Original Article

A Framework for Exploration and Cleaning of Environmental Data – Tehran Air Quality Data Experience

Mansour Shamsipour MSc PhD Candidate[1,2,3], Farshad Farzadfar MD MPH DSc[3,4], Kimiya Gohari BSc[5,3], Mahboubeh Parsaeian MSc PhD Candidate[1,3], Hassan Amini MSPH PhD Candidate[6,7,8], Katayoun Rabiei MD MPH PhD Candidate[9], Mohammad Sadegh Hassanvand PhD[2,10], Iman Navidi BSc[1,3], Akbar Fotouhi MD PhD[1], Kazem Naddafi PhD[2,10], Nizal Sarrafzadegan MD[9], Anita Mansouri BSc[5,3], Alireza Mesdaghinia PhD[2,11], Bagher Larijani MD[4], Masud Yunesian MD PhD•[2,10]

Abstract

Background: Management and cleaning of large environmental monitored data sets is a specific challenge. In this article, the authors present a novel framework for exploring and cleaning large datasets. As a case study, we applied the method on air quality data of Tehran, Iran from 1996 to 2013.

Methods: The framework consists of data acquisition [here, data of particulate matter with aerodynamic diameter ≤10 μm (PM_{10})], development of databases, initial descriptive analyses, removing inconsistent data with plausibility range, and detection of missing pattern. Additionally, we developed a novel tool entitled spatiotemporal screening tool (SST), which considers both spatial and temporal nature of data in process of outlier detection. We also evaluated the effect of dust storm in outlier detection phase.

Results: The raw mean concentration of PM_{10} before implementation of algorithms was 88.96 μg/m^3 for 1996–2013 in Tehran. After implementing the algorithms, in total, 5.7% of data points were recognized as unacceptable outliers, from which 69% data points were detected by SST and 1% data points were detected via dust storm algorithm. In addition, 29% of unacceptable outlier values were not in the PR.

The mean concentration of PM_{10} after implementation of algorithms was 88.41 μg/m^3. However, the standard deviation was significantly decreased from 90.86 μg/m^3 to 61.64 μg/m^3 after implementation of the algorithms. There was no distinguishable significant pattern according to hour, day, month, and year in missing data.

Conclusion: We developed a novel framework for cleaning of large environmental monitored data, which can identify hidden patterns. We also presented a complete picture of PM_{10} from 1996 to 2013 in Tehran. Finally, we propose implementation of our framework on large spatiotemporal databases, especially in developing countries.

Keywords: Air pollution, air quality data management, EBD-NASBOD, Iran, outlier detection, Tehran

Cite this article as: Shamsipour M, Farzadfar F, Gohari K, Parsaeian M, Amini H, Rabiei K, Hassanvand MS, Navidi I, Fotouhi A, Naddafi K, Sarrafzadegan N, Mansouri A, Mesdaghinia A, Larijani B, Yunesian M. A Framework for Exploration and Cleaning of Environmental Data – Tehran Air Quality Data Experience. *Arch Iran Med*. 2014; **17(12):** 821 – 829.

Introduction

Over the past decades, there have been numerous global concerns regarding environmental determinants of health and their attributable health impacts.[1] These adverse events may generate a range of effects from mild effects, such as annoyance, to severe impacts on morbidity and mortality.[2]

The Environmental Burden of Disease (EBD) study[3] is a part of the National and Sub-national Burden of Diseases, Injuries, and Risk Factors (NASBOD) study from 1990 to 2013 in Iran.[4] The EBD and NASBOD aim to estimate the distribution of province year specific exposure of general Iranian population to major environmental risk factors and estimate the burden of diseases attributable to these explanatory determinants of health from 1990 to 2013.[3,4] Air pollution is one of the risk factors that the EBD has looked into. It includes many gaseous compounds and particles. Generally, the major air pollutants that have shown up so far in epidemiologic studies and risk assessment of air pollution are particulate matter with aerodynamic diameter smaller than 10 μm (PM_{10}), sulfur dioxide (SO_2), nitrogen oxides ($NOx = NO + NO_2$), carbon monoxide (CO), and ozone (O_3).[5–7] Notably, PM_{10} has been one of the most important air pollutants with respect to its health effects.[8–10]

Exposure assessment is a crucial component in investigating the relationship between air pollutant and health outcome. Over the past decades, various approaches of exposure assessment have been developed from crude estimates to refined integrated methods that estimate more accurate and precise exposures.[11] Jerrett, et al., (2004) classified within city air pollution exposure models as

Authors' affiliations: [1]Department of Epidemiology and Biostatistics, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran. [2]Center for Air Pollution Research (CAPR), Institute for Environmental Research (IER), Tehran University of Medical Sciences, Tehran, Iran. [3]Non-Communicable Diseases Research Center, Endocrinology and Metabolism Population Sciences Institute, Tehran University of Medical Sciences, Tehran, Iran. [4]Endocrinology and Metabolism Research center, Endocrinology and Metabolism Research Institute, Tehran University of Medical sciences, Tehran, Iran. [5]Department of Biostatistics, Faculty of Paramedical Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [6]Departmentof Epidemiology and Public Health, Swiss Tropical and Public Health Institute (Swiss TPH), Basel, Switzerland.[7]University of Basel, Basel, Switzerland. [8]Kurdistan Environmental Health Research Center, Kurdistan University of Medical Sciences, Sanandaj, Iran. [9]Isfahan Cardiovascular Research Center, Cardiovascular Research Institute, Isfahan University of Medical Sciences, Isfahan, Iran. [10]Department of Environmental Health Engineering, School of Public Health, Tehran University of Medical Sciences, Tehran, Iran.[11]Center for Water Quality Research (CWQR), Institute for Environmental Research (IER), Tehran University of Medical Sciences, Tehran, Iran.
•**Corresponding author and reprint:** Masud Yunesian MD PhD, Center for Air Pollution Research (CAPR), Institute for Environmental Research (IER), Tehran University of Medical Sciences, Tehran, Iran. Telefax: +98-21-8898-3698. E-mail: yunesian@tums.ac.ir.

