OPEN ACCESS
*Atmosphere*
ISSN 2073-4433
www.mdpi.com/journal/atmosphere

*Supplementary Information*

# $PM_{2.5}$ Prediction Based on Random Forest, XGBoost, and Deep Learning Using Multi-Source Remote Sensing Data

**Mehdi Zamani Joharestani [1,2,†], Chunxiang Cao [1,2,*], Xiliang Ni[1,2,†], Barjeece Bashir[1,2], and Somayeh Talebiesfandarani [1,2]**

[1] State Key Laboratory of Remote Sensing Science, Institute of Remote Sensing and Digital Earth, Chinese Academy of Sciences, Beijing 100101, China; madiz@radi.ac.cn (M.Z.); nixl@radi.ac.cn (N.X.); barjeece@radi.ac.cn(B.B.); soma@radi.ac.cn(S.T.)
[2] University of Chinese Academy of Science, Beijing 100049, China
* Correspondence: caocx@radi.ac.cn; Tel.: +86-139-1161-0226 (C.C.)
† These authors contributed equally to this work.

Received: date; Accepted: date; Published: date

## List of Abbreviations

| | |
|---|---|
| Lat. | Latitude |
| Lon. | Longitude |
| APM | Air Pollution Monitoring |
| PM | Particulate Matter |
| T | Temperature |
| T_min | Minimum Temperature |
| T_max | Maximum Temperature |
| Windsp | Wind Speed |
| ST_windsp | Sustained Wind Speed |
| *RH* | Relative Humidity |
| XGBoost | Extreme Gradient Boosting |
| *RF* | Random Forest |
| AOD | Aerosol Optical Depth |
| AOD10 | Aerosol Optical Depth at 10 km spatial Resolution |
| AOD03 | Aerosol Optical Depth at 03 km spatial Resolution |
| WHO | World Health Organization |
| ML | Machine Learning |
| RMSE | root mean square error |
| MAE | mean absolute error |
| DNN | Deep Neural Network |
| Rainfall_lag1 | Rainfall with the lag of one day |
| Rainfall_lag2 | Rainfall with the lag of two days |
| PM2.5_lag1 | $PM_{2.5}$ observations with the lag of one days |
| PM2.5_lag2 | $PM_{2.5}$ observations with the lag of two days |

## S1. Model assessment

The rows of the dataset were shuffled and split into train and test dataset considering 70% of dataset for training and 30% for test dataset. The same random state is set to guarantee model performance comparability. After training the model, the performance of the model was evaluated by indicators such as $R^2$ value, mean absolute error (MAE) and root mean square error (RMSE) shown in formulas (1 to 3). MAE is less sensitive to outliers in compare to RMSE.

$$R^2 = 1 - \frac{\sqrt{\sum_{i=1}^{n}(y_i - \check{y}_i)^2}}{\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{1}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |y_i - \check{y}_i| \tag{2}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n} (y_i - \check{y}_i)^2} \tag{3}$$

where $y_i$ is the observations of $PM_{2.5}$, $\check{y}_i$ is the predicted value, $\bar{y}$ is mean and n is total sample count.

## S2. Correlation coefficient analysis

The correlation coefficient analysis of features was carried out to evaluate features association with $PM_{2.5}$. There are several methods to calculate the correlation coefficient. The Pearson correlation coefficient and Spearman correlation coefficient that are the most commonly used methods to compute the rate of a possible association between two variable. The correlation coefficient varies from -1 to 1 illustrating perfect negative to perfect positive correlation. In the case of no correlation between a pair of data, the correlation coefficient is zero. Pearson correlation coefficient is calculated using the Equations 4 to 7.

$$\rho_{x,y} = \frac{Cov(x,y)}{\delta_x \delta_y} \tag{4}$$

$$Cov(x,y) = \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}) \tag{5}$$

$$\delta_x = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2} \tag{6}$$

$$\delta_y = \sqrt{\sum_{i=1}^{n} (y_i - \bar{y})^2} \tag{7}$$

where $Cov$ is the covariance between two variable x and y, $\delta_x$ and $\delta_y$ are the standard deviation of x and y subsequently. $x_i$ $and$ $y_i$ are individual sample points. $\bar{x}$ and $\bar{y}$ are the mean value of variables.

