The Journal of Open Source Software

pollucheck v1.0: A package to explore open-source air pollution data

Adithi R. Upadhya[1], **Pratyush Agrawal**[1], **Sreekanth Vakacherla**[2], **and Meenakshi Kushwaha**[1]

1 ILK Labs, Bengaluru, India **2** Center for Study of Science, Technology and Policy, Bengaluru, India

DOI: 10.21105/joss.03435

Software

- Review
- Repository
- Archive

Editor: Andrew Stewart

Reviewers:

- @nmstreethran
- @ibarraespinosa

Submitted: 18 June 2021
Published: 23 July 2021

Summary

Air pollution impacts human health, quality of living, climate, and the economy (Hystad et al., 2020). To assess its impact and facilitate mitigation actions, quantification of air pollution is vital. Measurements are the most accurate way of quantifying air pollution. Many countries conduct regulatory measurements of various air pollutants (e.g., fine and respirable particulate matter, nitrogen dioxide, sulfur dioxide, and surface ozone) and make the data available publicly.

Air pollution data sets typically span several seasons or years and real-time data are recorded typically every hour or at a higher frequency. With the ever increasing amount of data and number of data providers, there is a clear need for tools to handle, analyze, and visualize large data sets. The current Shiny app `pollucheck` aims at a simple workflow to generate a suite of statistical plots and summary statistics (Chang et al., 2021). Users do not need any programming background to analyze time series data and generate a variety of plots.

`pollucheck` can handle real-time pollution and co-located meteorological data (if available) from the three most popular open-source air pollution databases: OpenAQ, AirNow, and Indian Central Pollution Control Board (CPCB) dashboard. While CPCB data are specific to Indian regulatory monitoring stations, OpenAQ hosts the global open-source pollution databases and AirNow hosts the global $PM_{2.5}$ (mass concentration of particulate matter with an aerodynamic diameter less than or equal to 2.5 microns) data, collected under the United States Embassy and Consulates' air quality monitoring programmes.

The output of `pollucheck` is displayed in seven tabs. Different packages used for building `pollucheck` include `tidyverse`, `openair`, `shiny`, `bslib`, `forecast`, `biwavelet`, `readxl`, `DT`, `data.table`, `nortest`, and `zoo`.

Statement of Need

Pollution data from these sources are typically in different file formats and templates that require customised codes or programmes for analysis. Also, a rigorous quality check of the data is preferred before visualization (plotting) and reporting. `pollucheck` offers a single-stop solution for

(i) handling the pollution data from the open-source databases,
(ii) applying a suite of quality check options,
(iii) generating a variety of summary statistics at various averaging intervals,
(iv) performing time series analysis,

(v) generating a bunch of temporal and statistical plots, and
(vi) comparing data from two input files.

To our knowledge, currently there is no application that can generate utilisable summary statistics and plots using the data from the pollution databases. However, there are a few Shiny apps that deal with data cleaning and visualization of pollution data collected from single/multiple air quality instruments (Salmon et al., 2017; Upadhya et al., 2020).

App Display

i) The `File` tab is used to upload the input file and to specify the source and time resolution of the input data. The default time zone is set to *Asia/Kolkata*. For OpenAQ and AirNow data sets, appropriate time zones need to be selected based on the input file. For the CPCB data set, the time zone option is default and inactive. A set of quality check options for the (a) removal of negative values, (b) removal of consecutive duplicate values, and (c) detection of outliers are provided. Data completeness criteria (minimum percentage of data required) for computing daily mean values can be specified. If the input file contains simultaneous $PM_{2.5}$ and PM_{10} (mass concentration of particulate matter with an aerodynamic diameter less than or equal to 10 microns) data, the app computes the $PM_{2.5}/PM_{10}$ ratio, a useful metric in the air pollution field to identify sources of PM and to estimate $PM_{2.5}$ when only PM_{10} is available (Chan & Yao, 2008; Chu et al., 2015; Spandana et al., 2021). The selected quality check or completeness criteria will be applied to all the parameters of the input file. Hourly or daily mean values of all the parameters can be displayed and downloaded (as `.csv`) from this tab.

ii) The `Summary` tab provides various statistics (central tendencies, percentiles, minimum, maximum, standard deviation, interquartile range, etc.) for all the parameters in the input file at three different averaging intervals. The averaging intervals can be selected using the drop-down menu. The displayed statistics can be downloaded.

