*Journal of Air Pollution and Health (Winter 2019); 4(1): 73-80*

**Software Developments and News**

Available online at http://japh.tums.ac.ir

# Developing codes for validation of $PM_{10}$, $PM_{2.5}$, and $O_3$ datasets using R programming language

***Ramin Nabizadeh*[1], *Mostafa Hadei*[1,*]**

[1] Department of Environmental Health Engineering, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran

**ARTICLE INFORMATION**

*Article Chronology:*
Received 21 December 2018
Revised 5 January 2019
Accepted 29 January 2019
Published 29 March 2019

*Keywords:*
Exposure assessment; Particulate matter; Air pollution; Epidemiology; Health impact assessment

**CORRESPONDING AUTHOR:**

Mostafa.hadei@gmail.com
Tel: (+98 21) 88954914
Fax: (+98 21) 88954914

**ABSTRACT:**

**Introduction:** The wide range of studies on air pollution requires accurate and reliable datasets. However, due to many reasons, the measured concentrations may be incomplete or biased. The development of an easy-to-use and reproducible exposure assessment method is required for researchers. Therefore, in this article, we describe and present a series of codes written in R Programming Language for data handling, validating and averaging of $PM_{10}$, $PM_{2.5}$, and $O_3$ datasets.

**Findings:** These codes can be used in any types of air pollution studies that seek for PM and ozone concentrations that are indicator of real concentrations. We used and combined criteria from several guidelines proposed by US EPA and APHEKOM project to obtain an acceptable methodology. Separate .csv files for $PM_{10}$, $PM_{2.5}$ and $O_3$ should be prepared as input file. After the file was imported to the R Programming software, first, negative and zero values of concentrations within all the dataset will be removed. Then, only monitors will be selected that have at least 75% of hourly concentrations. Then, 24-h averages and daily maximum of 8-h moving averages will be calculated for PM and ozone, respectively. For output, the codes create two different sets of data. One contains the hourly concentrations of the interest pollutant ($PM_{10}$, $PM_{2.5}$, or $O_3$) in valid stations and their average at city level. Another is the final 24-h averages of city for $PM_{10}$ and $PM_{2.5}$ or the final daily maximum 8-h averages of city for $O_3$.

**Conclusion:** These validated codes use a reliable and valid methodology, and eliminate the possibility of wrong or mistaken data handling and averaging. The use of these codes are free and without any limitation, only after the citation to this article.

## Introduction

Air pollution has been introduced as the fifth risk factor for health in the world [1]. High concentrations of particulate matter with aerodynamic diameter <10 µm ($PM_{10}$) and <2.5 µm ($PM_{2.5}$), ozone ($O_3$), nitrogen dioxide ($NO_2$), etc. have been recorded worldwide [2-4]. This has brought a significant attention to air pollution research. Many epidemiological studies have been investigating the relationships between exposure to various air pollutants and different health outcomes [5, 6]. Health impact assessment studies evaluates the impact of reduction or increase in air pollution levels caused by the environmental, industrial and political strategies and decisions. In addition, the study on the temporal trends, spatial variability, and affecting factors of air pollution is always important in urban environments [7].

Please cite this article as: Nabizadeh R, Hadei M. Developing codes for validation of $PM_{10}$, $PM_{2.5}$, and $O_3$ datasets using R programming language. Journal of Air Pollution and Health. 2019; 4(1): 73-80.

The wide range of studies about air pollution requires accurate and reliable datasets [8]. In addition to short-term campaigns of air quality measurements for research purposes, many cities provide continuous measurements using air quality monitoring stations. However, due to many reasons such as power outage, device failure, lack of calibration, etc., the collected data may be incomplete or biased [9]. The use of this raw data in research leads to the inaccurate, unreliable, and biased results. Therefore, an important step in air pollution studies using the data from air quality monitors is to handle and evaluate the validity of raw datasets [10]. In addition, since different types of air quality monitors have been designed and operated, the certain types of monitors should be selected for different research objectives [11-13]. In should be noted that these procedures are different from those used in the data audit of monitors after the data obtained.

