Analyzing Air Pollutant Concentrations in New Delhi, India

B. Alparslan¹,† M. Jain²,³† J. Wu²,³,* S. Dev²,³,⁴

¹Middle East Technical University (METU), Ankara, Turkey
²ADAPT SFI Research Centre, Dublin, Ireland
³School of Computer Science, University College Dublin, Ireland.
⁴Beijing-Dublin International College, Beijing, China.

*Presenting author and corresponding author, e-mail: jiantao.wu@ucdconnect.ie
†Authors contributed equally
Table of contents

1 Dataset

2 Data Analysis
   Trend & Seasonality
   White Noise Time Series
   Stationarity Test

3 Conclusion
Introduction

- Air pollutants have long been known to cause major health problems across humans and all living organisms.
- Apart from that, they also play a crucial role in temperature inversion situations in the atmospheric layers thereby seriously impacting the radio communications, increased fog levels and decreased visibility [1].
- Consecutively, remote sensing analysts continuously monitor the amount of pollutants in the atmosphere - usually performed via satellite images [2, 3].
- Due to low temporal and low spatial resolution of satellite images, ground-based observations offer a useful alternative [4].
• Low-cost sensors are placed at various places across the globe to continuously monitor pollutant concentration levels in the atmosphere.

• The National Air Quality Index of India (NAQII)\(^1\) has records of ground-based observations of some pollutants: PM\(_{10}\), PM\(_{2.5}\), CO, Ozone, etc. from Delhi, India - which has consistently secured a position in the list of world’s top 10 most polluted cities over the years [5, 6].

• The dataset from this website is systematically scraped and archived in a user-friendly comma-separated-values (CSV) format.

\(^1\)https://app.cpcbccr.com/AQI_India/
Data is scraped for 2 years (2017 and 2018) from 2 stations based in Delhi, namely, the ‘Anand Vihar’ station and the ‘Punjabi Bagh’ station.

The readings were taken at a 15 minute interval for 7 different type of pollutants:

- Particulate Matter 10 ($PM^{10}$)
- Particulate Matter 2.5 ($PM^{2.5}$)
- Nitrogen Dioxide ($NO_2$)
- Ammonia ($NH_3$)
- Sulphur Dioxide ($SO_2$)
- Carbon Monoxide ($CO$)
- Ground-level Ozone ($O_3$)

The data is organized month-wise for each pollutant and each station.
Sample monthly readings of different pollutants as obtained from the dataset from 2 different stations. All 3 plots are shown along with their day-wise rolling mean and standard deviation.
Heat map showing the ratio of missing values to the total number of recordings for different pollutants in different months across the two stations of ‘Anand Vihar’ and ‘Punjabi Bagh’ for 2017 and 2018. Due to huge proportion of missing data, all pollutant-month combinations with more than 5% missing values were removed, leaving 58 such monthly entries.
While no continuous upward or downward trend was observed in the data (as depicted in Fig. 1), decomposition plots were made to study the seasonality. The degree of seasonality is noted to be very close to 96, i.e. data has seasonality.

Decomposition plots of pollutants from the ‘Anand Vihar’ station, captured in November, 2018.

(a) $PM_{10}$

(b) $PM_{2.5}$
White Noise Time Series

- Ljung-Box test [7] to check if data is a white noise
- DoF, $h = \min(2m, T/5)$; where $m$ = period of seasonality, and $T$ = length of time series [8]
- Adjoining table shows the test results for one of the monthly entries that are used in this paper
- Since $Q > c$-value in each lag, the time series is not a white noise
- Similar results were obtained for other monthly entries

| Lag | p-value | $Q$    | c-value |
|-----|---------|--------|---------|
| 1   | 0.000   | 2663.383 | 2.706   |
| 2   | 0.000   | 5075.472 | 4.605   |
| 3   | 0.000   | 7267.983 | 6.251   |
| 4   | 0.000   | 9201.765 | 7.779   |
| 5   | 0.000   | 10887.555 | 9.236 |
| 6   | 0.000   | 12321.927 | 10.645 |
| 7   | 0.000   | 13528.789 | 12.017 |
| 8   | 0.000   | 14524.748 | 13.362 |
| 9   | 0.000   | 15316.793 | 14.684 |
Stationarity Test - Augmented Dickey-Fuller (ADF) Test

- ADF test [9, 10] determines how strongly a time series is defined by a trend.
- If p-value > 0.05, data is stationary and vice versa.
- Adjoining table shows the test results for one of the monthly entries that are used in this paper.
- Since p-value > 0.05, and value of Test Statistic < Critical Value (1%), time series is stationary with a significance level of less than 1%.

