

1 **Title: Clustering the concentrations of PM₁₀ and O₃: Application of spatio-temporal model-**
2 **based clustering**

3 **Corresponding author:**

4 **Parvin Sarbakhsh**, PhD, Associate Professor of Biostatistics.

5 Health and Environment Research Center, Department of Statistics and Epidemiology, Faculty of
6 Health, Tabriz University of Medical Sciences, Tabriz, Iran

7 Email: p.sarbakhsh@gmail.com

8 Golgasht St, Attar Neyshabori St, Tabriz, Islamic Republic of IRAN. Tel: 041-33355781 (392)

9 **Authors' name:**

10 1. **Parisa Saeipourdizaj**, MSc, Department of Statistics and Epidemiology, Faculty of Health,
11 Tabriz University of Medical Sciences, Tabriz, Iran.

12 Email: s_parisa72@yahoo.com

13 2. **Saeed Musavi**, PhD, Assistant Professor of Biostatistics.

14 Department of Statistics and Epidemiology, Faculty of Health, Tabriz University of Medical
15 Sciences, Tabriz, Iran.

16 Email: saeedmusavi@ymail.com

17 3. **Akbar Gholampour**, PhD, Associate Professor of Environmental Health Engineering.

18 Health and Environment Research Center, Department of Environmental Health Engineering,
19 School of Public Health, Tabriz university of medical sciences, Tabriz, Iran.

20 E-mail: Gholampoura@tbzmed.ac.ir

21 4. **Parvin Sarbakhsh**, PhD, Associate Professor of Biostatistics.

22 Health and Environment Research Center, Department of Statistics and Epidemiology, Faculty of
23 Health, Tabriz University of Medical Sciences, Tabriz, Iran

24 Email: p.sarbakhsh@gmail.com

25

26

27

28

29

30

31

32

33 **Abstract**

34 Air pollution data are large-scale dataset which can be analyzed in low scales by clustering to recognize the
35 pattern of pollution and have simpler and more comprehensible interpretation. So, this study aims to cluster
36 the days of year 2017 according to the hourly O_3 and PM_{10} amounts collected from four stations of Tabriz
37 by using spatio-temporal mixture model-based clustering (STMC). Besides, mixture model-based
38 clustering with temporal dimension (TMC) and mixture model-based clustering without considering spatio-
39 temporal dimensions (MC) were utilized to compare with STMC. To evaluate the efficiency of these three
40 models and obtain the optimal number of clusters in each model, BIC and ICL criteria were used. According
41 to BIC and ICL, STMC outperforms TMC and MC. Three clusters for O_3 and four clusters for PM_{10} were
42 selected as the optimal number of clusters to fit STMC models. Regarding PM_{10} , the average concentration
43 was the highest in cluster 4. Regarding O_3 , all summer days were in cluster 3 and the average concentration
44 of this cluster was the highest. Cluster 2 had the lowest concentration with a high difference from clusters
45 1 and 3 and its average temperature was the lowest. Autumn days make up about 84% of this cluster. The
46 clustering of polluted and clean days into separate groups and observing the effect of meteorological factors
47 on the amount of concentration in each cluster clearly prove the efficiency of the model. Results of STMC
48 showed that efficiency of clustering in air pollution data increases by considering both spatio-temporal
49 dimensions.

50 **Key words:** Air quality, Spatial, Temporal, Meteorological factor, Mixture, BIC.

51 **1. Introduction**

52 Today, air pollution is one of the major environmental problems in the world [1]. The presence of various
53 sources of pollution and airborne particles causes the millions of people's death throughout the world every
54 year. According to the World Health Organization (WHO), about 92% of the world's population live in the
55 polluted areas, and of the total deaths in the world, 11.6% of them are due to air pollution exposure [2].

56 Air pollution amount is intensified by some elements; Urban factors play a major role in aggravating
57 the circumstances by a lack of strategic planning in the urbanization issues, using the private cars instead
58 of public transportation, and developing various industrial areas around cities [3, 4]. Moreover,
59 meteorological factors affect air quality, e.g., temperature, speed and direction of wind, relative humidity,
60 and rainfall [4–6]. As a citation, several studies have reported the potential impact of meteorological factors
61 on the ambient air quality [4, 7–10].

62 Considering that the world is currently dealing with air pollution issues, Iran, as a developing country,
63 is no exception. As a result, air pollution killed thousands of Iranians in 2017 [11]. According to the
64 conducted studies, Tabriz has been considered among the most polluted cities of Iran [2, 12, 13].