Health-related agencies and organizations have published several guidelines for air quality data handling and validating, i.e. exposure assessment [8, 11, 14]Work Package 5, Deliverable D5, April 2011. The main objective of these guidelines is to obtain the air pollution concentrations that are representative of the population's average exposure to air pollution [15]. The first criteria is the selection of proper stations. For instance, the proper stations for health impact assessment of particulate matter, $NO_2$, and ozone are urban, urban, and urban + suburban stations, respectively. The second and main criteria in these guidelines is the percentage of data completeness that are proposed to be more than 50%, 75% or even 90%. In addition, zero, negative and other logically invalid values that sometimes are present in dataset should be deleted. Fourth, the concentrations of pollutants should be averaged over certain time periods. For particulate matter and ozone, 24-h averages and maximum of 8-h moving averages during a day are normally accepted. The averages should be calculated only for days that 75% of hourly concentrations (18 h) are present. In case of 8-h averaging, the 8-h averages should be calculated for those parts of a day that 75% of data (6 h) are present.

In conclusion, exposure assessment is a highly critical step in air pollution studies. Several theoretical guidelines are developed by national and international agencies. However, those guidelines may not be applicable and feasible in most cases, especially in other parts of the world. In addition, the development of an easy-to-use and reproducible exposure assessment method is required for researchers. Therefore, in this article, we describe and present a series of codes written in R Programming Language for data handling, validating and averaging of $PM_{10}$, $PM_{2.5}$, and $O_3$ datasets.

## Software Description

### *The usage of software*

We developed easy-to-use codes written in R Programming Language for data handling, validating and averaging of $PM_{10}$, $PM_{2.5}$, and $O_3$. These codes can be used in any types of air pollution studies that seek for PM and ozone concentrations that are indicator for those concentrations have been experienced in urban environments. These studies can be epidemiological studies, health impact assessments, spatiotemporal investigations, etc.

### *The methods and equations used*

We used and combined criteria from several guidelines proposed by US EPA and APHEKOM project to obtain an acceptable methodology [8, 11]Work Package 5, Deliverable D5, April 2011 plicable in many cities and for many datasets. Fig. 1 shows the flow diagram of inputs, calcula-

tion steps, and outputs of codes presented in this article. It should be noted that all of these steps will be accomplished automatically, once after the user opens the codes in R Programming software, and runs them.

In case of input file, for PM and $O_3$ separate .csv files should be prepared. This comma delimited .csv file that should be named "Dataset" consists of unlimited number of columns that each column contains hourly concentrations of PM or $O_3$ in each monitor. Each column should have a title, i.e. the name of monitor. The user should set the working folder of R software to the folder that "Dataset" exists, using these steps: file > Change dir.

After the file was imported to the R Programming software, first, negative values of concentrations within all the dataset will be removed. Second, zero concentrations in the dataset will be deleted. After the logically invalid values were removed from each monitor, only monitors will be selected that have at least 75% of hourly concentrations, i.e. do not have more than 25% missing concentrations. Fourth, the concentration of pollutant in each hour among all the monitors will be averaged to obtain the hourly concentration at city levels.

Then, 24-h averages and daily maximum of 8-h moving averages will be calculated for PM and ozone, respectively. For calculating 24-h averages of $PM_{10}$ and $PM_{2.5}$, the codes calculates the arithmetic averages of concentrations during each day. Finally, 365 or 366 daily (24-h) averages will be calculated for a whole year. For calculating daily maximum of 8-h moving averages of $O_3$, first, the codes calculates the arithmetic 8-h moving averages during each day. For each day, twenty-four 8-h moving averages will be obtained in this step. Then, the maximum of those 8-h moving averages in each day will be selected and reported as the "daily maximum of 8-h moving average" of ozone. Finally, 365 or 366 daily maximum of 8-h averages will be selected for a whole year.

