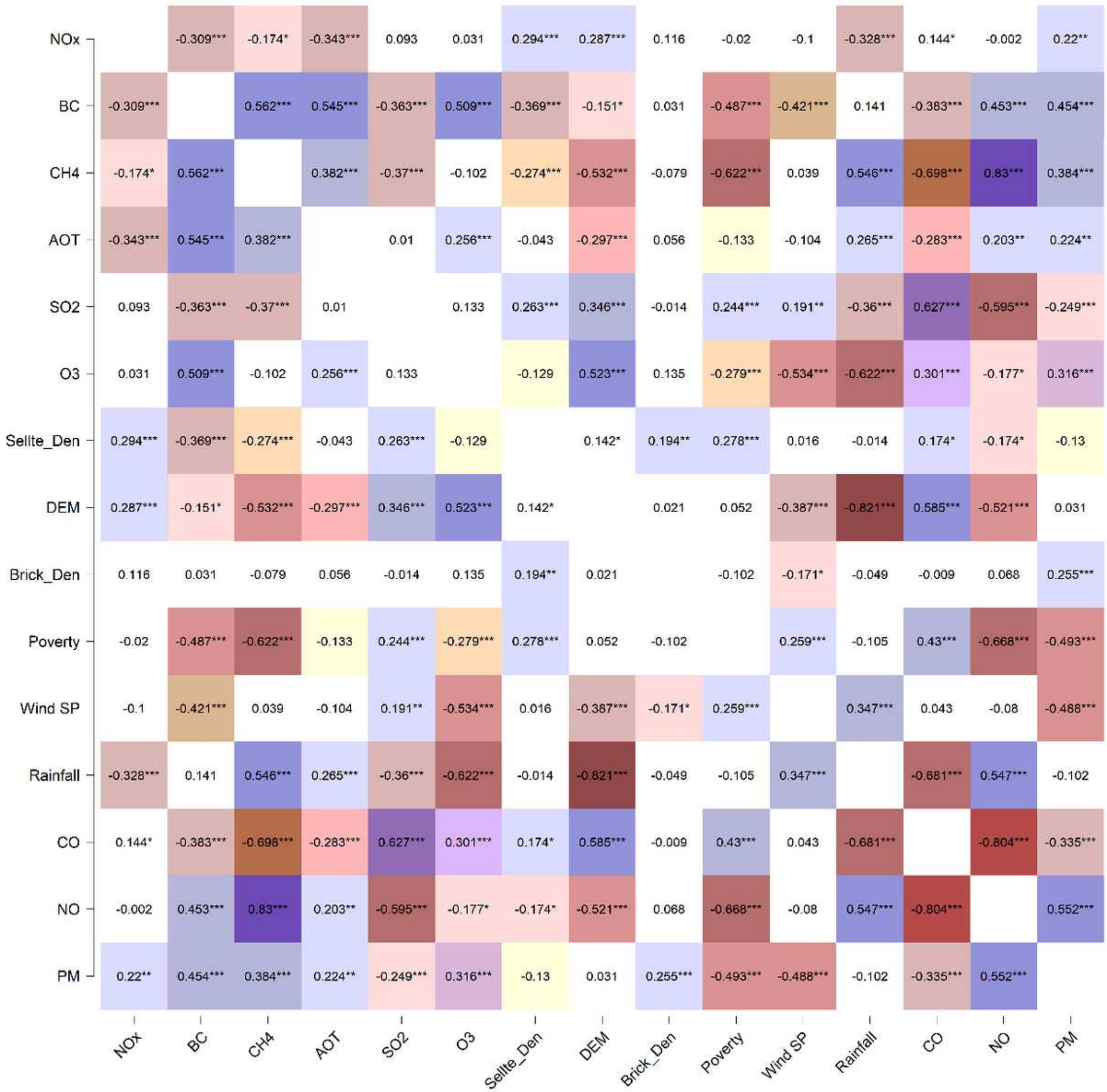

Figure 2

Results of Spearman's Correlation of all variables, where *p < .05, **p < .01, ***p < .001