65 The Air Quality Monitoring Stations (AQMS) of Tabriz as a part of Environmental Protection
66 Organization has set up 9 stations (spatial) in Tabriz. These stations record the air pollutant concentrations
67 on an hourly basis (temporal) resulting in collecting the huge amounts of spatio-temporal data. This
68 recorded information should be analyzed to monitor and control the pollution level and finally take
69 appropriate actions. Due to the complexity of voluminous data with spatiotemporal dependencies, the
70 traditional methods are unable to properly process and analyze them. Therefore, appropriate data mining
71 approaches are needed to analyze this heterogeneous data and access to meaningful information.

72 Clustering, as one of the most important statistical methods in data mining, analyzes the massive datasets
73 such as air pollutant concentrations. Clustering is defined as the grouping of objects in datasets with similar
74 properties. In other words, it is a function on data separating observations in terms of similarity into subsets
75 and identifying the pattern of data. Due to clustering's feature reducing the size of data, obtained clusters
76 can be easily analyzed and interpreted. Therefore, in air pollution data where the size is large, clustering by
77 reducing its size can recognize the pattern and give simpler and more comprehensible interpretation.
78 Obtained clusters of air pollution data can be investigated according to the effects of meteorological factors
79 on the pollutant concentrations within each cluster.

80 Many studies have used clustering techniques to analyze spatial and temporal data. Some have used
81 cluster analysis without considering the temporal dependency; they clustered the observed sites to identify
82 the spatial pattern [14–17]. Some have considered the data as a set of time series in different places [14, 16,
83 18, 19], which in fact did not consider the spatial nature of data. Some of them have also done clustering
84 regardless of spatial and temporal nature of data [20–22]. Due to considering only one dimension (time or
85 place) or none of them, models appeared to have insufficient accuracy and validity.

86 Recently, a new model-based clustering method for spatio-temporal data has been introduced by Cheam
87 [1] which considers both spatial and temporal dimensions for fitting a mixture model for spatio-temporal
88 data. In this method, variations in time and space were considered simultaneously, so we can cluster the
89 days of the year according to the spatio-temporal information of pollutant concentrations.

90 Many studies have been conducted in Tabriz to investigate the level of air pollution concentration and
91 its relationship with meteorological factors [23–26]. However, we did not find any study that clusters the
92 pollutant concentrations according to the temporal and spatial dimensions and justifies extracted clusters
93 with meteorological factors in Tabriz so far.

94 Therefore, the purpose of this study was to cluster the days of 2017 according to the hourly
95 concentrations of O_3 and PM_{10} collected from four stations of Tabriz using the mixture model-based
96 clustering, for spatio-temporal data, and assessing association of meteorological factors and concentration
97 within detected clusters.

98 **2. Method**

99 2.1 Data

100 Tabriz, as one of the largest and industrial cities, is the capital of East Azerbaijan in Iran. According to the
101 last census in 2017, population of Tabriz is approximately 1.8 million and its area is 320 square kilometers
102 [24]. Tabriz has four seasons and is semi-arid in terms of climate, so that it is rainy and moderate in spring,
103 hot and dry in summer, rainy in autumn, cold and snowy in winter. On hot summer days, the ambient

104 temperature reaches 30-35°C and on winter days, it reaches -10 to -20°C [27]. Additionally, the annual
105 wind speed is about 1.65 meters per second [24, 28]. According to the reports of the Meteorological
106 Organization of Tabriz in 2017, the average temperature was about 13.32°C with minimum and maximum
107 (-5, 39.40°C), average relative humidity 20.32% and average wind speed 7.02 kilometers per hour.

108 Based on the information received from AQMS, the recorded hourly data on the pollutant concentrations
109 in 2017 were more accurate and complete than other recent years. Therefore, the pollutant information
110 recorded in 2017 was selected to assess and analyze.

111 Concentration Information of PM₁₀ (Particulate matter with an aerodynamic diameter smaller than 10
112 µm) and O₃ (surface ozone) were measured by beta attenuation and UV-spectrophotometry methods at each
113 station, respectively. Table 1 shows geographical coordinates of four quality monitoring stations in Tabriz
114 (see Table 1).

