Data Article

Variability analysis in PM2.5 monitoring

Rashmi Bhardwaj*, Dimple Pruthi

Guru Gobind Singh Indraprastha University, India

ARTICLE INFO

Article history:
Received 20 September 2018
Received in revised form 11 February 2019
Accepted 13 February 2019
Available online 14 March 2019

ABSTRACT

United States Environmental Protection Agency (US EPA) and Central Pollution Control Board (CPCB) are two major air quality monitoring agencies in India that measure the concentration of particulate matter of size up to 2.5 μm (PM2.5). PM2.5 study over southern Asia has significance from the environment and ecosystem viewpoint (Abdullah et al., 2007; Dockery and Stone, 2007). In order to raise alert and controlling of pollutants, not only forecasting but the accuracy of forecasting has attracted attentions from various departments of research and air quality monitoring agencies. Quest for reducing error in forecasting has never come to pause. The precursor in forecasting is data monitoring. Keeping in focus the initial phase of data analysis, PM2.5 concentration was collected from both agencies within an area of radius 3.1 miles for the year 2016. Using the data, variability analysis is carried out for the efficiency of vital environment protection agencies.

© 2019 The Author(s). Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

1. Data

The daily concentration of fine particulate matter monitored by US EPA and CPCB from 1st January 2016 to 19th December 2016 has been taken for the present study. The data has been obtained from http://cpcb.nic.in/ and https://www.epa.gov/ for CPCB and US EPA respectively with time series in Fig. 1 and descriptive statistics shown in Table 1 [1,3].
1.1. Study area

The U.S Embassy and Consulates manage airborne fine particulate matter monitoring. PM2.5 is a standard recognized by US EPA and permit to examine against U.S. standard measures [5]. US EPA covers Chanakyapuri area in Delhi. Central Pollution Control Board (CPCB) of India is the apex organization in country for monitoring pollution [6]. One of its monitoring stations is at RK Puram, Delhi. RK Puram and Chanakyapuri are 3.1 miles away. New Delhi, the capital of India and has Latitude, longitude coordinates as 28.644800, 77.216721 respectively. Chanakyapuri in Delhi has Latitude, longitude coordinates as 28.593853, 77.188736 and RK Puram as 28.566008 and 77.176743 respectively.

**Fig. 1.** Daily variation of the US EPA and CPCB monitored PM2.5 values.
2. Experimental design, materials and methods

The study is divided into two sections of statistical and predictive analysis. Descriptive and inferential statistics are a vital part of data analysis. Analyzing data includes studying the statistics of data. The descriptive analysis describes big data using different measurements as indicated in Table 1.

A further step is to observe if there is any significant relation between the data considered. The correlation coefficient is a measure of the strength of the linear relationship between two such variables and is calculated as

$$r_{uv} = \frac{\sum u_iv_i - m_u m_v}{\sqrt{m \sum u_i^2 - (\sum u_i)^2} \sqrt{m \sum v_i^2 - (\sum v_i)^2}}$$

(1)

$r_{uv}$ lies between $-1$ and $+1$ inclusive as discussed in Bhardwaj and Pruthi, 2016 [2]. The value of Pearson correlation is 0.933 significant at 0.01 level proving the reliability of data observed.

Location and scale (estimated normal distribution parameters) of US EPA and CPCB data sets for unweighted cases using Blom's proportion estimation formula is calculated in Table 2. The probability-probability plot in Fig. 2 depicts deviation from the normal distribution. Location and scale values show a persistent trend for the data sets of PM2.5.

The extensively used t-test is carried out to analyze the difference between data monitored by USEPA and CPCB. Null hypothesis, $H_0$: No mean difference between USEPA and CPCB monitored data i.e. $\mu_{USEPA} - \mu_{CPCB} = 0$ and alternative hypothesis, $H_a$: $\mu_{USEPA} - \mu_{CPCB} \neq 0$.

$$t = \frac{\mu_{USEPA} - \mu_{CPCB}}{\sqrt{SD_{USEPA}^2 + SD_{CPCB}^2}}$$

(2)

Calculated t-value is compared to t-value corresponding to degree of freedom (see Table 3):

$$df = \frac{\left(\frac{SD_{USEPA}^2 + SD_{CPCB}^2}{N}\right)^2}{N-1 \left(\frac{SD_{USEPA}^2 + SD_{CPCB}^2}{N}\right)^2 + 2\frac{SD_{USEPA}SD_{CPCB}}{N^2}}$$

(3)

where, $\mu$ represents mean and SD standard deviation. The F-test statistic using one-way ANOVA is evaluated to emphasize the mean difference between two datasets (Table 4).