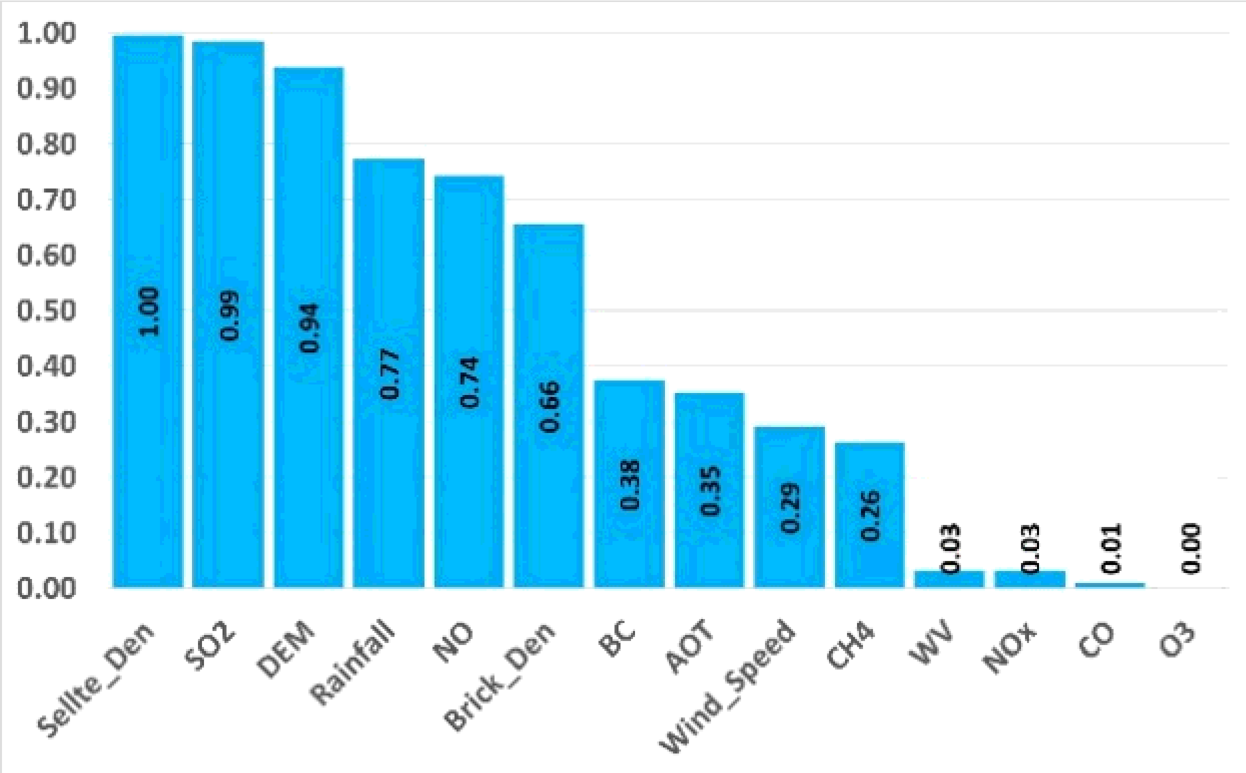

Figure 3

Relationship between the observed and predicted of each pollutant generated from the PMF modeling.