115 Air quality information usually contains inaccurate, missing and outlier data. Thus, the available
116 information was examined in terms of the presence of outlier and missing data. The outlier and inaccurate
117 data were removed by z-score method before imputing the missing data, estimating the parameters and data
118 mining [2, 29, 30]. Thus, the hourly concentrations of monitoring stations were compared with the time
119 trend data of the same stations and information of neighboring monitoring stations. First, the original data
120 were converted into z-scores (with mean=0 and SD=1) based on the following settings: (1) having $|z| >$
121 $4(|z_t > 4|)$, (2) the difference from the prior value being greater than 9 ($z_t - z_{t-1} > 9$), (3) the ratio of the
122 z-score value to its centered moving average of order 3(MA3) being greater than 2 ($z_t/MA3(z_t) > 2$) and
123 (4) the difference between the singular monitoring air quality station and the prior value ($z_t - z_{t-1}$) being
124 at least twice greater than the difference between the city's monitoring air quality station's averaged
125 increment ($city(z_t - z_{t-1})$) and the prior value $[(z_t - z_{t-1})/city(z_t - z_{t-1})] > 2$, and then the outlier
126 data were deleted from the original data.

127 According to WHO guidelines, there must be at least 50% valid information of one year and the desired
128 station for statistical analysis, so the information for January to December 2017, which had at least 50% of
129 the complete data, was selected (250 days). According to the instructions mentioned for the valid data in
130 AQMS, O₃ and PM₁₀ data of four monitoring stations were considered in this research (Abrasan,
131 Baghshamal, Rahahan and Rastekoche). It should be noted that only the hourly information of 4 out of 9
132 stations was available. Of the total data available for pollutants (PM₁₀, O₃), the percentage of missing data
133 was (11.88%, 0.6%) for Abrasan, (3.3%, 3.22%) for Baghshamal, (2.22%, 1.68%) for Rahahan and (2.38%,
134 1.58%) for Rastekoche. To increase the accuracy and validity of the results, the missing data were imputed
135 using linear interpolation method [31] with R (4.0.2) software (package: imputeTS version 3.1). In order to
136 calculate the diurnal concentration of each pollutant (PM₁₀, O₃), hourly concentrations are considered. Then,
137 the obtained diurnal data are used to evaluate the air pollutant variations in each cluster. Meteorological
138 factors for Tabriz in 2017 were collected from <https://en.tutiempo.net>.

139 2.2 Statistical analysis

140 The clustering of the spatio-temporal air quality data was performed by fitting a mixture model considering
141 the spatio-temporal dimensions (STMC). In order to fit the mentioned model, spatial (geographical
142 coordinates of data recording stations) and temporal (day and hour of the recorded data) information is
143 considered, optimal number of clusters was determined according to the Bayesian Information Criterion
144 (BIC) [32] and Integrated Completed Likelihood criterion (ICL) [33], and the parameters of the model were
145 estimated using the EM algorithm. Finally, each observation was assigned to an appropriate cluster.

146 In order to assess the performance of STMC, it was compared with temporal mixture model-based
147 clustering considering only the temporal dimension (TMC), and mixture model-based clustering without
148 considering the spatio-temporal dimensions (MC). The criteria used for comparing the fitted models are
149 BIC and ICL. Fitting the TMC and MC models were performed at the same way as STMC, except a
150 difference in considering the dimensions.

151 The nonparametric Mann-Kendall test (MK) [2, 34–36] was calculated to investigate the relationship
 152 between meteorological factors and pollutant concentrations (PM₁₀ and O₃) within the detected clusters.

153 R (4.0.2) software with “SpaTimeClus” (version 1.0) and “mclust” (version 5.4.6) packages were used for
 154 statistical analysis. The following is a brief explanation on STMC.

155 2.2.1 Mixture models for spatio-temporal data

156 Mixture model-based clustering is a broad family of algorithms designed for modelling an unknown
 157 distribution as a mixture of simpler distributions, sometimes called basis distribution [37, 38]. In this
 158 method, unlike other clustering methods that cluster the data based on some similarity measures, a Gaussian
 159 statistical distribution was considered for data. Thus, the purpose of model-based clustering is to estimate
 160 the parameters of the statistical distribution and hidden variables, considered as a cluster label of the data
 161 [37, 39]. Moreover, this model can be applied for different types of data such as continuous, ordinal,
 162 categorical, mixed and functional.

163 Extension of mixture model-based clustering for spatio-temporal data was introduced in 2016 by
 164 Cheam [1], which is a generalization for mixture model-based clustering considering only the temporal
 165 dimension [40]. Besides, the STMC method considers the variations in space and time simultaneously.