For output, the codes create two different sets of data. One contains the hourly concentrations of the interest pollutant ($PM_{10}$, $PM_{2.5}$, or $O_3$) in valid stations and their average at city level. Another is the final 24-h averages of city for $PM_{10}$ and $PM_{2.5}$ or the final daily maximum 8-h averages of city for $O_3$. These two file will be in the .csv format.

***Advantages and limitations***

These codes solve a major problems for researchers working on air pollution. These validated codes use a reliable and valid methodology, and eliminate the possibility of wrong or mistaken data handling and averaging. The user has the least involvement in the process, and only should prepare a raw dataset, and enter and run the codes in the R Programming software. Another advantage is the generation of two different sets of data that can cover the users' requirements for their research purposes. In addition, the use of these codes do not need any other packages, and all the calculations are set to be performed only by using built-in codes of R.

However, the use of these codes may have some limitations. First, the user should have some basic knowledge about how to work with R Programming software. Although, these codes require very low levels of R knowledge. Second, researchers may want to conduct their studies on other pollutants rather than $PM_{10}$, $PM_{2.5}$, and $O_3$, and this methodology cannot cover their needs. Of course, that can be the subject of future developments of these series of codes.

***Practical usage of software***

In this section, we present and test the instruction and performance of the codes developed for the validation of $PM_{10}$, $PM_{2.5}$, and $O_3$ concentrations. Fig. 2 shows an example of input data for

the codes. This prototype dataset is actually the real $PM_{2.5}$ concentrations recorded in 27 air quality monitors of Tehran during 2017-2018. Every column belongs to the hourly concentrations of $PM_{2.5}$ in a specific monitor. The same format should be provided for $PM_{10}$ and $O_3$ data. This comma delimited .csv file should be named as "Dataset".

Then, the user should set the working folder of R software to the folder that the "Dataset" file exist. Fig. 2 shows the procedure for setting the direction of working folder in R software.