Accepted for publication: 12 November 2014

proximity, dispersion, land use regression, interpolation, integrated meteorological-emission, and hybrid models.[12] Spatio-temporal and remote sensing approaches are more recent state-of-the-science methods, which have been used for exposure assessment in epidemiologic research.[9,13,14] Overall, measurements of ground monitoring network, except for remote sensing approaches, are an essential input for abovementioned methods.

Air quality monitoring systems are a series of measurements for air pollutants, taken continuously or intermittently over a short- or long-term period through a single station or air quality-monitoring network (AQMN). In such settings, data volume may increase rapidly, especially when new monitoring stations/devices are installed. Although AQMN should pursue specific quality control/quality assurance (QA/QC) procedures, the quality of big retrieved data from monitoring stations is of concern for investigators. Typical challenges that can be found in these time series are: finding outlier data, duplication, huge missing (usually due to faults in data acquisition, machine failure, routine maintenance, changes of the site location, and human errors), and so forth. In the EBD study, the major data sources for exposure assessment of air pollutants are AQMN sites throughout the country. Therefore, it is critical to establish a framework for data quality assessment in such a large-scale study. Tehran is the capital and the most populated city of Iran, and one of the largest cities in South West Asia with more than 8.2 million residents.[15] It is geographically located in valleys and surrounded by medium to high mountains to the north, northwest, east, and southeast. Owing to its large population, numerous cars, and specific geographic location, air pollution is one of the major common problems in the city.[15–17] In this paper, we present a novel framework that can simplify management processes of large datasets through its key components. We present this framework systematically through its implementation on the large datasets of Tehran air quality.

Materials and Methods

The key components of our framework are data acquisition from data source, development of databases, initial descriptive analysis, and detection of missing and outlier data. Outlier detection algorithm include spatio-temporal screening tool, checking for dust, and checking of plausibility range. The flow chart presented in Figure 1 shows the steps of the developed framework.

AQMN: Geographic distribution and equipment of AQMN

The first part of process was to determine the source of data. In our case, AQMN was the source of data. A wide network, comprised of two government agencies including Air Quality Control Company (AQCC) and Department of Environment (DoE), monitors air quality in Tehran. This network has been established to collect and archive air quality data continuously on an hourly basis throughout the year.

According to data we collected from these two agencies, AQCC has 27 active air quality monitoring stations, and DoE has 12 monitoring stations. These stations are distributed within 22 administrative counties of Tehran. Figure 2 shows the locations of these stations. The number of the abovementioned stations has been increasing since their inception in 1996. PM_{10}, is measured in both AQCC and DoE monitoring stations using beta-radiation attenuation instruments or beta-gauge monitors (model MP 101M of Environment SA, France; FH 62 IN, FAG Kugelfischer, Germany; APDA-351E of Horiba, Japan; and instruments of Ecotech, Australia).[16] Noteworthy, measurement of air quality was not under one umbrella until mid-2014 and each of the AQCC and DoE authorities operated their monitoring stations independently. However, currently all stations are operated by one private company.

Data acquisition

The next step in our framework was obtaining available monitoring data through contacting AQCC and DoE offices in Tehran. All available hourly records of PM_{10} obtained from the first measurement in 1996 up to 2013 were collected from all monitoring stations. Table 1 shows the tabulated description of each monitoring station.

Development of database

Hourly measured datasheets from different months and years, obtained from all stations, were stored in scattered excel files. First, we merged all dispersed excel files in a long shaped data frame for all stations (one by one), and created a single datasheet for each station. Moreover, all dates were converted to Gregorian calendar dates. In the original obtained datasheets, there was no missing row for those hours for which there had been no measurement. For example, if one station was active for only two months in a year, there was no row for other months and it was difficult to assess frequency and patterns of missing data. Hence, we separately created a full frame dataset from 1996 to 2013 on an hourly basis (consisting of 212,280 Million cells) and merged the data into this full frame, so the deleted missing hours were restored. Finally, we merged datasheets of all stations together and created a final single dataset needed for the next step of analyses.

Descriptive analysis

Looking into descriptive measures and plots is the first step in data analysis. After developing the final dataset, to get familiar with the data and to detect data quality problems, we carried out some basic descriptive analyses including mean, skewness, kurtosis, SD (standard deviation), minimum, and maximum (Table 2).

Using per station box plots, we investigated the distribution of all measurements before (Figure 3) and after (Figure 4) implementation of algorithms.

Missing values

AQMN stations in Tehran automatically record the concentration of various pollutants on an hourly basis. Therefore, for a complete year, we expected to have a complete datasets with 8760 measured values (24 h/day × 365 days except leap year, which is 8784, 24 h/day × 366 days), but the values had not been measured for many occasions and their related data were missing. As explained above, through merging every station's datasheet into the full frame dataset, we calculated the percentage of annual missing data for all stations and the results are presented in Figure 5. In addition, we evaluated the pattern of missing data by day, month, hour, season, and year to find systematic patternsof missing data.

Outlier detection methods

Spatio-temporal screening tool

We developed a new methodology for detection of spatio-temporal outliers in large environmental monitoring databases, which considered both spatial and temporal relationships. In our algo-

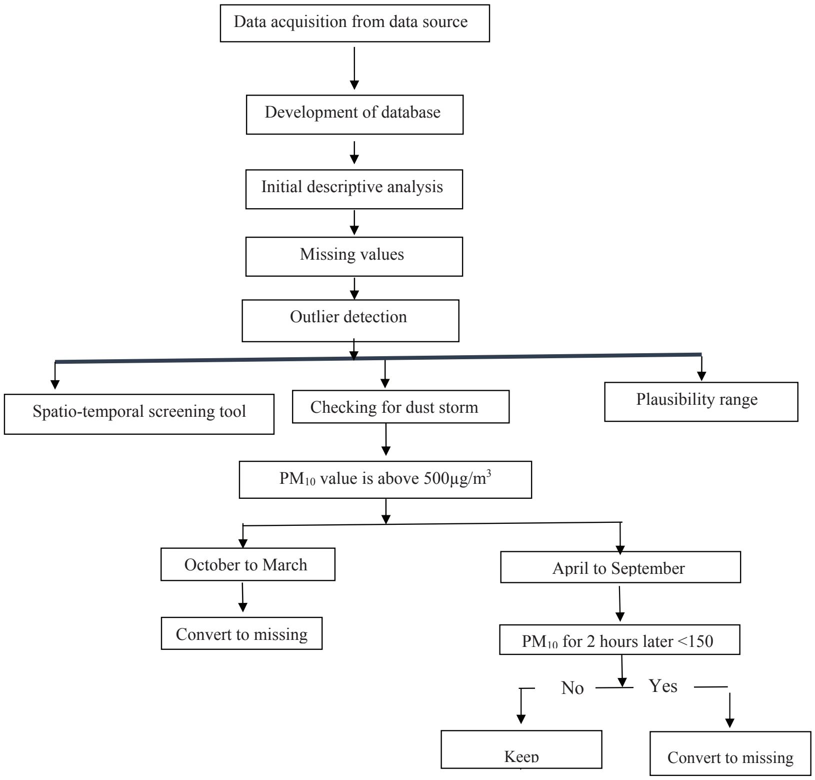

Figure 1. Flow chart representing the different steps of the framework.