166 Suppose each observation $x_i = \{x_{ijt}\}_{j=1,\dots,J}^{t=1,\dots,T}$, as spatio-temporal data, including $J \times T$ observations on
 167 predetermined temporal framework $m = (m_1, \dots, m_T)$ and the spatial framework $s = (s_1, \dots, s_J)$.
 168 Therefore, $x_{ijt} \in \mathbb{R}$ is the observation i at location j and time t . For spatial coordinates, location j is a two-
 169 dimensional vector $s_j = (v_j, w_j)$.

170 The density function of the c^{th} element (or cluster) for the spatio-temporal model is as follows:

$$171 f_c(x_i | \mathcal{G}_c) = \prod_{j=1}^J \prod_{t=1}^T \sum_{h=1}^H \gamma_{cjth}(\alpha_c) \times \varphi(x_{ijt} | [M_t - M_{t-1}]' \beta_{ch} + x_{ij(t-1)}, \sigma_{ch}^2), \quad (1)$$

172 so that $\mathcal{G}_c = (\alpha_c, \beta_{ch}, \sigma_{ch}^2, h = 1, \dots, H)$ is a set of parameters of the element c^{th} , $M_t = (1, m_t, \dots, m_t^Q)$

173 the polynomial vector with Q -degree of M_t , $\beta_{ch} = (\beta_{ch0}, \dots, \beta_{chQ})'$ is the coefficient vector of the h^{th}

174 regression model for the c^{th} element, and $\phi(\cdot | \mu, \sigma^2)$ is the univariate Gaussian distribution density with
 175 mean μ and variance σ^2 . The weight of this mixture, γ_{cjh} , depends on both dimensions (space and time) via
 176 a logistic function. In Expectation and Maximization steps of EM algorithm, the parameters of the model,
 177 which includes $\mathcal{G}_c = (\alpha_c, \beta_{ch}, \sigma_{ch}^2, h = 1, \dots, 5)$, were estimated. In this study, each x_{ijt} represents the
 178 hourly concentration of desired pollutant, e.i., observation i in time t (T=24 hour per day for n=250 days)
 179 and location j (J=4 stations).

180 2.2.2 Model selection

181 Suppose M represents a set of selectable models with the optimal number of clusters. Each model $\mathcal{M} \in M$,
 182 is defined by three elements. C: number of elements (number of clusters), H: number of regressions of each
 183 element, and Q: degree of polynomial regression. Therefore, for model M we have:

$$184 \mathcal{M} = (C, H, Q) \quad ; C, H, Q \in \mathbb{N}^*. \quad (2)$$

185 So, M is obtained by applying the maximum number of each element. Thus, assuming $C_{\min} = H_{\min} =$
 186 $Q_{\min} = 1$, the number of selectable models is as follows:

$$187 \text{models}(M) = C_{\max} \times H_{\max} \times Q_{\max}. \quad (3)$$

188 The selection of the best model based on BIC and ICL criteria, calculated for all models of M . Therefore,
 189 in this study, suitable elements for selecting the best models in STMC, TMC and MC, based on the
 190 mentioned criteria, is obtained. Finally, each model with suitable number of clusters is analyzed.

191 3. Results

192 Table 2 presents the information criteria and the number of free parameters for O_3 and PM_{10} concentrations
 193 in all models (see Table 2). BIC for STMC was the highest in both pollutants. The number of free
 194 parameters of STMC in both pollutants was less than the other models. Since STMC had better fitting on
 195 O_3 and PM_{10} concentrations which is spatio-temporal data, it was selected as the final model. The number
 196 of optimal clusters in STMC was three for O_3 and four for PM_{10} .

197 3.1 Description of clusters obtained from STMC

198 Figure 1 shows the scatter plot of the diurnal average concentrations of PM₁₀ (right) and O₃ (left), where
199 each representative dot corresponds to the day profile of pollutant measurements for a cluster (see Figure
200 1). Also, Table 3 presents descriptive information on the detected clusters and meteorological factors (see
201 Table 3). It is important to mention that, the number of available days for analyzing the concentration
202 variations of pollutants in each season was: spring 93 days, summer 31 days, autumn 90 days, and winter
203 36 days.

204 3.1.1 Description of clusters obtained for O₃ pollutant

205 In clustering O₃ concentrations based on BIC and ICL criteria, STMC with three clusters was selected as
206 the best model. O₃ contains three clusters with average concentrations of low (cluster2), moderate
207 (cluster1), and high (cluster3). O₃-cluster1 encompasses 102 days which is the most member days.