iii) The `Summary Plots` tab generates (a) a data availability plot for all the parameters (based on daily mean values), (b) a time series plot, (c) box and whisker plots, (d) a vertical bar plot, and (e) diurnal variability plots. Except for the data availability plot, the parameter of interest to a plot needs to be selected from the drop-down menu. Plots can be generated using hourly or daily mean data. The diurnal variability plots can be plotted either by aggregating the whole data in the input file or month wise. Considering the general log-normal nature of the pollution data, an option is provided for the diurnal variability plots to be plotted using mean and standard deviations or median and interquartile ranges. The title and y-axis labels of the plots are editable.

iv) The `Statistical Plots` tab can be used to conduct normality tests (Anderson-Darling and Shapiro-Wilk), generate density and quantile-quantile (QQ) plots, generate auto-correlogram, and conduct trends and periodicity analysis on the parameter selected. While autocorrelogram is generated based on monthly mean values, trend (the Mann-Kendall test) and periodicity (wavelet analysis) analyses are conducted on daily mean values of the selected parameter. For trend and periodicity analyses and generating autocorrelogram, the missing daily mean values are imputed using the `forecast` package (Hyndman & Khandakar, 2008).

v) The `Linear Regression` tab allows a user to perform univariable and multiple linear regression analyses among the parameters of choice. For univariable linear regression, a scatter plot will be generated with least-squares linear fit. For multiple linear regression, multiple independent parameters can be selected. A scatter plot between the dependent variable and fitted data (using regression coefficients) will be generated. Relevant statistical coefficients are provided along with the plots.

vi) The `Compare` tab allows users to upload a second data file to compare data between the selected parameters from the two input files. The selected quality check criteria conditions applied on the parameters of the first input file will be automatically applied to the parameters in the second input file. Time-series, scatter, and diurnal variability plots of the two parameters of interest will be generated.

vii) Some features of the widely used `openair` package (Carslaw & Ropkins, 2012) are integrated into `pollucheck` with permission. Calendar and time variation plots of the selected parameter are generated in this tab. Daily data will be used for Calendar plots and hourly data will be used for time variation plots.

An extensive list of frequently asked questions (FAQs) is provided as a separate tab for a better understanding of the `pollucheck` functioning, detailed features of the plots, and analysis and the various packages used to build `pollucheck`.

Limitations

1) `pollucheck` does not download data automatically from the cloud. Downloaded files need to be provided as input.
2) Multiple files cannot be uploaded to `pollucheck` at a given time.
3) The current version of `pollucheck` is limited to accepting real-time data files from only three data sources.
4) Some analyses (e.g., periodicity analysis) can be performed using daily mean values only.
5) Caution needs to be exercised when using the averaged wind direction data, since wind direction is a vector quantity; hence it needs to be processed in a different way which has not been implemented here. Only wind direction data at one hour resolution is processed correctly.
6) Any manipulation or alteration to the downloaded file before giving it as input to the app can lead to erroneous results.

Installation

`pollucheck` is hosted online on *shinyapps.io* and can be installed to serve locally from GitHub.

Load and run `pollucheck` as follows:

```
install.packages("devtools")
devtools::install_github("adithirgis/pollucheck")
pollucheck::pollucheck_run()
```

`pollucheck` is furnished with a preloaded data set for a quick user tour of the analysis, plotting options, and the functions available. In the `Compare` tab, the preloaded data set acts as the second input file if no second file is uploaded.

Case Study

For better understanding of the major functionalities of `pollucheck`, we present a case study based on 18 months of pollution data set. This data is downloaded from the Central Pollution Control Board dashboard for the monitoring station located at Hebbal, Bengaluru, India at a time resolution of 60 minutes. Only plots related to $PM_{2.5}$ data generated through the app

are shown here. Figure 1 depicts the efficiency of the app in detecting the outliers. The top panel of Figure 1 shows the hourly time series of the raw $PM_{2.5}$ (few outliers were synthetically added to the data), while the bottom panel depicts the quality checked data. Almost all the sporadically high values were detected by the app as outliers and removed.