Table 1. Air quality monitoring stations in Tehran, Iran.

Station no.	Station name	District	Period of operation	Operated by	x	y
1	Aghdasieh	1	2004–2013	AQCC	543825	3961881
2	Baharan	17	2010–2012	AQCC	532064	3944606
3	Bazar	12	1996–1997,2000-2007	AQCC	538298	3948233
4	Darous	3	2010–2011	AQCC	541051	3958533
5	Fatemi	6	1996–2008	AQCC	536893	3953105
6	Fath	9	2010–2013	AQCC	530544	3948378
7	Geophisics	6	2006–2013	AQCC	534928	3955927
8	Golbarg	8	2008–2013	AQCC	545771	3954234
9	Mahallati	14	2010–2013	AQCC	541878	3947227
10	Masoudyeh	15	2008–2013	AQCC	545185	3943029
11	Ostandari	7	2009–2010	AQCC	538565	3954101
12	Park Rose	22	2000–2013	AQCC	524223	3955132
13	Pirozi	13	2011–2013	AQCC	544672	3950343
14	Pounak	5	2007–2013	AQCC	528415	3957696
15	Setadbohran	12	-------	AQCC	538996	3950343
16	Shadabad	18	2011–2013	AQCC	526561	3947468
17	Shahrdari Mantaghe2	2	2013	AQCC	533276	3959287
18	ShahrdariMantaghe 4	4	2009–2010, 2012-2013	AQCC	544531	3955425
19	Shahrdari Mantaghe7	7	2010–2013	AQCC	540725	3952814
20	ShahrdariMantaghe 10	10	2009–2013	AQCC	532391	3950455
21	ShahrdariMantaghe 11	11	2009–2013	AQCC	535270	3947748
22	ShahrdariMantaghe 16	16	2009–2013	AQCC	536001	3944615
23	ShahrdariMantaghe 19	19	2009–2011	AQCC	532823	3943549
24	Share Rey	20	2006–2013	AQCC	538547	3940175
25	Sanatisharif	2	2012–2013	AQCC	531281	3951389
26	Tarbyatmodares	6	2013	AQCC	534703	3952983
27	Tehransar	21	2007, 2012–2013	AQCC	519401	3952134
28	Emam Khomeini	12	2005–2006, 2008–2013	DOE	538318	3949050
29	Ghadir	22	2011–2012	DOE	523231	3956097
30	Bahman	16	2005–2012	DOE	535737	3944857
31	Gholhak	3	2005–2012	DOE	539717	3958807
32	Tajrish	1	2005–2010	DOE	539235	3962529
33	Pardisan	2	2005–2007, 2010, 2012, 2013	DOE	531885	3956144
34	Azadi	5	2006–2012	DOE	530506	3951224
35	SorkheHesar	13	2006–2007	DOE	552708	3952968
36	ShahidBeheshti University	1	-------	DOE	535714	3962208
37	ShahrakeCheshmeh	22	2010–2013	DOE	523757	3956530
38	ElmuSanat University	4	------	DOE	546217	3955170
39	Ghaem	17	------	DOE	532241	3946115

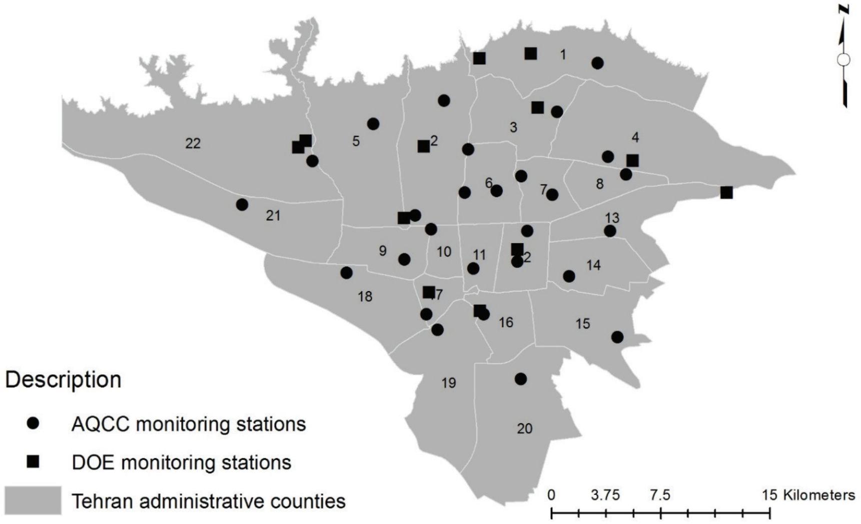
N
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
Description
● AQCC monitoring stations
■ DOE monitoring stations
Tehran administrative counties
0
3.75
7.5
15 Kilometers

Figure 2. Air quality monitoring network in Tehran, Iran.