208 Spring forms approximately 60% of this cluster and the rest of the members are in O₃-cluster3, whereas
209 average concentration for spring in O₃-cluster3 ($65.51 \mu\text{g}/\text{m}^3$) is higher than O₃-cluster1 (55.87
210 $\mu\text{g}/\text{m}^3$). O₃-cluster2 includes approximately 85% of autumn days (59 days) and the rest of the members
211 are in O₃-cluster1. Although O₃-cluster1 encompasses less number of autumn days (22 days), the average
212 concentration of O₃-cluster1 ($48.39 \mu\text{g}/\text{m}^3$) members is higher than O₃-cluster2 ($18.68 \mu\text{g}/\text{m}^3$)
213 members. Additionally, most of winter days (22 days) are in O₃-cluster1.

214 3.1.2 Description of clusters obtained for PM₁₀ pollutant

215 In clustering PM₁₀ concentrations based on BIC and ICL criteria, STMC with four clusters was selected
216 as the best model. PM₁₀-cluster4 has the highest number of days (90 days) and highest average concentration
217 ($78.77 \mu\text{g}/\text{m}^3$). PM₁₀-cluster1 and PM₁₀-cluster2 were almost the same in member days (70 and 69 days,
218 respectively), whereas the average concentration of PM₁₀-cluster1 ($49.69 \mu\text{g}/\text{m}^3$) was higher than the
219 average concentration of PM₁₀-cluster2 ($43.81 \mu\text{g}/\text{m}^3$). PM₁₀-cluster3 ($41.87 \mu\text{g}/\text{m}^3$) contained the
220 lowest member days (21 days) and its average concentration was close to PM₁₀-cluster2.

221 Spring days were present in all clusters and this season's the highest and lowest average concentrations
222 belonged to PM₁₀-cluster4 and PM₁₀-cluster2, respectively. It is important to note that, regarding the spring,
223 PM₁₀-cluster1 and PM₁₀-cluster3 had moderate average concentrations and were close in values. However,
224 PM₁₀-cluster3 had slightly lower average concentration than PM₁₀-cluster1. the summer days were present
225 in PM₁₀-clusters: 1, 2 and 4, but PM₁₀-cluster1 included only two days which can be disregarded.
226 Additionally, the PM₁₀-cluster4 had higher average concentration than PM₁₀-cluster2. Autumn days were
227 seen in PM₁₀-cluster2 (approximately 60%) and PM₁₀-cluster4 (approximately 37%). PM₁₀-cluster4
228 containing low member days exhibited higher average concentration than PM₁₀-cluster2. In regard to
229 winter, PM₁₀-cluster1 included 50% of days and PM₁₀-cluster4 30% of them. Furthermore, PM₁₀-cluster4
230 had higher average concentration than PM₁₀-cluster1.

231 3.2 Analysis of the correlation between pollutant concentrations and meteorological factors within clusters 232 in STMC

233 The MK test to assess the correlation between meteorological factors and the concentrations of O₃ and PM₁₀
234 was calculated, shown in Table 4 (see Table 4).

235 According to the MK's test results, all clusters presented a statistically significance correlation between
236 the average concentration of O₃ and meteorological factors, except O₃-cluster2 for rainfall. The relationship
237 between the average concentration of O₃ and temperature in all clusters was positive and weak to moderate
238 with a very high statistical significance (p-value<0.000), in respect of relative humidity, it was negative and
239 moderate with a very high statistical significance (p-value<0.000). Moreover, all clusters showed a positive
240 and weak relationship between the average concentration of O₃ and wind speed.

241 The MK's test results for PM₁₀ revealed a statistically significance correlation between average
242 concentration and meteorological factors (temperature and relative humidity) in PM₁₀-cluster2 and 3. In
243 PM₁₀-cluster2, the relationship between temperature and average concentration was positive and weak to
244 moderate with a very high statistical significance (p-value<0.000), and in respect of relative humidity, it
245 was negative and weak. In PM₁₀-cluster3, the relationship between temperature and average concentration

246 was positive and moderate to strong, and in respect of relative humidity, it was negative and moderate to
247 strong. Furthermore, in PM_{10} -cluster1,2 and 3, the relationship between rainfall and average concentration
248 was negative and weak. It is important to note that there was no statistically significance correlation between
249 PM_{10} and wind speed.

250 **4. Discussion**

251 The goal of our study was to apply mixture model-based clustering method for clustering the days according
252 to hourly concentrations of PM_{10} and O_3 over the one-year period in Tabriz. This method does not consider
253 the type of data and can be used for various types such as continuous, ordinal, categorical, mixed and
254 functional. Spatial and temporal information is always available in the air pollution datasets, which is a type
255 of functional data. Therefore, mixture model-based clustering method is a proper choice for analyzing this
256 functional data with spatial and temporal dependencies.