Despite the Pearson's correlation method that calculates the intensity of a linear association, Spearman's correlation coefficient focuses on the monotonic association of paired data by computing Pearson's correlation on the sorted scored data. It is less sensitive to outliers and not limited to the linear correlation of the data as it is in Pearson's method (see Equation 8)).

$$\rho_{rx,ry} = \frac{Cov(rx,ry)}{\delta_{rx} \delta_{ry}} \tag{8}$$

where $Cov$ is the covariance, $\delta_{rx}$ and $\delta_{ry}$ are the standard deviation of ranked values of x and y subsequently.

## S3. Tables S1-S2

**Table S1.** Descriptive Statistics of climatic parameters, $PM_{2.5}$, and AODs

| Parameter | Mean | STD | Min | 25% | 50% | 75% | Max | Count | Missing % |
|---|---|---|---|---|---|---|---|---|---|
| Temperature (°C) | 19.0 | 9.8 | 0.1 | 10.3 | 18.9 | 28.2 | 36.5 | 1462 | 0.75 |
| Temperature max (°C) | 23.9 | 10.4 | 0.5 | 14.4 | 24 | 33.6 | 42.2 | 1462 | 0.75 |
| Temperature min (°C) | 13.6 | 8.6 | 0 | 5.8 | 13.6 | 21.8 | 30.6 | 1462 | 0.75 |
| RH (%) | 32.0 | 17.1 | 8 | 18 | 29 | 42 | 93 | 1462 | 1.30 |
| Precipitation (mm) | 0.5 | 2.1 | 0 | 0 | 0 | 0 | 28.5 | 1462 | 0.75 |
| Visibility (km) | 8.9 | 1.6 | 0.8 | 8.4 | 9.7 | 10 | 11.3 | 1462 | **10.19** |
| Wind speed (m/s) | 11.3 | 4.9 | 2.8 | 8 | 10.2 | 13.5 | 35.6 | 1462 | 0.75 |
| Sustained wind speed (m/s) | 26.1 | 11.9 | 11.1 | 18.3 | 22.2 | 33.5 | 85.2 | 1462 | 0.82 |
| Air pressure (hPa) | 882 | 4.3 | 844 | 879 | 882 | 885 | 896 | 1462 | 0.00 |
| Dew point (°c) | 4.3 | 3.2 | 0 | 2 | 4 | 7 | 16 | 1462 | 0.00 |
| $PM_{2.5}$ ($\mu g\ m^{-3}$) | 86.8 | 33.5 | 3 | 62 | 82 | 107 | 500 | 29276 | **54.11** |
| AOD03 | 0.07 | 0.04 | 0.00 | 0.04 | 0.06 | 0.09 | 0.31 | 3621 | **94.09** |
| AOD10 | 0.04 | 0.03 | 0.00 | 0.02 | 0.03 | 0.06 | 0.22 | 22607 | **63.13** |

**Table S2.** Descriptive Statistics of $PM_{2.5}$ at APM stations of Tehran. The list is sorted based on the rate of missing values for $PM_{2.5}$ parameter.