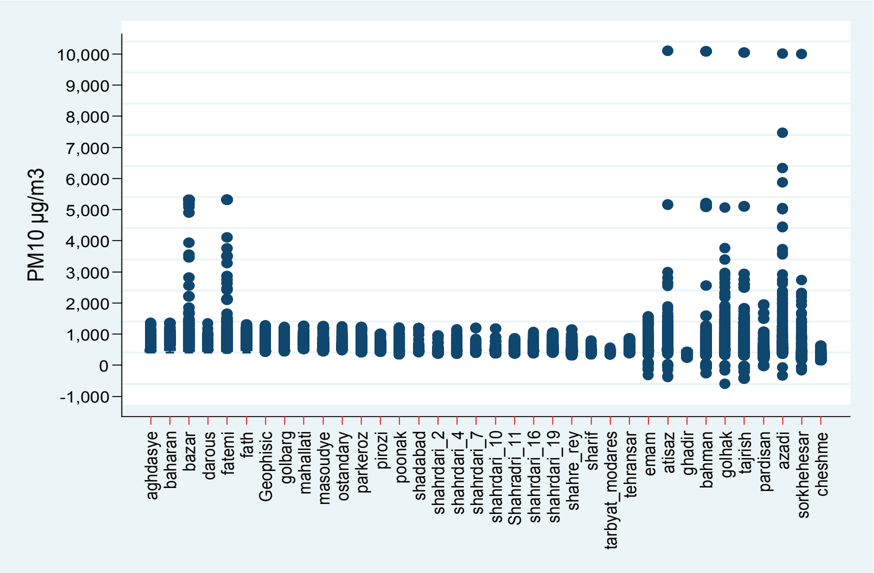

Figure 3. Boxplots per station, distribution of all measurements before implementation of algorithms.

rithm, we defined a spatiotemporal neighborhood domain for each observation; this domain was limited by time (+/- 2 hours) and distance (11 km) from the location of every individual ambient air monitoring station. To make it clear, the neighborhood domain is illustrated in Figure 6.

Based on our proposed algorithm within the given spatio-temporal domain, there is an interrelationship between the attribute values of neighbors; moreover, abnormal values can be detected by a comparison of any measurement with the attribute values of their neighbors in a defined domain. For any measurement in proposed methodology, first, algorithm calculates mean and standard deviation for all abovementioned spatial and temporal neighbors for any measurement.

Equation (1):

$$\bar{x}=\frac{1}{N}\sum_{i=1}^{N} x_i=\frac{x_1+x_2+\ldots+x_N}{N}$$

Where $\bar{x}$ denotes the mean value of spatio-temporal neighbors, x_i denotes every observation value for spatio-temporal neighbors ($X_i \ldots X_n$) and N denotes the total number of neighbors.

Next, the algorithm calculates the upper control limit (UCL) and lower control limit (LCL) values as follow:

$$\text{UCL}=\bar{x}+3\sigma$$
$$\text{LCL}=\bar{x}-3\sigma$$

Equation (2):

$$\sigma=\sqrt{\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2}$$

Table 2. Descriptive analysis summery measures before and after implementation of outlier detection algorithms.

	Before outlier detection algorithm implementation					**After outlier detection algorithm implementation**						
Station name	**Obs**	**Mean**	**Std.**	**Min**	**Max**	**Obs**	**Mean**	**Std.**	**Min**	**Max**	**a**	**b**
Aghdasieh	70679	72.97	54.65	0	986.45	68079	73.77	51.96	10	986.45	2600	0.04
Baharan	7909	131.60	81.64	4	985	7815	129.11	74.84	11	985	94	0.01
Bazar	50731	106.80	172.56	0	4965	47911	102.20	75.34	10	1860	2820	0.06
Darous	6523	100.35	65.17	2	999	6357	99.61	62.19	10	705	166	0.03
Fatemi	89174	85.11	88.09	0	5000	86386	85.38	81.52	10	2430	2788	0.03
Fath	24684	119.34	68.54	4	1000	24237	117.45	64.93	10	903	447	0.02
Geophisics	41628	62.05	47.34	0	986.85	41018	62.15	43.66	10.05	986.85	610	0.01
Golbarg	39934	76.51	65.41	0	956.53	38965	74.90	52.47	10	941.89	969	0.02
Mahallati	25819	119.21	65.76	3	1000	24921	115.60	59.76	10	826	888	0.03
Masoudyeh	24850	88.62	64.62	0	997.22	24613	88.32	61.79	10.07	997.22	237	0.01
Ostandari	10699	115.16	95.18	1.43	995.38	10282	106.05	70.54	10	995.38	417	0.04
Park ROZ	33187	81.25	65.90	0	998.66	32932	80.84	62.50	10.05	998.66	255	0.01
Pirozi	17382	97.43	52.90	6	792	17264	96.63	49.92	10	513	118	0.01
Pounak	43207	66.70	47.03	0	997.72	42946	66.87	46.05	10.01	997.72	261	0.01
Shadabad	7502	77.34	92.44	3	985	7294	70.73	58.48	11	491	208	0.03
Sh.Mantaghe2*	2485	70.65	49.24	0	758.91	2331	68.72	41.42	10.07	381.52	154	0.06
Sh. Mantaghe 4*	21064	74.30	46.62	0.12	961.27	20679	74.88	42.83	10.03	773.92	385	0.02
Sh.Mantaghe7*	9236	95.90	57.07	2	998	9191	95.08	49.18	10	643	45	0.00
Sh. Mantaghe 10*	14937	82.90	44.26	2.04	986.41	14823	82.92	42.13	10.1	441.27	114	0.01
Sh. Mantaghe 11*	20613	76.56	52.84	2.02	676.34	20207	77.56	51.10	10	602.59	406	0.02
Sh.Mantaghe 16*	32964	78.15	58.81	2.28	892.97	31815	79.99	56.03	10	703.13	1149	0.03
Sh.Mantaghe 19*	18344	104.81	71.32	3.15	873.23	18187	102.90	61.24	10	688.71	157	0.01
Share Rey	51760	66.01	47.91	0.02	999.63	50512	67.07	46.22	10.01	837.35	1248	0.02
Sanatisharif	10585	97.27	48.07	12	647	10516	96.60	46.66	12	605	69	0.01
Tarbyatmodares	4353	93.64	44.39	12	422	4324	92.90	43.01	12	379	29	0.01
Tehransar	12109	114.61	66.33	0.88	730	11529	110.47	59.77	12	727	580	0.05
Emam Khomeini	37904	87.37	73.51	-447.9	1451	36162	88.31	67.20	10	1451	1742	0.05
Ghadir	3095	64.14	36.34	0	321.91	3085	64.34	36.22	10.12	321.91	10	0.00
Bahman	47385	93.76	147.90	-367	9999	41397	100.61	61.17	10.03	1054	5988	0.13
Gholhak	52252	105.57	97.96	-690.5	5000	46760	109.50	70.51	10.02	1626	5492	0.11
Tajrish	40368	113.41	119.99	-496.9	9999	37464	114.68	69.40	10.01	1106	2904	0.07
Pardisan	41907	76.89	59.75	-78.97	1922	36689	82.31	52.75	10.01	1047	5218	0.12
Azadi	39558	140.32	158.22	-361.4	9999	33483	143.46	86.68	10.05	1447	6075	0.15
SorkheHesar	41297	58.59	83.77	-177.4	9999	31689	70.98	51.90	10.01	968	9608	0.23
ShahrakeCheshmeh	9974	71.76	40.38	0.02	635.57	9946	71.71	39.45	10.01	464.78	28	0.00
Total	**1006098**	**88.41**	**90.86**	**-690.5**	**9999**	**951809**	**89.04**	**64.36**	**10**	**2430**	**54279**	**0.05**
Variance	8257.009					4141.932						
Skewness	28.85					8.9						
Kurtosis	2095.6					458.28						