257 To evaluate the appropriateness of statistical models for our data, STMC with spatial and temporal
258 dependencies, TMC with temporal dependency, and MC lacking in both dependencies were fitted to data.
259 BIC, ICL and a number of free parameters were calculated for these models. The results showed that the
260 STMC compared to TMC and MC had better fitting on spatio-temporal data of PM_{10} and O_3 . In the
261 following, we have a review of the various conducted studies on the pollutant clustering where none of the
262 dependencies (spatial and temporal) are simultaneously considered.

263 Jin [41] established the use of k-means clustering method to recognize O_3 spatial regimes (or cluster)
264 over San Joaquin Valley of California. Clusters demonstrate the days of similar O_3 spatial distribution. In
265 terms of concentration, of total six recognized regimes, two corresponded to low- O_3 cluster, three to
266 moderate- O_3 cluster, and one to high- O_3 cluster. Moreover, Meteorological measurements were used to
267 describe O_3 spatial distributions, and their correlation to those in San Francisco Basin.

268 Pandey [42] by using average linkage clustering method, showed the spatial and temporal variations of
269 $PM_{1.0}$, $PM_{2.5}$, PM_{10} , NO_2 and SO_2 in India. Clusters including monitoring sites represent similar behavior

270 in terms of the pollutant dispersions and spatial variations. Based on the results, in all the sites,
271 concentrations of all types of PM exhibited the highest value in the winter and the lowest one in the rainy
272 days.

273 Huang [43] by using a hierarchical clustering method, investigated the characteristics of PM_{2.5} in China.
274 According to the results, PM_{2.5} information Collected from 13 monitoring sites were arranged in 3 clusters.
275 As one of the important results, temporal distribution of PM_{2.5} revealed that the winter has the highest
276 concentration, autumn is higher than spring in concentration value and the lowest concentration belongs to
277 summer. In terms of spatial distribution, three out of 13 monitoring sites exhibited the highest concentration
278 of PM_{2.5}.

279 In a study conducted in the northern China [44], cluster analysis was used to reveal the spatial and
280 temporal distribution of PM_{2.5}, SO₂, NO₂, CO and O₃ pollutants. Pearson correlation coefficient was used
281 to investigate the relationship between pollutant concentrations with each other. In this study, hierarchical
282 cluster analysis was used to classify 9 cities in different groups based on monthly average of above-
283 mentioned pollutants in each city. According to cluster analysis, considering the monthly average of
284 pollutants, 9 cities were grouped in 4 clusters. Clustering analysis shows that air pollution is mainly
285 associated with industrial city structures along with their geographical and socioeconomic factors.

286 Many studies have been conducted on the relationship between the pollutant concentrations and
287 meteorological factors representing that the hourly, daily, monthly and seasonal variations of air pollutants
288 in a residential area can be caused by meteorological factors such as atmospheric temperature, relative
289 humidity, wind speed, solar radiation intensity and etc. By increasing or decreasing of the above-mentioned
290 factors during the seasons, the amount of pollutant concentration will be affected [2, 45, 46].

291 Among the meteorological factors, atmospheric stability and wind speed have the most influence on the
292 atmospheric dispersion and decreasing air pollution [30, 34]. The intensification of air pollution in Tabriz
293 can be attributed to the temperature inversions and calm conditions. Current situation of the city is generated
294 by the geographical characteristics, meteorological factors, and residential constructions in the direction of

295 the wind entering the city. It should be noted that based on the results, there was a fluctuated variation for
296 PM_{10} and O_3 concentrations from 2006 to 2017 in Tabriz [25].

297 Based on the results of this study, days with close values in terms of average temperature, relative
298 humidity and rainfall are placed in a cluster. O_3 -cluster3 with the highest temperature and lowest relative
299 humidity has the hottest days of the year while the coldest days belong to O_3 -cluster2 with the lowest
300 temperature and highest relative humidity. Therefore, based on the average concentration of each cluster
301 and the MK test, the effect of meteorological factors can be clearly observed, so that concentration has a
302 direct and positive relationship with temperature and has an inverse and negative relationship with relative
303 humidity and rain. In the following, the reasons for increasing and decreasing O_3 concentration in hot and
304 cold days of the year are briefly indicated.