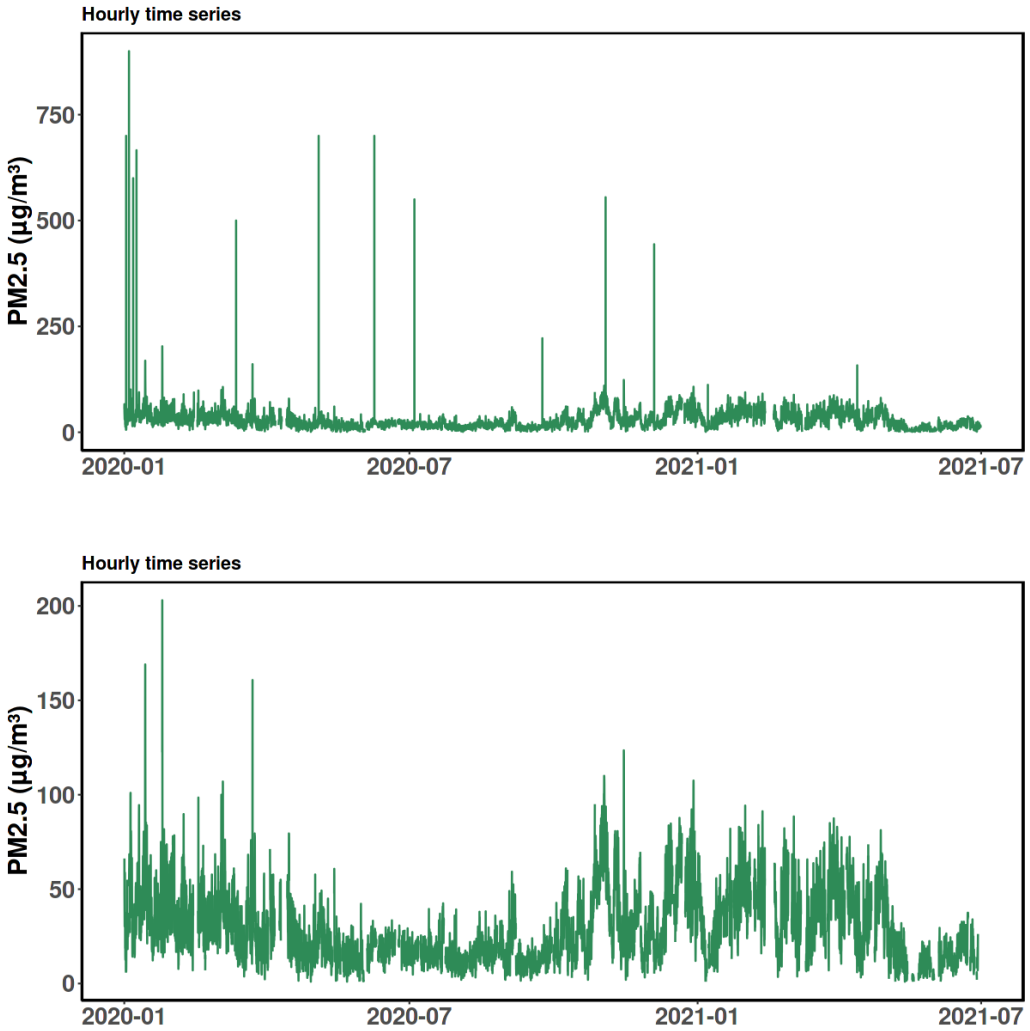

Figure 1: Hourly time series of raw (top panel) and cleaned (bottom panel) $PM_{2.5}$.

Figure 2 depicts the difference between **Month and year box plot** and **Monthly box plot**. These plots are highly useful if the dataset length is more than a year. **Monthly box plot** (bottom panel) partitions all the data points into the calendar month bins irrespective of the year. While **Month and year box plot** (top panel) accounts for the entire timeline, i.e., including the year.