a Number of detected unacceptable outlier ; b Percent of data that detected as an unacceptable outlier in every stations; * ShahrdariMantaghe

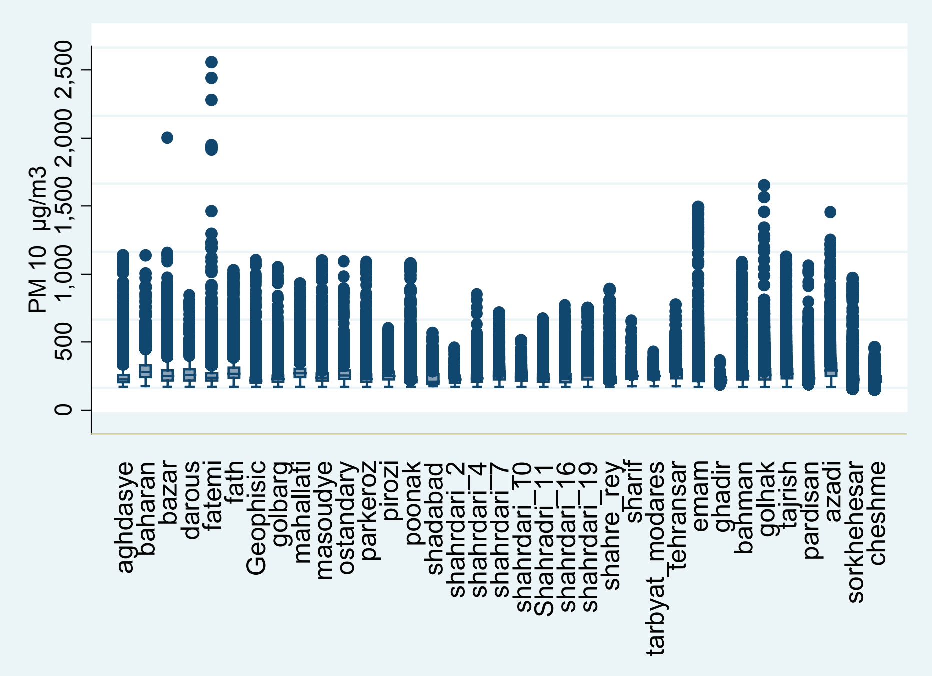

Figure 4. Boxplots per station, distribution of all measurements after implementation of algorithms.

Station Year	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013	†	‡
Aghdasieh									41.1	20.3	17.1	1.24	6.97	5.3	8.58	24.1	37.6	31.5	10	1
Fatemi	22.8	12.4	13.7	13.4	81.3	22.9	12.1	11.8	11.5	8.98	2.13	10.6	59.2						13	0.92
Bazar	28	79.2			53.2	82.2	50.6	8.38	17.9	4.12	8.55	89.3							10	0.5
Geophisics											98.3	23.3	10.2	3.79	74.2	11.2	26.7	77.6	8	0.63
Emam Khomeini										0.22	66.7		94	3.84	7.79	37.1	59.5	98.3	8	0.63
Bahman										5.57	31.4	2.43	42.9	30.8	50.2	24.9	71.1		8	0.75
Gholhak										0.71	33.4	1.28	6.22	39.3	12.7	48.3	61.9		8	0.88
Azadi											75.3	59.7	3.1	3.23	10.1	28.5	74.8		7	0.71
Park ROZ												97.9	99.1	9.89	19.8	20.6	23.3	50.8	7	0.57
Pounak												61.4	11.5	6.42	2.01	3.76	80.9	41.1	7	0.71
Share Rey											37.5	5.95	79.1	27.6	8.3	7.96	14.6	28.4	8	0.75
Pardisan										9.6	22.2	14.7			24	28.6	22.6		6	1
Golbarg													44.9	10.8	21.6	7.01	19.4	40.8	6	1
Masoudyeh													46.1	1.72	82.9	14.9	88.2	82.7	6	0.5
Tajrish										1.06	28.7	44.9	4.52	29.3	31				6	0
Sh.Mantaghe 10 *														80.5	82	33.8	39.9	93.4	5	0.4
Sh. Mantaghe 11 *														85.7	10.5	64.1	50.3	54.2	5	0.4
Sha. Mantaghe 16 *														76.4	5.55	8.21	13.4	20.3	5	0.8
Sorkhe Hesar											18.5	57.2			15.6		47.9	84.8	5	0.6
Fath															54.8	15.2	7.66	40.9	4	0.75
Mahallati															56.3	10.8	6.31	32.1	4	0.75
Sh.Mantaghe 4 *														76.1	13.6		52.8	17.2	4	0.5
Sh. Mantaghe 19 *														81.5	7.03	36.1		66	4	0.5
Shahrake Cheshmeh															63.5	96.6	39.3	86.9	4	0.25
Baharan															80.3	39.3	90.1		3	0.33
Pirozi																80.7	3.69	17.4	3	0.67
Shadabad																79.7	79.5	55.2	3	0
Sh.Mantaghe7*															94	25.5	75.1		3	0.33
Tehransar												87.3					40	35.7	3	0.67
Darous															61.1	64.5			2	1
Ostandari														54.4	23.5				2	0.5
Sanatisharif																	43	36.4	2	1
Ghadir																86.3	78.5		2	0
Sh. Mantaghe2*																		71.6	1	0
Tarbyatmodares																		50.3	1	0
†: Number of years that stations have data																			183	0.64
‡:proportion of years that missing percent were lower than 50%																				
* Shahrdari Mantaghe																				
Spectrum of missing percent of data					🟥 76 -100		🟦 50 -75		🟨 26 -50		🟩 0 -25									

Figure 5. The percentage of annual missing data for all stations in all years.