305 based on most conducted studies, there was a clear and logical trend in the monthly and seasonal
306 variations of O_3 concentration. According to the variation pattern of O_3 concentration, the highest and lowest
307 values were observed in summer and winter, respectively. This highest value may be due to high
308 atmospheric temperature, high intensity of sunlight, long days and long sunny hours in hot seasons which
309 increase photochemical reactions and O_3 production, while the lowest value is related to reduced daylight
310 (Sunlight time), lower temperature and sunshine duration [2, 45, 47]. Various processes and activities such
311 as transport, deposition and NO_x titration also participate in O_3 formation [48]. The increasing of O_3
312 concentration in the warm seasons is directly related to the increment amount of temperature, which is one
313 of the important parameters in controlling the O_3 formation [49].

314 According to the results of this study, PM_{10} -cluster4 had the highest temperature. PM_{10} -cluster 3 and 2
315 had lower temperatures, higher rainfall, and higher wind speed. It is important to note that PM_{10} -cluster 3
316 and 2 were close in values with slight differences. In the following, the effect of meteorological factors on
317 PM_{10} concentration in different seasons is summarized.

318 PM_{10} may increase in spring and summer due to the occurrence of the Asian dust phenomenon [36, 50].
319 In addition, in hot seasons due to high atmospheric temperature and reduced relative humidity, the

320 concentration of PM_{10} increases. Furthermore, the correlation between PM_{10} concentration with temperature
321 and relative humidity is positive and negative, respectively [8].

322 The conducted studies on the relationship between PM_{10} and rainfall demonstrate that PM_{10}
323 concentration reduces in atmosphere due to the washout effects of rainfall. Eventually, rainfall and relative
324 humidity have negative correlation with PM_{10} [51].

325 According to the conducted analysis in relation to the average concentrations of O_3 and PM_{10} in each
326 cluster, the relationship between meteorological factors and average concentration in each cluster, and
327 member days of seasons in individual clusters, it can be concluded that the clustering method used in this
328 study had a good fitness and its results were interpretable. Finally, model-based clustering has a strong
329 background and shows high efficiency in analyzing the data with spatio-temporal dimensions. In datasets
330 with spatial and temporal information, using the above-mentioned method lead to more reliable and
331 accurate results [1].

332 **5. Conclusion**

333 Data mining methods are very applicable in analyzing the air pollution data and its results play a major role
334 in controlling and preventing the increment of pollutants. Clustering, as one of the most important Data
335 mining methods, reduces the amount of examined data, and reveals the hidden information of them.
336 Consequently, analyzing individual clusters including low amount of data reduces the errors of the results.
337 In this study, we fitted the STMC, a mixture model-based clustering method with considering spatio-
338 temporal dimensions, to classify the days according to their PM_{10} and O_3 concentrations in Tabriz, 2017.
339 The results showed that in our spatio-temporal air pollution data, considering the dimensions of the data in
340 analyzing could improve the performance of the clustering. The results of the evaluation of the obtained
341 clusters in terms of available meteorological factors indicated an acceptable and meaningful clustering by
342 STMC.

343 **Limitation**

344 This study may have some limitations:

- 345 • Incomplete measurements of pollutant concentrations were seen in winter and summer, over the
346 year 2017. This deficiency was highlighted in the recorded data of January, February, July and
347 September.
- 348 • Due to the different sources for pollutant production in each region, geographical conditions and
349 meteorological factors which have influence on the pollutant level variations, it was impossible to
350 directly compare with the results of the other studies.

351 **Funding**

352 This work was supported by the Tabriz University of Medical Sciences.

353 **Conflicts of interest/Competing interests**

354 The authors declare that they have no competing interests.

355 **Ethics approval**

356 We appreciate the Health and Environment Research Center because of financial support. This article has
357 been extracted from the thesis submitted for MSc degree in Biostatistics which has been approved by the
358 ethics committee of Tabriz University of Medical Sciences (Ethic number: IR.TBZMED.REC.1398.352)

359 **Availability of data and material/ Data availability**

360 The datasets analyzed during the current study are available from the corresponding author on reasonable
361 request.

362 **Code availability**

363 Not applicable

364 **Authors' contributions**

365 **Parisa saeipourdizaj (first author):** formulation and evaluation of overarching research goals and aims;
366 setting the data in software package format; application of statistical, computational, and other formal

367 techniques to analyze; Application of available software codes; Preparation (drafting, reviewing,
368 translating, and revising the paper), and presentation of the manuscript.