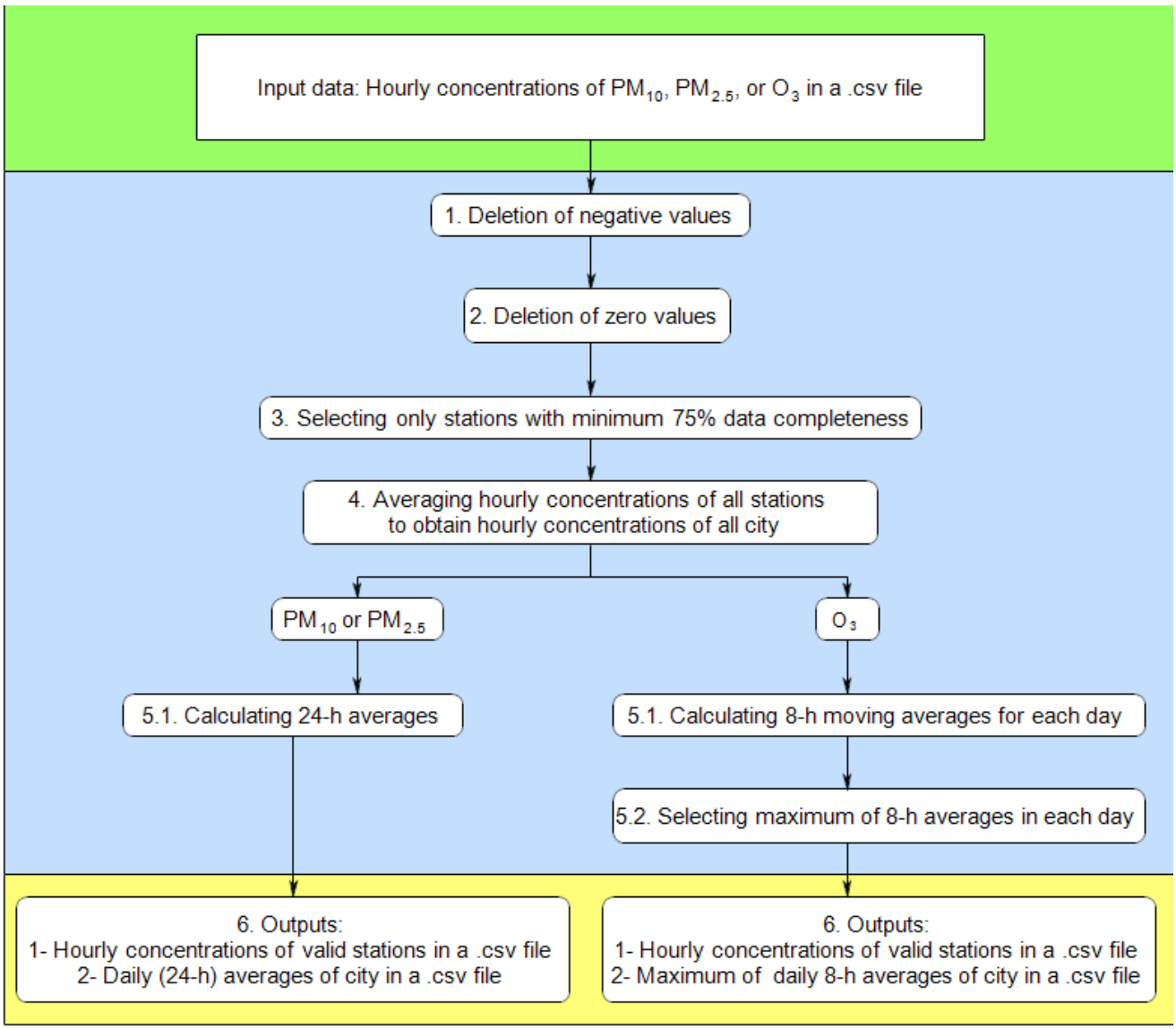
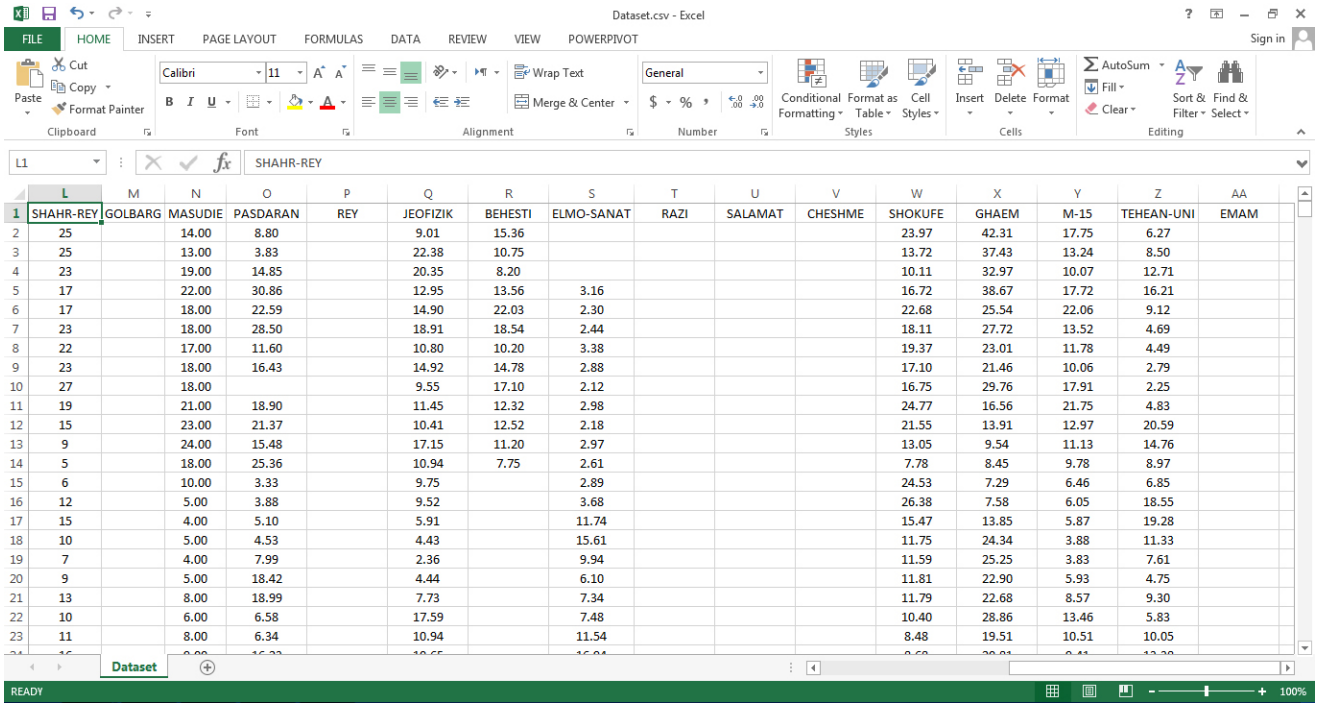