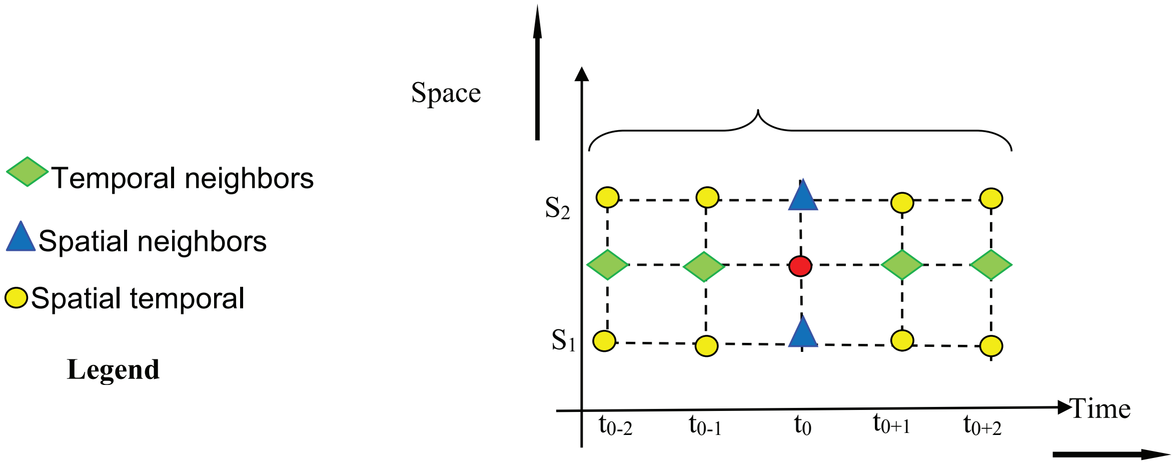

Figure 6. Spatial and temporal outliers - definition of neighborhood domain

Where σ is standard deviation of neighbors.

Finally, for detecting spatiotemporal outlier values, the algorithm searches the data and finds whether they are out of UCL and LCL limit values.

Equation (3):

$$UCL < \theta < LCL$$

Then, the algorithm detects unacceptable outlier values and transforms them to missing values. This process was done for all measurements in the dataset.

Dust storm's data

Dust storm is a natural phenomenon that occurs mostly in dry and bare lands. Recently, the amount of dust coming from Arabian countries to Iran has increased. Dust storms especially affect western and central parts of Iran.[18,19] On dust storm days, PM_{10} concentration increases and leads to large values of PM_{10}. In our study, we evaluated the effect of this phenomenon in outlier detection phase as follows:

Considering the data obtained from April to September, we detected the time points when PM_{10} value was above 500 μg/m^3; if two hours after the detected time point, PM_{10} value decreased to below 150 μg/m^3, we considered the large value as unacceptable and converted it to missing data, but if it did not reach below 150μg/m^3 after two hours, that large value was kept as PM_{10} pollution due to dust storm.

We selected 500 μg/m^3 as the cut-off point, because, according to the results of previous studies published until now, 500 μg/m^3 is the maximum PM_{10} concentration that has been reported during dusty days in Tehran. The interval from April to September was considered as the desired time of study because normally dust storm phenomenon occurs during this period and leads to large PM_{10} values in Tehran.[20,21] If PM_{10} values were above 500 µg/m^3 between October and March, we considered the large values as missing.

Plausibility range

Following the advice of a group of experts and the literature,[7] plausible concentrations of PM_{10} in different cities such as Tehran were determined and values below 10 µg/m^3 and greater than 5000 µg/m^3 were converted to missing data, because such values are not in a plausible range and the occurrence of such PM_{10} values is impossible in Tehran. We used R and STATA software for all data management including preparing data sets, performing descriptive analyses and creating spatiotemporal neighborhood tool.

Results

Using the data obtained from the two agencies (AQCC & DoE), the operation times for each station are presented in Table 1; it can be observed that operation time varies between stations. As shown, we obtained data from 39 stations. Moreover, there was no data about PM_{10} pollutant in four stations (Ghaem, Elmosanat, ShahidBeheshti, and Shadabad).

In Figure 3, a box plot is given for every measurement station considered. As illustrated in Figure 3, PM_{10} concentration ranges from negative values to 9999 and data values of DoE are more diverse than data of AQCC. In the dataset, 21553 point of data was zero. The result of initial descriptive analysis is presented in the first part of Table 2, in which the total hourly (Mean ± SD) concentration of PM_{10} determined across all available data (1006098 hour) was 88.41 ± 90.86 µg/m^3. As shown in Table 2, in seven stations (four from DoE and three from AQCC), standard deviation is larger than mean and the majority of negative values are also reported from DoE stations.

In our study, missing pattern was not significantly different by day, month or year. However, there was a small difference between seasons and there was a large proportion of missing data in winter (45.1%), compared to summer (41.6%), spring (38.0%), and autumn (41.5%). Detailed information about the proportion of missing data in every station/year is presented in Figure 5.

In total, 54,279 (5.7%) data points were recognized as unacceptable outliers, from which 37,600 (69%) data points were detected by spatiotemporal screening tool and 632 (1%) data points were detected via dust storm algorithm. Also, 16047 (29%) of data were not in the plausible range of values (below plausible lower level 10 µg/m^3) and were detected via the implementation of plausible range program and converted to missing data. Table 2 presents the detailed information about descriptive statistics of PM_{10} before and after the implementation of all algorithms and the number and proportion of data that were detected as outlier in each station. In brief, SorkheHesar station had the maximum outliers with 9608 data points (23% of station data) and Ghadir had the minimum outliers with 10 cases. The number of outliers in DoE-related stations was significantly more than that in AQCC stations. According to the results presented in Table 2, standard deviation has decreased significantly (32 %). Also, mean of measurements before the implementation of algorithms (88.41 ± 90.86 µg/m^3) was significantly different from (88.96 ± 61.64 µg/m^3) the mean after implementation of algorithms ($P < 0.001$). Figure 4 shows the detailed changes that occurred after implementation of all outlier algorithms.