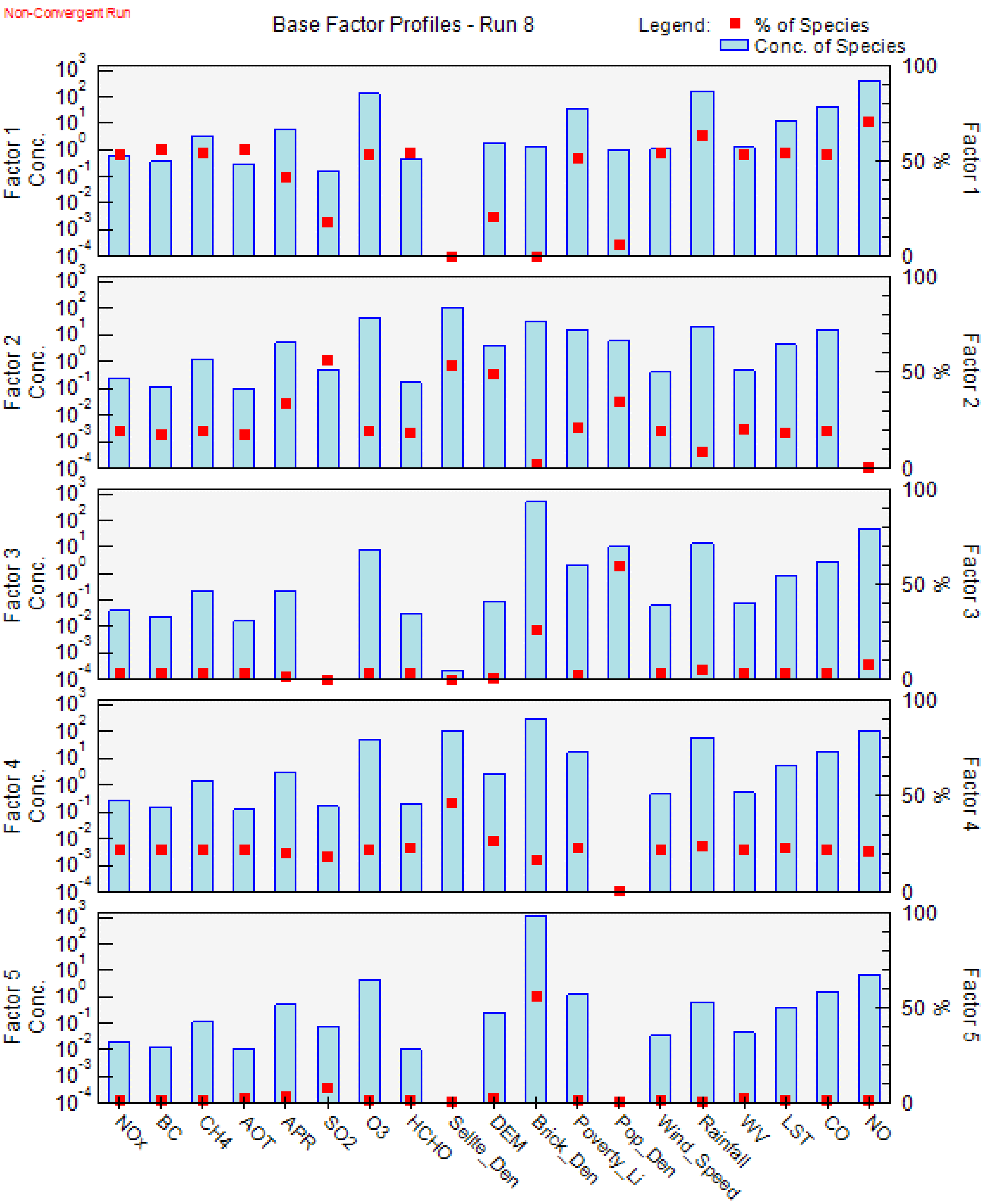

Figure 4

Factor profile and source contribution from the PMF modeling.