Fig. 1. Flow diagram of inputs, calculation steps, and outputs



| L | M | N | O | P | Q | R | S | T | U | V | W | X | Y | Z | AA |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| SHAHR-REY | GOLBARG | MASUDIE | PASDARAN | REY | JEOFIZIK | BEHESTI | ELMO-SANAT | RAZI | SALAMAT | CHESHME | SHOKUFE | GHAEM | M-15 | TEHEAN-UNI | EMAM |
| 25 | | 14.00 | 8.80 | | 9.01 | 15.36 | | | | | 23.97 | 42.31 | 17.75 | 6.27 | |
| 25 | | 13.00 | 3.83 | | 22.38 | 10.75 | | | | | 13.72 | 37.43 | 13.24 | 8.50 | |
| 23 | | 19.00 | 14.85 | | 20.35 | 8.20 | | | | | 10.11 | 32.97 | 10.07 | 12.71 | |
| 17 | | 22.00 | 30.86 | | 12.95 | 13.56 | 3.16 | | | | 16.72 | 38.67 | 17.72 | 16.21 | |
| 17 | | 18.00 | 22.59 | | 14.90 | 22.03 | 2.30 | | | | 22.68 | 25.54 | 22.06 | 9.12 | |
| 23 | | 18.00 | 28.50 | | 18.91 | 18.54 | 2.44 | | | | 18.11 | 27.72 | 13.52 | 4.69 | |
| 22 | | 17.00 | 11.60 | | 10.80 | 10.20 | 3.38 | | | | 19.37 | 23.01 | 11.78 | 4.49 | |
| 23 | | 18.00 | 16.43 | | 14.92 | 14.78 | 2.88 | | | | 17.10 | 21.46 | 10.06 | 2.79 | |
| 27 | | 18.00 | | | 9.55 | 17.10 | 2.12 | | | | 16.75 | 29.76 | 17.91 | 2.25 | |
| 19 | | 21.00 | 18.90 | | 11.45 | 12.32 | 2.98 | | | | 24.77 | 16.56 | 21.75 | 4.83 | |
| 15 | | 23.00 | 21.37 | | 10.41 | 12.52 | 2.18 | | | | 21.55 | 13.91 | 12.97 | 20.59 | |
| 9 | | 24.00 | 15.48 | | 17.15 | 11.20 | 2.97 | | | | 13.05 | 9.54 | 11.13 | 14.76 | |
| 5 | | 18.00 | 25.36 | | 10.94 | 7.75 | 2.61 | | | | 7.78 | 8.45 | 9.78 | 8.97 | |
| 6 | | 10.00 | 3.33 | | 9.75 | | 2.89 | | | | 24.53 | 7.29 | 6.46 | 6.85 | |
| 12 | | 5.00 | 3.88 | | 9.52 | | 3.68 | | | | 26.38 | 7.58 | 6.05 | 18.55 | |
| 15 | | 4.00 | 5.10 | | 5.91 | | 11.74 | | | | 15.47 | 13.85 | 5.87 | 19.28 | |
| 10 | | 5.00 | 4.53 | | 4.43 | | 15.61 | | | | 11.75 | 24.34 | 3.88 | 11.33 | |
| 7 | | 4.00 | 7.99 | | 2.36 | | 9.94 | | | | 11.59 | 25.25 | 3.83 | 7.61 | |
| 9 | | 5.00 | 18.42 | | 4.44 | | 6.10 | | | | 11.81 | 22.90 | 5.93 | 4.75 | |
| 13 | | 8.00 | 18.99 | | 7.73 | | 7.34 | | | | 11.79 | 22.68 | 8.57 | 9.30 | |
| 10 | | 6.00 | 6.58 | | 17.59 | | 7.48 | | | | 10.40 | 28.86 | 13.46 | 5.83 | |
| 11 | | 8.00 | 6.34 | | 10.94 | | 11.54 | | | | 8.48 | 19.51 | 10.51 | 10.05 | |