Discussion

It is obvious that the results of every study depend on the quality of the used input data. Undoubtedly, the data with low accuracy lead to biased or failed outputs and may produce unreliable scientific information. We made a significant effort to create a systematic approach to deal with the missing data and the outlier values in the air pollution data of Tehran AQMN.

Missing data is a common problem which affects large databases. The pattern and proportion of the missing data are two aspects that should be considered in any analysis. In the study of Tehran air pollution data, we identified that some stations had substantial amounts of missing data. Many studies so far have investigated the effects of air pollution on various health outcomes using these datasets.[22–24] However, the majority of these studies did not mention any clear protocol for proper assessment of the missing data and discarded the time points that contained missing values. This method may be efficient only when the data contain a relatively small number of missing values.[25] Our findings suggest that such studies should place greater attention on this substantial proportion of missing data and outlier values. To solve this problem and obtain complete data series, strategies such as interpolation of the missing values, modeling, multiple imputation techniques, or satellite-derived aerosol optical depth values might be used. The second subject is the pattern of missing data. The non-random or systematic missing of the exposure data may lead to underestimation of attributable health effects.[26] However, in our study, the missing data was scattered with a random pattern but we should mention that the important concern was the volume of the missing data.

The well-known methods for detection of outlier values are distribution based, cluster-based, depth-based, distance based, and density-based methods[27] Recently, some of these methods which cover spatial dimensions of the databases have been applied on large databases.[28] As a drawback for many of these methods, they cannot consider temporal dependence structure within the data.

In the Tehran air quality data, many unacceptable large values were detected and the standard deviation was significantly decreased by implementation of our three-step process. Hence, we may conclude our method worked well for the Tehran air quality dataset. Methods such as Manhattan Distance Technique (MDT),[29] multi-scale wavelet transform and explorative moving window exists for detection of outlier values.[30] Application of these evaluation techniques on our data may introduce them as potential candidates for future works to better assess the performance of our approach.

One of the interesting findings of this data exploration was the significant difference between the quality of the AQCC data and the DoE data. As presented in the results through tables and figures, the majority of exceptionally large and negative values were observed in the obtained data of DoE and it might be said that QA/QC procedures of the DoE may not work well. The large proportion of the DoE data was detected as outlier values or noisy data.

The results of another study, which checked the PM_{10} concentrations in a short time interval at each station using case and control instruments simultaneously, showed that the air quality monitoring network used for PM_{10} measurements in Tehran has not been adequately valid. Indeed, they have reported poor correlation between the measured data of the AQMN and the measured data by the investigators in Tehran.[31] As another advantage of our algorithm, we took dust storm into account while detecting abnormal large values, because every large value is not an outlier; hence, if large values would be removed without considering the possibility of dust storm occurrence, it might result in underestimation of the health effects of the pollutants. The results of our study also provided an inclusive picture about the availability and quality of the data for future works. Our results showed that prior to the analysis on air quality data, extensive data exploration and cleaning is essential to yield intelligent results.

In this study, we focused on only one single pollutant. Also, we did not use any covariate information. Our method might be improved by using these data but it may cause to complicate the model and reduce its applicability.

Data preparation and exploration is one of the most important steps prior to using them.

Our study evaluated the air quality data of Tehran AQMN and presented options to resolve their problems. Our study results can be used as a good guide for next full data mining analysis. In many developing countries, air quality management system is likely to receive lower priority than other programs and may be poorly funded. As a result, their data quality might not be desirable. Thus, our proposed framework can be a very helpful guide for researchers who work in developing countries.

Environmental spatio-temporal databases are growing very rapidly globally. Considering their important effect on health and policy, there is high need for proper extensive data exploration and preparation in these spatio-temporal databases. We proposed a consolidated methodology to explore and discover hidden patterns of noisy data in the large databases. The implementation of our framework on the Tehran AQMN data had interesting results. Our novel proposed methods might be a very helpful guide for those researchers who work on large spatio-temporal databases in developing countries.

Competing interests

All of the authors declare that they have no actual or potential personal or financial competing interests.

Authors' contributions

General designing of the paper was by Farshad Farzadfar, Masud Yunesian, and Mansour Shamsipour. The primary draft was prepared by Mansour Shamsipour and revised by all co-authors. All authors have given approval to the final version of the manuscript.

Acknowledgments

This study was part of a PhD dissertation completed by Mansour Shamsipour in Epidemiology and has been supported by Tehran University of Medical Sciences, Institute for Environmental Research (IER) and Setad-e-Ejraie Farmane Imam. We also thank Air Quality Control Company and Department of Environment for providing us with the necessary data. Noteworthy, we extend our unlimited gratitude to Ali Sheidaei, Sahar Saeedi Moghaddamand, and Ehsan Rezaei for their help and guidance during programming statistical functions in R. Finally, we thank Non Communicable Diseases Research Center's staff for their wholehearted cooperation.