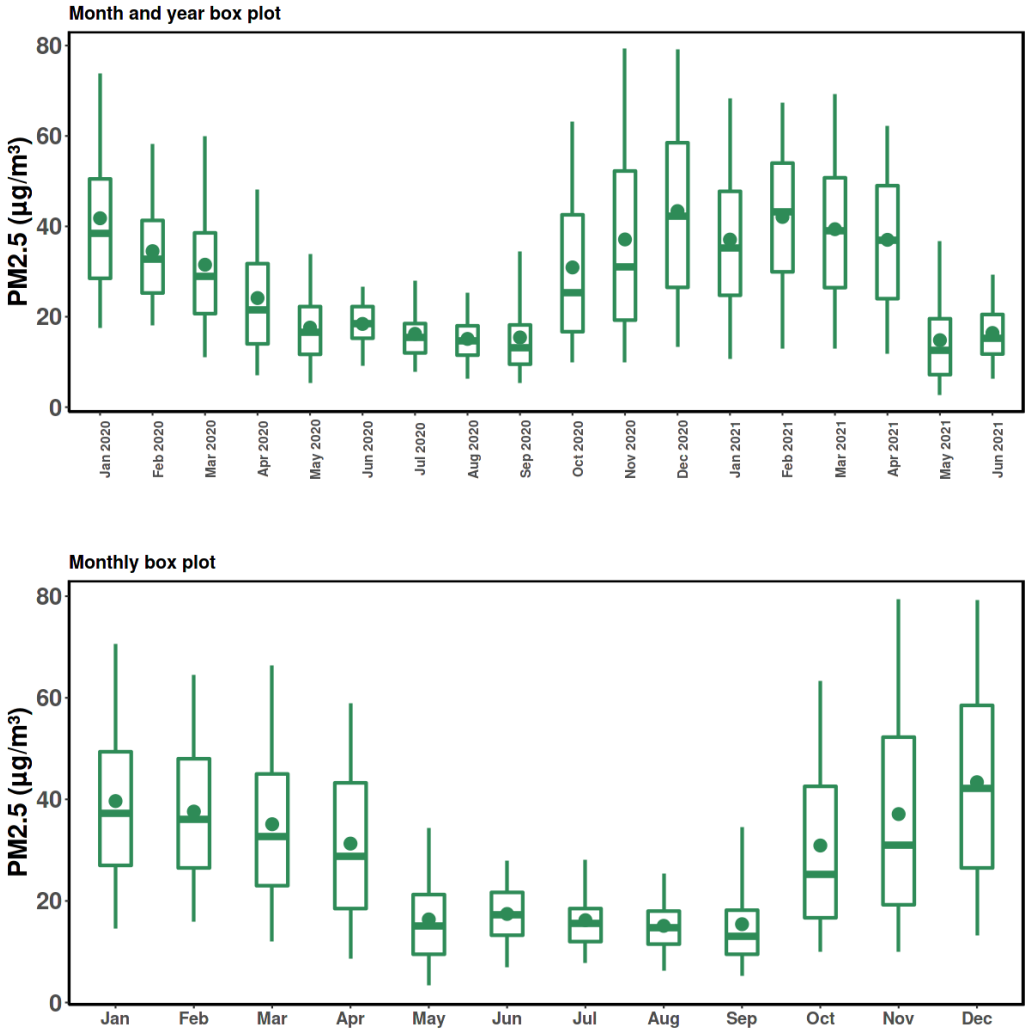

Figure 2: Box plots depicting the monthly variations in hourly $PM_{2.5}$.

Diurnal variation in $PM_{2.5}$ based on mean (and standard deviation) and median (and interquartile range) are shown in the top and bottom panels of Figure 3, respectively. The choice between mean and median is useful when the distribution of the data deviates from normal. In the top panel, the line depicts the mean and the vertical bars depict standard deviation. In the bottom panel, the line depicts the median and the vertical bars depict the interquartile range.