Fig. 2. An example of input data for the validation formula

Once the direction folder was set, the user should copy and paste the codes to R editor. Fig. 4 shows the codes for $PM_{2.5}$ validation in R editor. The six un-selected lines above are the guide to use the codes. The blue-highlighted part is the codes that user should select and run. After running the codes, R automatically loads the "Dataset", and performs all of the calculation steps, and creates two different sets of outputs as it was mentioned before. The same process can be performed for $PM_{10}$ and $O_3$.

Tables 1 and 2 show the outputs of validation codes in form of .csv files. In Table 1, the hourly concentrations of 15 valid air quality monitors and hourly averages of city are presented. Remember that the initial raw data contained the data from 27 monitors, that validation codes have excluded 12 invalid monitors. In Table 2, another outputs of validation codes are illustrated. This .csv file includes daily averages of $PM_{2.5}$ in the city. Similar files will be created for $PM_{10}$ and ozone, except that in case of ozone, the daily average file includes 365 (or 366) daily maximum of 8-h moving averages.

**Availability and requirements**

The package of three sets of codes for validation $PM_{10}$, $PM_{2.5}$ and $O_3$ datasets are available at: (https://bit.ly/2sJtAi5). The codes will be also freely available upon request from the corresponding author (Mostafa.hadei@gmail.com). These codes have been written using R Programming Language. The use of these codes are free and without any limitation, only after the citation to this article.