| Test Statistic      | -6.448849          |
|---------------------|--------------------|
| p-value             | 1.541445 x 10^{-08}|
| # Lags Used         | 28                 |
| # Observations Used | 2846               |
| Critical Value (1%) | -3.43265           |
| Critical Value (5%) | -2.862556          |
| Critical Value (10%)| -2.567311          |
Stationarity Test - Kwiatkowski-Phillips-Schmidt-Shin (KPSS) Test

- Similar to ADF test but with a different approach
- If p-value > 0.05, data is stationary and vice versa
- Adjoining table shows the test results for the same monthly entry that was used for ADF test
- Since p-value < 0.05, and value of Test Statistic < Critical Value (1%), time series is NOT stationary with a significance level of less than 1%

| Test Statistic | 0.534381 |
|----------------|----------|
| p-value        | 0.033923 |
| Number of Lags Used | 28 |
| Critical Value (10%) | 0.347 |
| Critical Value (5%)  | 0.463 |
| Critical Value (2.5%) | 0.574 |
| Critical Value (1%)   | 0.739 |
Eliminating Stationarity

- Having ADF test result of stationary and KPSS test result of non-stationary means this data is difference stationary
- Differencing is applied: $z_t = y_t - y_{t-1}$
- Post applying time differencing, both ADF and KPSS test indicate that the dataset is now completely stationary (see adjoining table)

|                          | ADF Test | KPSS Test |
|--------------------------|----------|-----------|
| Test Statistic           | $-16.203$ | $0.006$ |
| p-value                  | $4.1 \times 10^{-29}$ | $0.1$ |
| # Lags Used              | $28$     | $28$     |
| # Observations Used      | $2845$   | $-$      |
| Critical Value (10%)     | $-2.57$  | $0.35$   |
| Critical Value (5%)      | $-2.86$  | $0.46$   |
| Critical Value (2.5%)    | $-3.43$  | $0.57$   |
| Critical Value (1%)      | $-3.43$  | $0.74$   |
Conclusion & Future Work

- A unique dataset of 7 different pollutant values, gathered from 2 different stations in New Delhi over a period of 2 years (sourced from NQAI website) is presented.
- Pre-processed to remove missing data and segregate dataset as month-station-pollutant sets.
- Statistical tests to comment on the trend, seasonality, white noise similarity, and stationarity of the dataset.
- Provides assistance to other researchers who wish to use the dataset for further studies like pollutant forecasting and/or correlation analysis.
- In future, the authors would like to extend the dataset by collecting data for more years and more number of stations across the New Delhi area.
- Furthermore, the plan is to provide a completely pre-processed and cleared version of such datasets in the public domain so that the researchers in the community can make quick and easy use of the data.
N. David and H. O. Gao, “Using cellular communication networks to detect air pollution,” *Environmental science & technology*, vol. 50, no. 17, pp. 9442–9451, 2016.

D. Kaloni, Y. H. Lee, and S. Dev, “Impact of COVID19-induced lockdown on air quality in Ireland,” in *Proc. IEEE International Geoscience and Remote Sensing Symposium (IGARSS)*. IEEE, 2021.

S. Manandhar, S. Dev, Y. H. Lee, and Y. S. Meng, “Correlating satellite cloud cover with sky cameras,” in *Proc. Progress in Electromagnetics Research Symposium-Fall (PIERS-FALL)*. IEEE, 2017, pp. 2166–2168.

N. Danesi, M. Jain, Y. H. Lee, and S. Dev, “Monitoring atmospheric pollutants from ground-based observations,” in *Proc. USNC-URSI Radio Science Meeting (Joint with AP-S Symposium)*. IEEE, 2021.

A. Lal, “Environmental status of Delhi,” *Indian Journal of Environmental Protection*, vol. 16, no. 1, pp. 1–11, 1996.

Earth.org, “15 Most Polluted Cities in the World,” https://earth.org/most-polluted-cities-in-the-world/, January 2021, [Online; accessed 20-June-2021].

G. M. Ljung and G. E. Box, “On a measure of lack of fit in time series models,” *Biometrika*, vol. 65, no. 2, pp. 297–303, 1978.

R. J. Hyndman and G. Athanasopoulos, *Forecasting: principles and practice*. OTexts, 2018.
D. A. Dickey and W. A. Fuller, “Distribution of the estimators for autoregressive time series with a unit root,” *Journal of the American statistical association*, vol. 74, no. 366a, pp. 427–431, 1979.

Y.-W. Cheung and K. S. Lai, “Lag order and critical values of the augmented dickey–fuller test,” *Journal of Business & Economic Statistics*, vol. 13, no. 3, pp. 277–280, 1995.