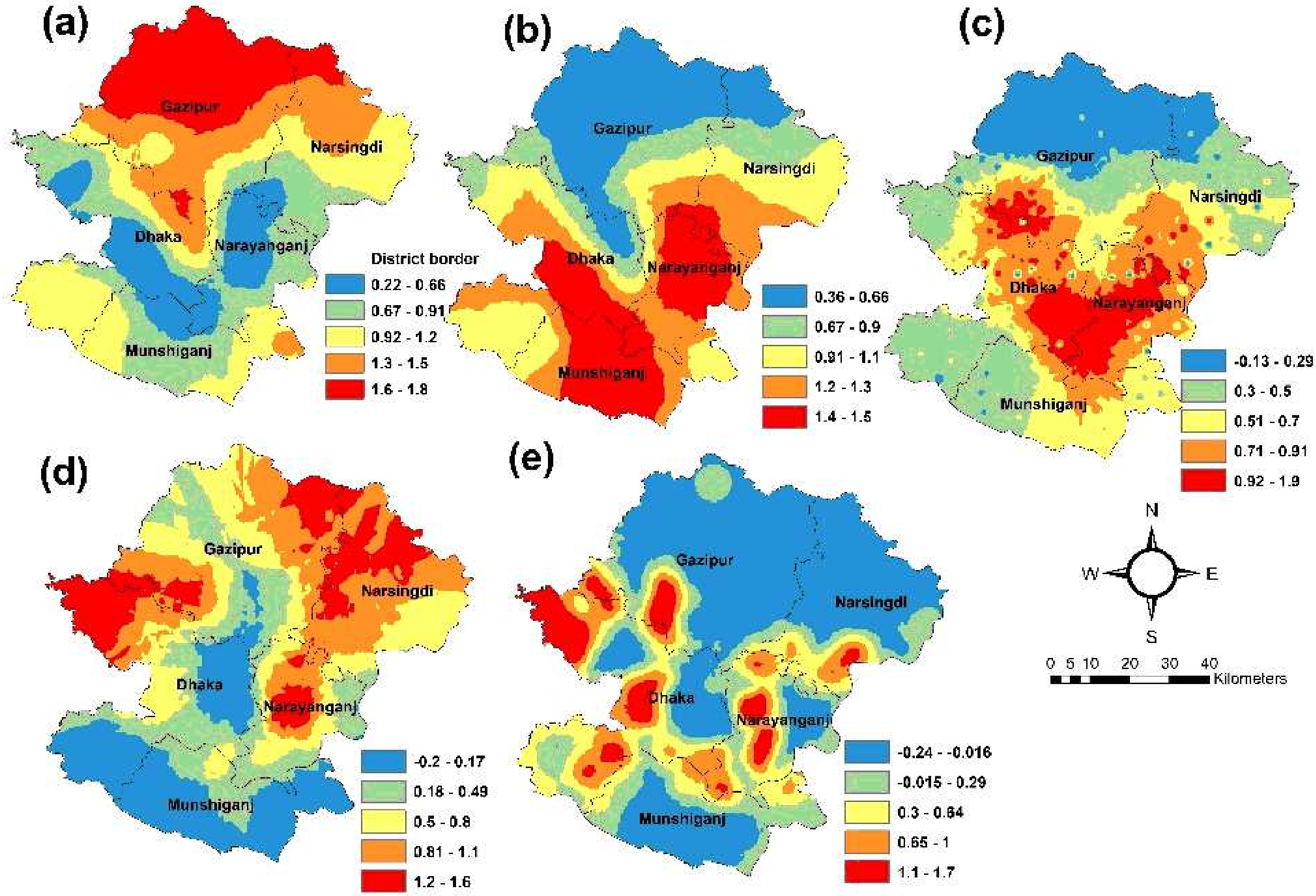

Figure 5

The spatial concentration mapping of all five factors from the PMF modeling, where factor 1 (5a), factor 2 (5b), factor 3 (5c), factor 4 (5d), and factor 5 (5e).