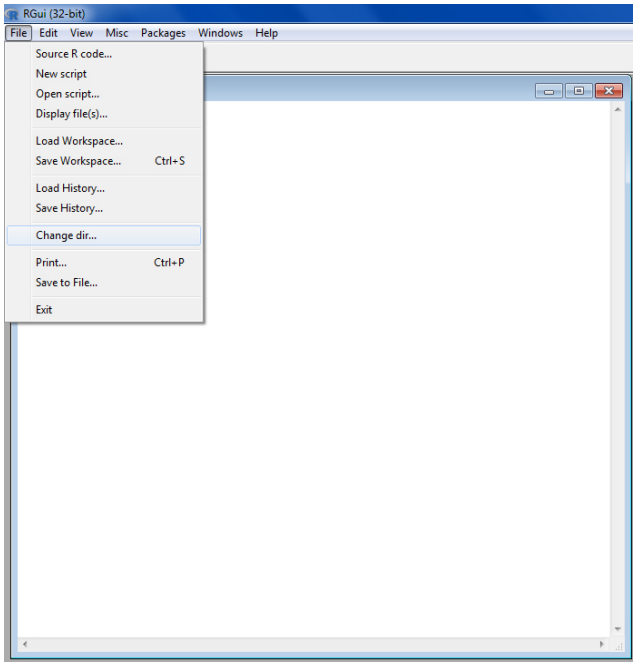



Fig. 3. Setting the direction of working folder in R software

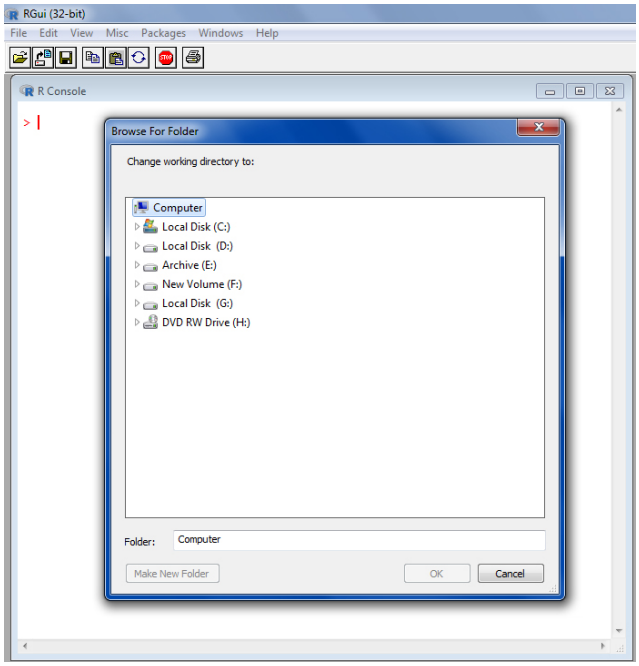



Fig. 4. Codes for $PM_{2.5}$ validation in R editor

Table 1. Output: hourly concentrations of valid air quality monitors and hourly averages of city

| | PONAK | SHARIF | MASUDIE | PASDARAN | BEHESTI | Hourly_Average |
|---|---|---|---|---|---|---|
| 1 | 16.00 | 18.00 | 14.00 | 8.80 | 15.36 | 16.10 |
| 2 | 20.00 | 19.00 | 13.00 | 3.83 | 10.75 | 14.43 |
| 3 | 22.00 | 19.00 | 19.00 | 14.85 | 8.20 | 16.00 |
| 4 | 23.00 | 10.00 | 22.00 | 30.86 | 13.56 | 17.03 |
| 5 | 18.00 | 8.00 | 18.00 | 22.59 | 22.03 | 16.05 |
| 6 | 11.00 | 13.00 | 18.00 | 28.50 | 18.54 | 16.31 |
| 7 | 11.00 | 14.00 | 17.00 | 11.60 | 10.20 | 13.68 |
| 8 | 16.00 | 14.00 | 18.00 | 16.43 | 14.78 | 14.40 |
| 9 | 21.00 | 12.00 | 18.00 | NA | 17.10 | 15.10 |
| 10 | 21.00 | 10.00 | 21.00 | 18.90 | 12.32 | 14.94 |
| 11 | 19.00 | 16.00 | 23.00 | 21.37 | 12.52 | 14.53 |
| 12 | 15.00 | NA | 24.00 | 15.48 | 11.20 | 13.14 |
| 13 | 20.00 | 15.00 | 18.00 | 25.36 | 7.75 | 13.62 |
| 14 | 12.00 | 14.00 | 10.00 | 3.33 | NA | 8.19 |
| 15 | 8.00 | 13.00 | 5.00 | 3.88 | NA | 7.07 |
| … | … | … | … | … | … | … |

Table 2. Output: daily averages of $PM_{2.5}$ in the city

| | Daily_Averages |
|---|---|
| 1 | 12.33 |
| 2 | 8.82 |
| 3 | 11.84 |
| 4 | 12.46 |
| 5 | 16.29 |
| 6 | 19.26 |
| 7 | 19.54 |
| 8 | 16.07 |
| 9 | 14.62 |
| 10 | 13.46 |
| 11 | 19.70 |
| 12 | 16.55 |
| 13 | 16.32 |
| 14 | 14.02 |
| 15 | 20.17 |
| … | … |

**Financial supports**

These codes were developed during a course of R programming in Tehran University of Medical Sciences.

**Competing interests**

The authors indicate that there are no conflicts of interests.