| Station | Lon. (° E) | Lat. (° N) | Elev. (m) | Dis. (km) | Org. | Mean μg/m$^3$ | STD μg/m$^3$ | Max μg/m$^3$ | Min μg/m$^3$ | Missing % |
|---|---|---|---|---|---|---|---|---|---|---|
| Golbarg | 51.51 | 35.73 | 1297 | 15 | TM[1] | 69.0 | 25.4 | 179.0 | 11.0 | 8.48 |
| Setad | 51.43 | 35.73 | 1305 | 9 | TM | 89.5 | 29.1 | 189.0 | 24.0 | 16.83 |
| Shad abad | 51.30 | 35.67 | 1151 | 5 | TM | 99.0 | 29.1 | 195.0 | 26.0 | 18.40 |
| Aqdasiyeh | 51.48 | 35.80 | 1562 | 18 | TM | 72.5 | 26.3 | 162.0 | 18.0 | 19.63 |
| Tarbiat | 51.39 | 35.72 | 1267 | 5 | TM | 88.6 | 30.3 | 200.0 | 15.0 | 21.20 |
| Sharif | 51.35 | 35.70 | 1187 | 2 | TM | 102.3 | 27.8 | 197.0 | 35.0 | 22.71 |
| Punak | 51.33 | 35.76 | 1492 | 9 | TM | 66.8 | 23.2 | 182.0 | 17.0 | 25.38 |
| Ray | 51.43 | 35.60 | 1063 | 11 | TM | 97.9 | 32.4 | 196.0 | 30.0 | 29.48 |
| District 2 | 51.37 | 35.78 | 1624 | 11 | TM | 62.3 | 28.6 | 189.0 | 15.0 | 29.82 |
| Piroozi | 51.49 | 35.70 | 1225 | 13 | TM | 89.0 | 31.4 | 200.0 | 21.0 | 30.71 |
| District 21 | 51.24 | 35.70 | 1225 | 10 | TM | 92.2 | 29.8 | 196.0 | 23.0 | 32.76 |
| Tehran_01 | 51.43 | 35.59 | 1057 | 12 | DOE[2] | 116.6 | 34.8 | 297.0 | 24.0 | 33.93 |
| Tehran_07 | 51.48 | 35.64 | 1123 | 13 | DOE | 94.1 | 38.4 | 326.0 | 3.0 | 35.23 |
| Tehran_02 | 51.52 | 35.80 | 1688 | 20 | DOE | 84.7 | 32.8 | 184.0 | 18.0 | 35.70 |
| Tehran_06 | 51.51 | 35.74 | 1330 | 16 | DOE | 89.3 | 30.4 | 185.0 | 26.0 | 35.91 |
| Tehran_08 | 51.47 | 35.79 | 1511 | 17 | DOE | 87.7 | 29.5 | 198.0 | 27.0 | 36.32 |
| Tehran_04 | 51.40 | 35.80 | 1734 | 14 | DOE | 69.5 | 27.4 | 182.0 | 14.0 | 36.59 |
| Tehran_03 | 51.40 | 35.70 | 1212 | 5 | DOE | 85.7 | 31.6 | 197.0 | 27.0 | 36.66 |
| District 10 | 51.36 | 35.70 | 1187 | 11 | TM | 80.3 | 27.2 | 231.0 | 10.0 | 37.55 |
| Tehran_05 | 51.45 | 35.69 | 1172 | 9 | DOE | 91.8 | 34.3 | 321.0 | 19.0 | 38.03 |
| Rose Park | 51.27 | 35.74 | 1292 | 10 | TM | 77.2 | 30.0 | 173.0 | 18.0 | 41.38 |
| Tehran_09 | 51.39 | 35.67 | 1131 | 4 | DOE | 106.8 | 30.1 | 269.0 | 34.0 | 42.95 |
| Tehran_10 | 51.26 | 35.75 | 1329 | 11 | DOE | 84.2 | 34.1 | 204.0 | 7.0 | 43.30 |

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Masoudieh | 51.50 | 35.63 | 1172 | 15 | TM | 69.4 | 24.1 | 157.0 | 13.0 | 44.46 |
| District 11 | 51.39 | 35.67 | 1126 | 4 | TM | 96.1 | 32.6 | 275.0 | 14.0 | 47.61 |
| Tehran_14 | 51.36 | 35.64 | 1106 | 5 | DOE | 102.3 | 31.9 | 184.0 | 23.0 | 48.84 |
| Tehran_16 | 51.33 | 35.66 | 1131 | 3 | DOE | 117.1 | 32.6 | 251.0 | 33.0 | 49.79 |
| District 4 | 51.51 | 35.74 | 1354 | 16 | TM | 85.1 | 33.5 | 232.0 | 10.0 | 52.87 |
| Tehran_18 | 51.39 | 35.75 | 1408 | 8 | DOE | 78.9 | 32.6 | 390.0 | 20.0 | 57.11 |
| Tehran_11 | 51.36 | 35.74 | 1360 | 7 | DOE | 57.4 | 25.4 | 161.0 | 11.0 | 68.67 |
| Tehran_19 | 51.42 | 35.69 | 1163 | 7 | DOE | 88.9 | 48.6 | 500.0 | 22.0 | 74.69 |
| Darrous | 51.45 | 35.77 | 1404 | 14 | TM | 96.3 | 31.6 | 167.0 | 27.0 | 88.17 |
| Sadr | 51.43 | 35.78 | 1501 | 13 | TM | 89.3 | 32.3 | 174.0 | 29.0 | 89.19 |
| District 19 | 51.36 | 35.64 | 1103 | 5 | TM | 82.4 | 36.2 | 244.0 | 15.0 | 89.60 |
| Region 22 | 51.24 | 35.72 | 1258 | 11 | TM | 69.2 | 13.5 | 115.0 | 35.0 | 91.18 |
| District 16 | 51.40 | 35.64 | 1109 | 6 | TM | 83.6 | 33.0 | 214.0 | 23.0 | 93.09 |
| Tehran_13 | 51.24 | 35.56 | 1065 | 17 | DOE | 73.60 | 15.04 | 130.00 | 40.00 | 93.37 |
| **Tehran_15** | 51.64 | 35.82 | 1758 | 30 | DOE | - | - | - | - | **100.00** |
| **Fath** | 51.34 | 35.68 | 1159 | 1 | TM | - | - | - | - | **100.00** |
| **Mahallati** | 51.47 | 35.66 | 1148 | 11 | TM | - | - | - | - | **100.00** |
| **Tehran_12** | 51.58 | 35.72 | 1471 | 21 | DOE | - | - | - | - | **100.00** |
| **Tehran_17** | 51.40 | 35.54 | 1023 | 17 | DOE | - | - | - | - | **100.00** |