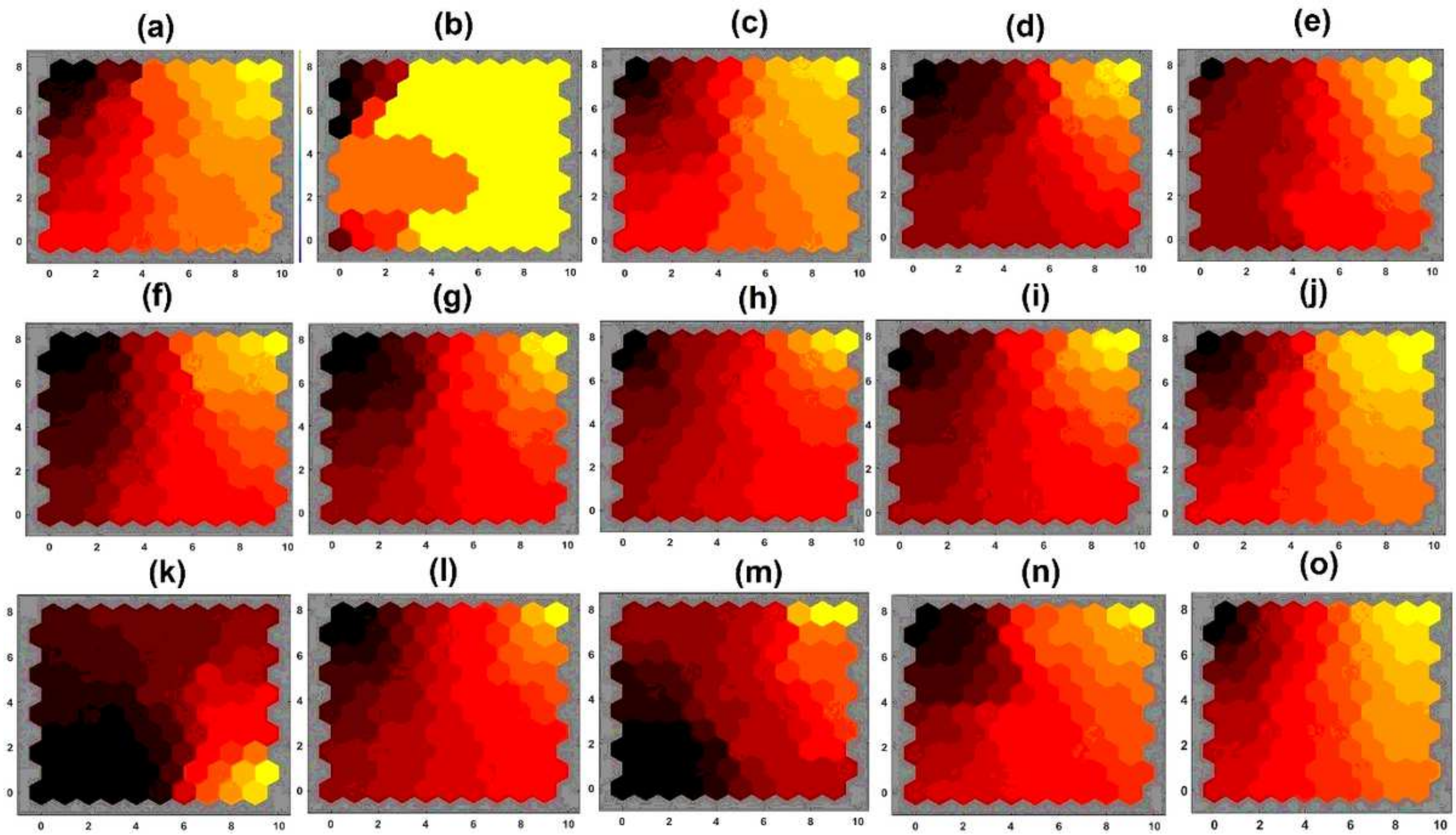

Figure 6

The spatial concentration of SOM map of all variables derived from the PMF modeling, tagging NOx (a), BC (b), CH_4 (c), AOT (d), CO (e), SO_2 (f), O_3 (g), Settlement density (h), DEM (i), NO (j), Brickfield (k), Poverty (l), Windspeed (m), Rainfall (n) and $PM_{2.5}$ (o).

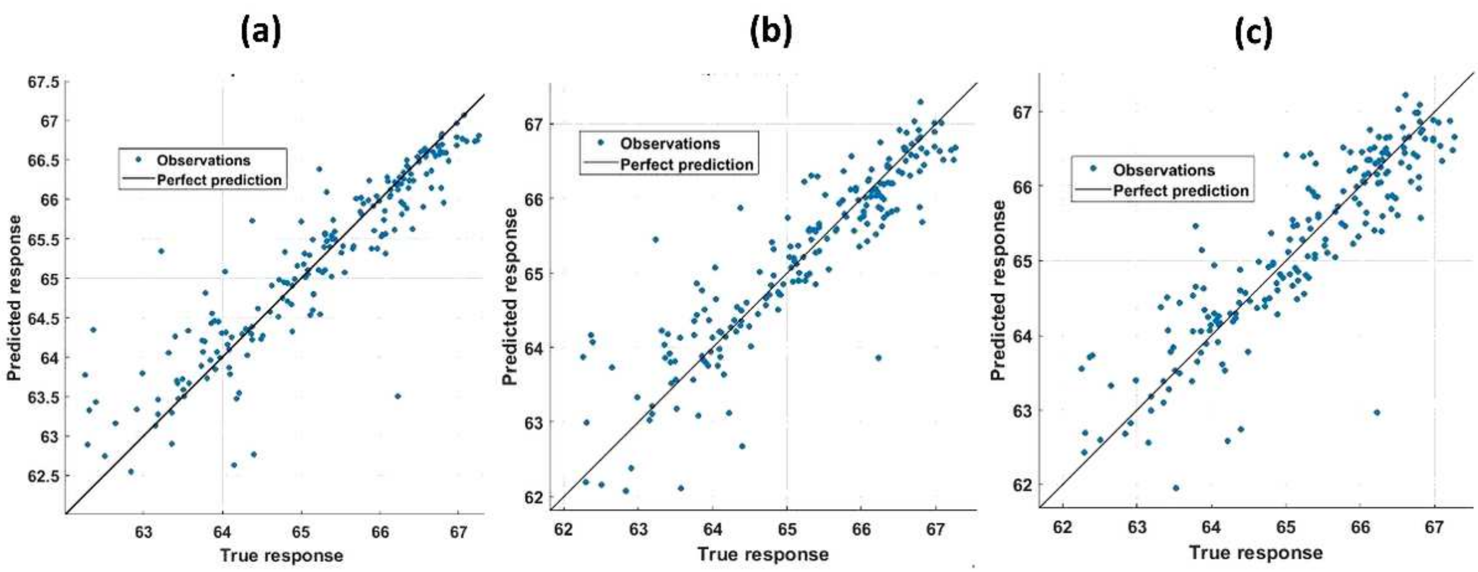

Figure 7

Scatterplot of predicted and observed values of $PM_{2.5}$. Figure 7a is the Random Forest, 7b is the Extreme gradient boosting while 7c is the Stepwise linear.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementarydata.docx