**Acknowledgements**

The authors wish to thank Department of Environmental Health Engineering, Tehran University of Medical Sciences.

**Ethical considerations**

Ethical issues (including plagiarism, informed consent, misconduct, data fabrication and/or falsification, double publication and/or submission, redundancy, etc) have been completely observed by the authors. The use of these codes are free and without any limitation, only after the citation to this article.

**References**

1. Forouzanfar MH, Afshin A, Alexander LT, Anderson HR, Bhutta ZA, Biryukov S, et al. Global, regional, and national comparative risk assessment of 79 behavioural, environmental and occupational, and metabolic risks or clusters of risks, 1990–2015: a systematic analysis for the Global burden of disease study 2015. The Lancet. 2016;388(10053):1659-724.
2. WHO. WHO's urban ambient air pollution database. 2016 update 2016.
3. Yarahmadi M, Hadei M, Nazari SSH, Conti GO, Alipour MR, Ferrante M, et al. Mortality assessment attributed to long-term exposure to fine particles in ambient air of the megacity of Tehran, Iran. Environmental science and pollution research. 2018;25(14):14254-62.
4. Hopke PK, Hashemi Nazari SS, Hadei M, Yarahmadi M, Kermani M, Yarahmadi E, et al. Spatial and temporal trends of short-term health impacts of PM2.5 in Iranian cities; a modelling approach (2013-2016). Aerosol and air quality research. 2018.
5. Hassanvand MS, Naddafi K, Malek M, Valojerdi AE, Mirzadeh M, Samavat T, et al. Effect of long-term exposure to ambient particulate matter on prevalence of type 2 diabetes and hypertension in Iranian adults: an ecologic study. Environmental science and pollution research. 2017:1-6.
6. Beelen R, Raaschou-Nielsen O, Stafoggia M, Andersen ZJ, Weinmayr G, Hoffmann B, et al. Effects of long-term exposure to air pollution on natural-cause mortality: an analysis of 22 European cohorts within the multicentre ESCAPE project. The lancet. 2014;383(9919):785-95.
7. Hadei M, Hopke PK, Nazari SSH, Yarahmadi M, Shahsavani A, Alipour MR. Estimation of mortality and hospital admissions attributed to criteria air pollutants in Tehran metropolis, Iran (2013–2016). Aerosol and air quality research. 2017;17:2474-81.
8. Pascal M, Corso M, Ung A. Guidelines for assessing the health impacts of air pollution in European cities. Aphekom project, Work package 5, deliverable D5, April 2011. 2011.
9. EPA. Quality assurance handbook for air pollution measurement systems; Ambient air quality monitoring program 2013.
10. Liu M, Huang Y, Ma Z, Jin Z, Liu X, Wang H, et al. Spatial and temporal trends in the mortality burden of air pollution in China: 2004–2012. Environment International. 2017;98:75-81.
11. USEPA. Chapter 11: Valid data and completeness requirements. In: Agency USEP, editor. USA2015.
12. Hadei M, Nazari SSH, Eslami A, Khosravi A, Yarahmadi M, Naghdali Z, et al. Distribution and number of ischemic heart disease (IHD) and stroke deaths due to chronic exposure to PM2.5 in 10 cities of Iran (2013-2015); and AIRQ+ modelling. Journal of air pollution and health. 2018;2(3):129-36.
13. Hadei M, Nazari SSH, Yarahmadi E, Kermani M, Yarahmadi M, Naghdali Z, et al. Estimation of lung can-

cer mortality attributed to long-term exposure to PM2.5 in 15 cities during 2015-2016; an AIRQQ+ modelling. Journal of air pollution and health. 2017;2(1):19-26.
14. World Health Organization. Monitoring ambient air quality for health impact assessment. 1999.
15. Brauer M, Freedman G, Frostad J, van Donkelaar A, Martin RV, Dentener F, et al. Ambient air pollution exposure estimation for the global burden of disease 2013. Environmental Science & Technology. 2016;50(1):79-88.