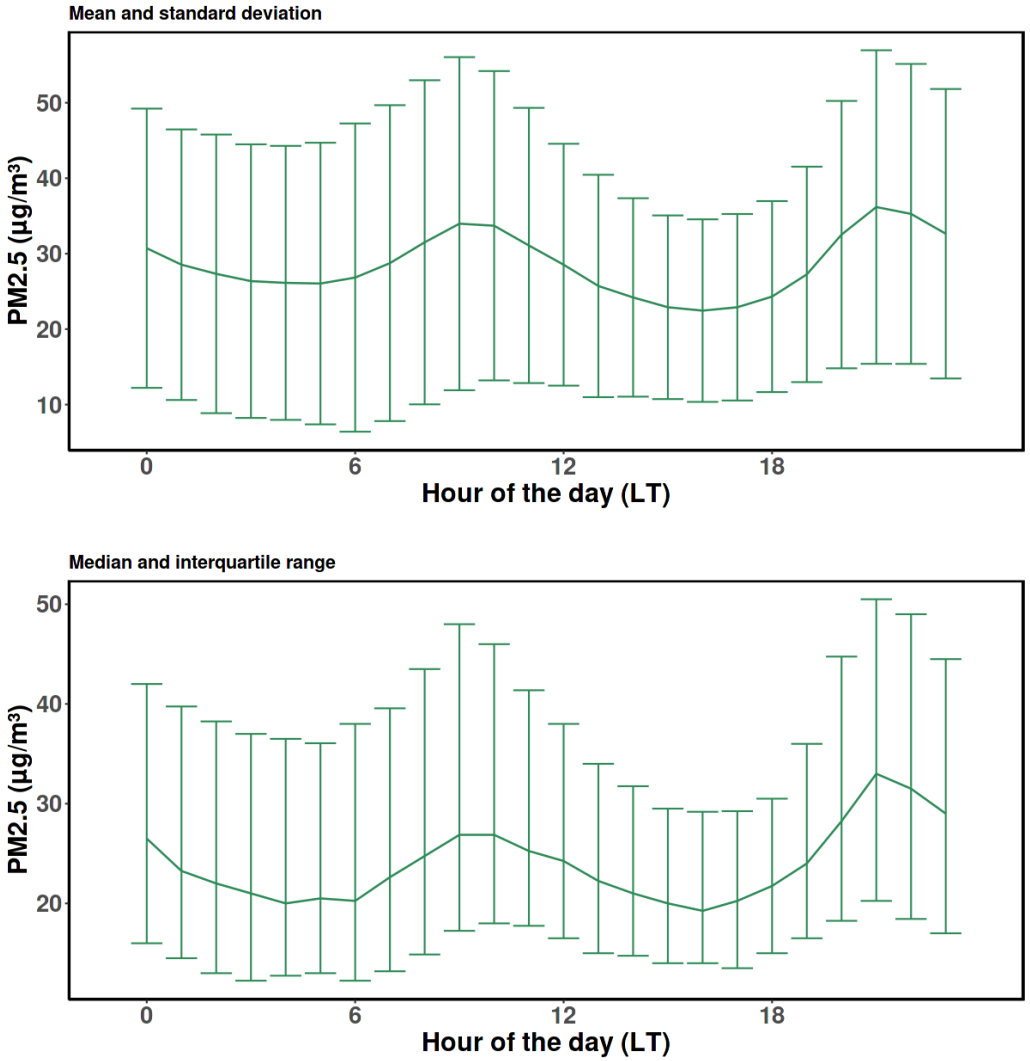

Figure 3: Diurnal variations in $PM_{2.5}$.

In Figure 4, a linear regression is shown between $PM_{2.5}$ and the $PM_{2.5}/PM_{10}$ ratio. The app computes the ratio using the individual $PM_{2.5}$ and PM_{10} data sets. The blue line depicts the least square linear fit. The R-square and the equation of the linear fit are also provided on the panel.

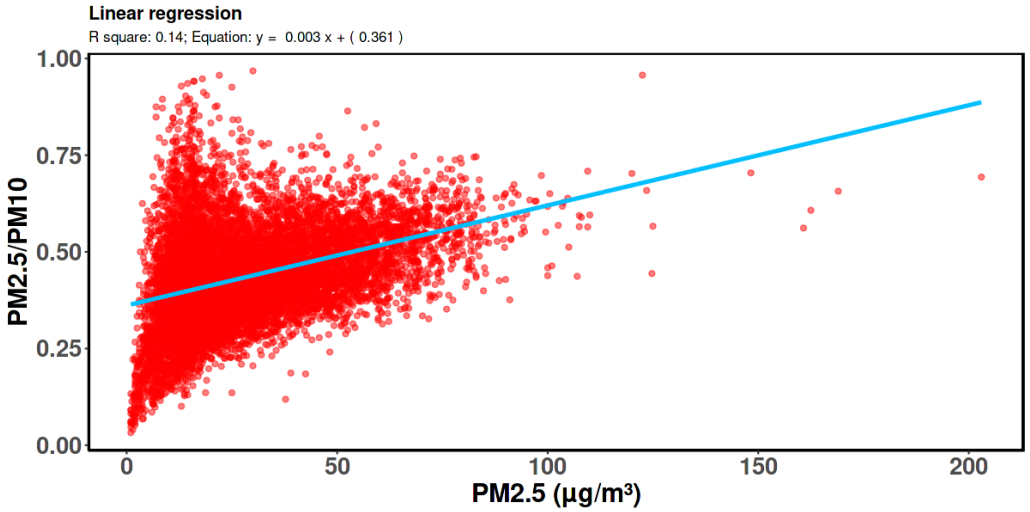

Figure 4: Linear regression analysis.