References

1. Liu GJ, Fu EJ, Wang YJ, Zhang KF, Han BP, ARROWSMITH C. A framework of environmental modelling and information sharing for urban air pollution control and management. *J China University of Mining and Technology.* 2007; **17:** 172 – 178.
2. Prüss Üstün A, Corvalán C, How much disease burden can be prevented by environmental interventions? *Epidemiology.* 2007; **18:** 167 – 178.
3. Amini H, Shamsipour M, Sowlat MH, Parsaeian M, Kasaeian A, Hassanvand MS, et al. National and sub-national Environmental Burden of Disease in Iran from 1990 to 2013-study profile. *Arch Iran Med.* 2014; **17(1):** 62 – 70.
4. Farzadfar F, Delavari A, Malekzadeh R, Mesdaghinia A, Jamshidi HR, Sayyari A, et al. NASBOD 2013: design, definitions, and metrics. *Arch Iran Med.* 2014; **17(1):** 7 – 15.
5. Chen H, Goldberg M, Villeneuve P. A systematic review of the relation between long-term exposure to ambient air pollution and chronic diseases. *Rev Environ Health.* 2008; **23:** 243 – 297
6. Kunzli N, Kaiser R, Medina S, Studnicka M, Chanel O, Filliger P, et al. Public-health impact of outdoor and traffic-related air pollution: a European assessment. *Lancet.* 2000; **356:** 795 – 801.
7. Ostro B. Outdoor air pollution. *WHO Environmental Burden of Disease Series.* 2004.
8. Begum BA, Biswas SK, Hopke PK. Assessment of trends and present ambient concentrations of PM2. 2 and PM10 in Dhaka, Bangladesh. *Air Qual Atmos.* 2008; **1:** 125 – 133.
9. Brauer M, Amann M, Burnett RT, Cohen A, Dentener F, Ezzati M, et al. Exposure assessment for estimation of the global burden of disease attributable to outdoor air pollution. *Environ Sci Technol.* 2012; **46:** 652 – 660.
10. Gharehchahi E, Mahvi AH, Amini H, Nabizadeh R, Akhlaghi AA, Shamsipour M, et al. Health impact assessment of air pollution in Shiraz, Iran: a two-part study. *J Environ Health Sci Eng.* 2013; **11:** 1 – 8.
11. Zou B, Wilson JG, Zhan FB, Zeng Y. Air pollution exposure assessment methods utilized in epidemiological studies. *J Environ Monit.* 2009; **11:** 475 – 490.
12. Jerrett M, Arain A, Kanaroglou P, Beckerman B, Potoglou D, Sahsuvaroglu T, et al. A review and evaluation of intraurban air pollution exposure models. *J Expo Sci Environ Epidemiol.* 2004; **15:** 185 – 204.
13. Kloog I, Koutrakis P, CoullBA, Lee HJ, Schwartz J. Assessing temporally and spatially resolved PM2.5 exposures for epidemiological studies using satellite aerosol optical depth measurements. *Atmos. Environ.* 2011; **45:** 6267 – 6275.
14. Kloog I, Nordio F, Coull BA, Schwartz J. Incorporating local land use regression and satellite aerosol optical depth in a hybrid model of spatiotemporal PM2.5 exposures in the Mid-Atlantic states. *Environ sci technol.* 2012; **46:** 11913 – 11921.
15. Naddafi K, Sowlat M, Safari M. Integrated assessment of air pollution in tehran, over the period from september 2008 to september 2009. *Iran J Public Health.* 2012; **41:** 77 – 86.
16. Amini H, Taghavi-Shahri SM, Henderson SB, Naddafi K, Nabizadeh R, Yunesian M. Land use regression models to estimate the annual and seasonal spatial variability of sulfur dioxide and particulate matter in Tehran, Iran. *Sci Total Environ.* 2014; **488-489:** 343 – 353.
17. Heydarpour P, Amini H, Khoshkish S, Seidkhani H, Sahraian MA, Yunesian M. Potential impact of air pollution on multiple sclerosis in Tehran, Iran. *Neuroepidemilogy. DOI:10.1159/000368553*
18. Ashrafi K, Shafiepour-Motlagh M, Aslemand A, Ghader S. Dust storm simulation over Iran using HYSPLIT. *J Environ Health Sci Eng.* 2014; **12:** 9.
19. Nazari Samani A, Dadfar S, Shahbazi A. A Study on Dust Storms Using Wind Rose, Storm Rose and Sand Rose (Case Study: Tehran Province). *Desert.* 2013; **18:** 9 – 18.
20. Givehchi R, Arhami M, Tajrishy M. Contribution of the Middle Eastern dust source areas to PM10 levels in urban receptors: Case study of Tehran, Iran. *Atmos. Environ.* 2013; **75:** 287 – 295.
21. Hassanvand MS, Naddafi K, Faridi S, Arhami M, Nabizadeh R, Sow-

lat MH, et al. Indoor/outdoor relationships of PM10, PM2.5, and PM1 mass concentrations and their water-soluble ions in a retirement home and a school dormitory. *Atmos. Environ.* 2014; **82:** 375 – 382.

22. Brajer V, Hall J, Rahmatian M. Air pollution, its mortality risk, and economic impacts in tehran, iran. *Iran J Public Health.* 2012; **41:** 31 – 38.
23. Gholizadeh M, Farajzadeh M, Darand M. The correlation between air pollution and human mortality in Tehran. *Hakim Research Journal.* 2009; **12:** 65 – 71.
24. Moridi M, Ziaei S, Kazemnejad A, The Association between Ambient Particulate Matters Pollutant and Spontaneous Abortion of the First Trimester of Pregnancy in Tehran. *Armaghane-danesh, Journal of Yasuj University of Medical Sciences.* 1390; **16:** 381 – 390.
25. Junninen H, NiskaH, Tuppurainen K, Ruuskanen J, Kolehmainen M. Methods for imputation of missing values in air quality data sets. *Atmos. Environ.* 2004; **38:** 2895 – 2907.
26. Samoli E, Peng RD, Ramsay T, Touloumi G, Dominici F, Atkinson RW, et al. What is the impact of systematically missing exposure data on air pollution health effect estimates? *Air Qual Atmos Health.* 2014: 1 – 6.
27. Kut A, Birant D. Spatio-temporal outlier detection in large databases. *CIT J Comput Inf Technol.* 2006; **14:** 291 – 297.
28. Kovács L, Vass D, Vidács A . In Improving quality of service parameter prediction with preliminary outlier detection and elimination, Proceedings of the second international workshop on inter-domain performance and simulation (IPS 2004), Budapest, 2004; 2004; 194 – 199.
29. Bakar ZA, Mohemad R, Ahmad A, Deris MM. A comparative study for outlier detection techniques in data mining, Cybernetics and Intelligent Systems, 2006 IEEE Conference on, 2006; IEEE: 2006; 1 – 6.
30. Li ST, Shue LY. Data mining to aid policy making in air pollution management. *Expert Syst Appl.* 2004; **27:** 331 – 340.
31. Goudarzi G, Naddafi K, Jonidi Jafari A. Yunesian M, Nabizadeh R, Jabbari H, Data Validity of PM_{10} Concentration Resulting from Air Quality Monitoring Network in Tehran. *World Appl Sci J.* 2009; **7:** 239 – 244.