Table 1
Description of statistical parameters

|               | US EPA     | CPCB       |
|---------------|------------|------------|
| Mean          | 122.1859   | 135.9284   |
| Median        | 84.1667    | 112.4575   |
| Standard Deviation | 110.502    | 99.11627   |
| Kurtosis      | 9.121813   | 6.596559   |
| Coefficient of Variation | 9.04376e-1 | 7.2918e-1 |
| Skewness      | 2.461981   | 2.036923   |

Table 2
Estimated Distribution Parameters of PM2.5 (US EPA and CPCB)

| Distribution | Location   | US EPA     | CPCB       |
|--------------|------------|------------|------------|
| Normal       |            | 122.1859   | 135.9284   |
| Scale        |            | 110.50204  | 99.11627   |

The cases are unweighted.
\[ F = \frac{\text{regression mean square (MSR)}}{\text{mean square error (MSE)}} \]  

where, \( MSE \) is error sum of squares divided by df associated and \( MSR \) is regression sum of squares.

ANOVA and \( t \)-test sufficiently emphasized the significant difference between PM2.5 data monitored by two major agencies US EPA and CPCB.

2.1. Auto regressive integrated moving average

The preferred method in time series modeling is Box-Jenkins. Box-Jenkins method constitutes three components “AR,” “I,” or “MA” thus resulting in ARIMA. An ARIMA model can be expressed as

\[
(1 - \sum_{i} \theta_i L^i)(1 - L)^d \epsilon_t = \left( 1 + \sum_{i} \phi_i L^i \right) \epsilon_t
\]  

The above equations were fitted to PM2.5 data. An approach of identification, estimation and diagnostic is carried out for ARIMA modeling. It defines large-scale variation in behavior of stationary time series. ARIMA is build upon present and past values of response and residuals. The main steps of
ARIMA methods are: • Identification — examining the data along with calculation and drawing a graph of auto-correlation and partial auto-correlation functions. The smallest values of parameters are sought. When the value is 0, corresponding AR or MA component is not requisite in the respective model. 'I' component in ARIMA (trend) is examined first. The objective is to determine whether the process is stationary (d = 0) or not. If not than it has to be transformed into such. No connection between every two sequential observations implies p = 0. • Constructing models and estimating its parameters [7]. • Diagnostics and selection of model — the residuals and the quality of approximation of the model are examined. Theoretically, it is assumed that residuals are random and normally distributed. • Application of the predictive model, forecasts, analysis of dependencies, and study problem-solving capabilities [4].

Using ARIMA algorithm, PM2.5 is forecasted. Tables 5–7 summarizes the output of fitted ARIMA Model.

### Table 3
- **t-test for PM2.5 (US EPA and CPCB)**

| Sample | N  | Mean | Std. Deviation | Std. Error Mean | T   | DF | p-value |
|--------|----|------|----------------|-----------------|-----|----|---------|
| USEPA  | 357| 122.1859 | 110.502 | 5.84839 | 6.479 | 356 | 0.000   |
| CPCB   | 357| 135.9284 | 99.11627 | 5.24579 | | | |

### Table 4
- **One-way ANOVA for PM2.5 (US EPA and CPCB)**

| Source of Variation | Df | SS          | MS          | F            | p-value |
|---------------------|----|-------------|-------------|--------------|---------|
| Model               | 1  | 32382.37    | 32382.37    | 2.97299      | 0.00    |
| Error               | 710| 7733453.143 | 10892.19    |              |         |

### Table 5
- **ARIMA model**

| ARIMA Model | Coefficients |
|-------------|--------------|
| AR1 | MA1 | MA2 | MA3 |
| US EPA (1,0,3) | 0.982 | 0.142 | 0.236 | 0.22 |
| CPCB (1,0,2)  | 0.994 | 0.237 | 0.247 | |