[1] TM stands for Municipality of Tehran city and [2] DOE stands for Department of Environment.

## S4. Figures S1

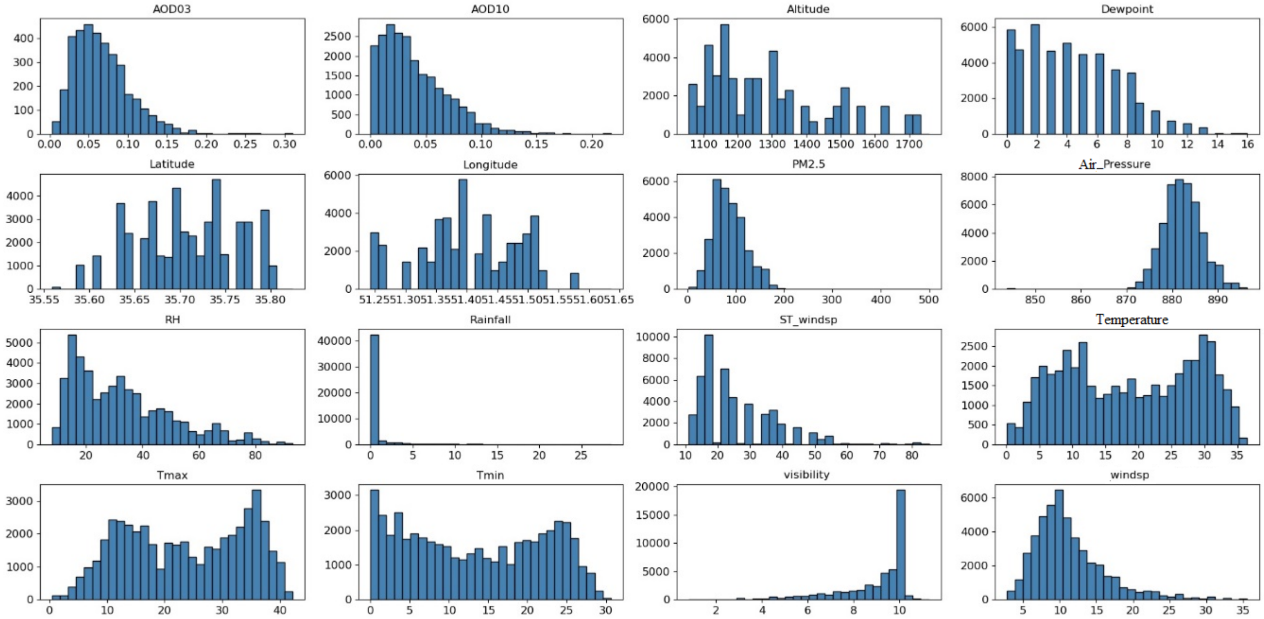

**Figure S1.** The histogram bar plot of features are illustrated in above figure. The histogram bar plot of features with the lag of one day and two days are not shown here, because their histograms are almost equal to the original features.