The periodicity in $PM_{2.5}$ is shown as a wavelet periodogram (Figure 5). Wavelet analysis is useful in analysing non-stationary time series data. Only daily averaged data will be used for this analysis and missing data is imputed to perform the wavelet analysis.

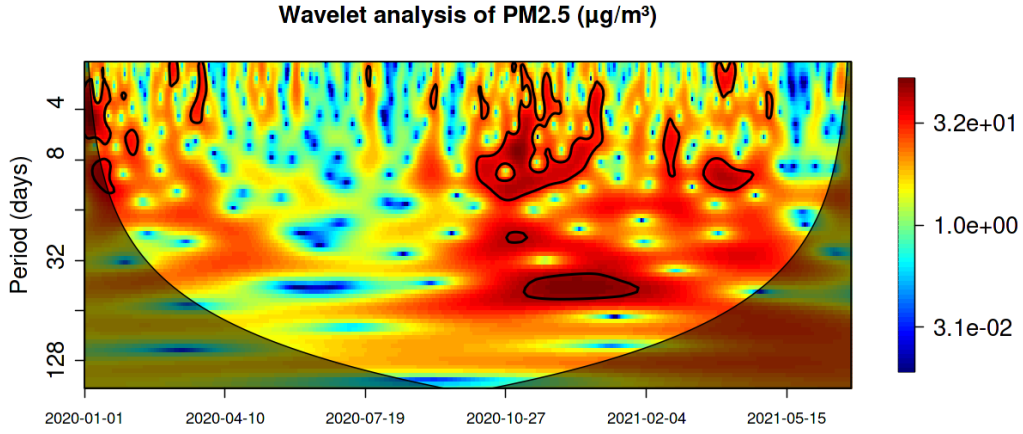

Figure 5: $PM_{2.5}$ periodicity analysis based on wavelet transform.

Acknowledgements

We wish to thank Prof. Julian D. Marshall (University of Washington, Seattle), Prof. Joshua Apte (University of California, Berkeley), Dr. Jai Asundi (Center for Study of Science, Technology and Policy, Bengaluru), Dr Saumya Singh (University of California, Berkeley), and the R community for their help and support.

References

Carslaw, D. C., & Ropkins, K. (2012). openair — An R package for air quality data analysis. *Environmental Modelling & Software*, *27–28*(0), 52–61. https://doi.org/10.1016/j.envsoft.2011.09.008

Chan, C. K., & Yao, X. (2008). Air pollution in mega cities in China. *Atmospheric Environment*, *42*(1), 1–42. https://doi.org/10.1016/j.atmosenv.2007.09.003

Chang, W., Cheng, J., Allaire, J., Sievert, C., Schloerke, B., Xie, Y., Allen, J., McPherson, J., Dipert, A., & Borges, B. (2021). *shiny: Web Application Framework for R*. https://CRAN.R-project.org/package=shiny

Chu, H.-J., Huang, B., & Lin, C.-Y. (2015). Modeling the spatio-temporal heterogeneity in the PM10-PM2.5 relationship. *Atmospheric Environment*, *102*, 176–182. https://doi.org/10.1016/j.atmosenv.2014.11.062

Hyndman, R. J., & Khandakar, Y. (2008). Automatic time series forecasting: the forecast package for R. *Journal of Statistical Software*, *26*(3), 1–22. https://doi.org/10.18637/jss.v027.i03

Hystad, P., Yusuf, S., & Brauer, M. (2020). *Air pollution health impacts: the knowns and unknowns for reliable global burden calculations*. Oxford University Press. https://doi.org/10.1093/cvr/cvaa092

Salmon, M., Vakacherla, S., Milà, C., Marshall, J., & Tonne, C. (2017). rtimicropem: an R package supporting the analysis of RTI MicroPEM output files. *Journal of Open Source Software*, *2*(16), 333. https://doi.org/10.21105/joss.00333

Spandana, B., Rao, S. S., Upadhya, A. R., Kulkarni, P., & Sreekanth, V. (2021). PM2.5/PM10 ratio characteristics over urban sites of India. *Advances in Space Research*, *67*(10), 3134–3146. https://doi.org/10.1016/j.asr.2021.02.008

Upadhya, A. R., Agrawal, P., Vakacherla, S., & Kushwaha, M. (2020). mmaqshiny v1.0: R-Shiny package to explore Air-Quality Mobile-Monitoring data. *Journal of Open Source Software*, *5*(50), 2250. https://doi.org/10.21105/joss.02250