### Table 6
- **Forecasted values**

| Date       | US EPA | CPCB |
|------------|--------|------|
|            | Observed | Forecasted | LCL | UCL | Observed | Forecasted | LCL | UCL |
| 20/12/2016 | 122.08  | 146.74 | 41.54 | 251.93 | 128.99 | 177.95 | 86.85 | 269.05 |
| 21/12/2016 | 163.34  | 161.8  | 24.43 | 299.17 | 171.13 | 192.67 | 78.39 | 306.96 |
| 22/12/2016 | 181.08  | 160.54 | 9.88  | 311.19 | 197.02 | 191.53 | 68.28 | 314.77 |

LCL and UCL stand for lower and upper confidence limits respectively.

### Table 7
- **R-squared values**

| R-squared | PM2.5 Forecasted |
|-----------|------------------|
|            | USEPA | CPCB |
| PM2.5 Observed | 0.9842 | 0.8520 |
| USEPA | 0.8656 | 0.9768 |
Fig. 3. Air Quality Index corresponding to PM2.5 concentration monitored by US EPA and CPCB.
To check the accuracy of modeling following statistics are used:

\[
R^2 = 1 - \frac{\text{sum of squared error of model}}{\text{sum of squared error of baseline model}}
\]  

(6)

2.2. Air quality index

Air Quality Index is not just a number signifying the quality of air but also explaining what we are inhaling. AQI was introduced in 1968. The objective was to aware public about deteriorating air quality and raise alarm in order to take precautionary measures. AQI is calculated using www.cpcb.nic.in and www.epa.gov. To focus on the effects of monitoring variability air quality index is calculated. AQI is divided into the following categories:

In Fig. 3 AQI is represented for those days in which they fall in different categories. The first and second bar in Fig. 3 represents calculated AQI corresponding to USEPA and CPCB data respectively. In a short span of 354 days AQI falls in different categories for 58 days.

2.3. Concluding remark

It is noted that PM2.5 monitored by US EPA and CPCB show a significant difference. Mathematically, US EPA measured PM2.5 data can be formed from CPCB by adding 13.7425 ± 7.85631729 and vice-versa by subtracting. AQI calculated fall in different categories as per National Standards. This might have lead to the issuance of wrong public health advisories in the past and if this difference is not observed as carried out in the present study, it may have a significant adverse impact on human health in future as well.

Acknowledgements

The authors are thankful to Guru Gobind Singh Indraprastha University, Delhi, India for providing research facilities and financial support.

Transparency document

Transparency document associated with this article can be found in the online version at https://doi.org/10.1016/j.dib.2019.103774.

References

[1] L.C. Abdullah, L.L. Wong, M. Saari, A. Salmiaton, M.S. Abdul Rashid, Particulate matter dispersion and haze occurrence potential studies at a local palm oil mill, Int. J. Environ. Sci. Technol. 4 (2) (2007) 271–278.
[2] R. Bhardwaj, D. Pruthi, Predictability and wavelet analysis of air pollutants for commercial and industrial regions in Delhi, Indian J. Ind. Appl. Math. 7 (2) (2016) 165–174.
[3] S. Tiwari, D.M. Chate, A.K. Srivastava, D.S. Bisht, B. Padmanabhamurty, Assessments of PM1, PM2.5 and PM10 concentrations in Delhi at different mean cycles, Geofizika 29 (2) (2012) 125–141.
[4] M.P. Stoimenova, Stochastic modeling of problematic air pollution with particulate matter in the city of pernik, Bulgaria, Ecol. Balk. 8 (2) (2016) 33–41.
[5] D.W. Dockery, P.H. Stone, Cardiovascular risks from fine particulate air pollution, N. Engl. J. Med. 356 (2007) 511–513.
[6] CPCB Central Pollution Control Board, Air quality in Delhi (1989–2000), National Ambient Air Quality Monitoring Series-NAAQMS/17/2000–2001, Parivesh Bhawan, Delhi, India, 2001.
[7] L. Zhang, J. Lin, R. Qiu, X. Hu, H. Zhang, Q. Chen, H. Tan, D. Lin, J. Wang, Trend analysis and forecast of PM2.5 in Fuzhou, China using the ARIMA model, Ecol. Indic. 95 (2018) 702–710.