369 **Saeed Musavi:** Statistical analysis, manuscript preparation and reviewing the paper

370 **Akbar Gholampour:** Reviewing the paper

371 **Parvin Sarbakhsh (corresponding author):** formulation and evaluation of overarching research goals and
372 aims; Statistical analysis; Preparation (drafting, reviewing, translating, and revising the paper), and
373 presentation of the manuscript.

374 All authors have read and approved the final manuscript.

375

376 **References**

- 377 1. Cheam, A.S.M., Marbac, M., McNicholas, P.D.: Model-based clustering for spatiotemporal data
378 on air quality monitoring, (2017). <https://doi.org/10.1002/env.2437>.
- 379 2. Faridi, S., Shamsipour, M., Krzyzanowski, M., Künzli, N., Amini, H., Azimi, F., Malkawi, M.,
380 Momeniha, F., Gholampour, A., Hassanvand, M.S., Naddafi, K.: Long-term trends and health
381 impact of PM_{2.5} and O₃ in Tehran, Iran, 2006–2015. *Environmental International*. 114, 37–49
382 (2018). <https://doi.org/10.1016/j.envint.2018.02.026>.
- 383 3. Manju, A., Kalaiselvi, K., Dhananjayan, V., Palanivel, M., Banupriya, G.S., Vidhya, M.H.,
384 Panjakumar, K., Ravichandran, B.: Spatio-seasonal variation in ambient air pollutants and
385 influence of meteorological factors in Coimbatore, Southern India. *Air Quality, Atmosphere and*
386 *Health*. 11, 1179–1189 (2018). <https://doi.org/10.1007/s11869-018-0617-x>.
- 387 4. Zhang, H., Wang, Y., Hu, J., Ying, Q., Hu, X.M.: Relationships between meteorological
388 parameters and criteria air pollutants in three megacities in China. *Environmental Research*. 140,
389 242–254 (2015). <https://doi.org/10.1016/j.envres.2015.04.004>.
- 390 5. Shukla, J.B., Misra, A.K., Sundar, S., Naresh, R.: Effect of rain on removal of a gaseous pollutant
391 and two different particulate matters from the atmosphere of a city. *Mathematical and Computer*

392 Modelling. 48, 832–844 (2008).

393 6. Goyal, S.K., Rao, C.V.C.: Assessment of atmospheric assimilation potential for industrial
394 development in an urban environment: Kochi (India). *Science of the total environment*. 376, 27–
395 39 (2007).

396 7. Owoade, O. K.; Olise, F. S.; Ogundele, L.T.; Fawole, O.G., Olaniyi, H.B.: Correlation between
397 particulate matter concentrations and meteorological parameters at a site in Ile-Ife, Nigeria. *Ife*
398 *Journal of Science* no. 1 (2012). 14, 83–93 (2012).

399 8. Dominick, D., Latif, M.T., Juahir, H., Aris, A.Z., Zain, S.M.: An assessment of influence of
400 meteorological factors on PM sub (10) and NO sub (2) at selected stations in Malaysia.
401 *Sustainable Environment Research*. 22, 305–315 (2012).

402 9. Islam, M.M., Afrin, S., Ahmed, T., Ali, M.A.: Meteorological and seasonal influences in ambient
403 air quality parameters of Dhaka city. *J. Civ. Eng.* 43, 67–77 (2015).

404 10. Galindo, N., Yubero, E., Nicola, J.F., Crespo, J.: Chemical characterization of PM1 at a regional
405 background site in the western Mediterranean. *Aerosol and Air Quality Research*. 16, 530–541
406 (2015).

407 11. 2017, O.: <https://ourworldindata.org/>.

408 12. Naddafi, K., Hassanvand, M.S., Yunesian, M., Momeniha, F., Nabizadeh, R., Faridi, S.,
409 Gholampour, A.: Health impact assessment of air pollution in megacity of Tehran, Iran. *Iranian*
410 *journal of environmental health science & engineering*. 9, 28 (2012).

411 13. Hassanvand, M.S., Naddafi, K., Faridi, S., Arhami, M., Nabizadeh, R., Sowlat, M.H., Pourpak, Z.,
412 Rastkari, N., Momeniha, F., Kashani, H.: Indoor/outdoor relationships of PM10, PM2. 5, and PM1
413 mass concentrations and their water-soluble ions in a retirement home and a school dormitory.
414 *Atmospheric Environment*. 82, 375–382 (2014).

415 14. Lavecchia, C., Angelino, E., Bedogni, M., Bravetti, E., Gualdi, R., Lanzani, G., Musitelli, A.,
416 Valentini, M.: The ozone patterns in the aerological basin of Milan (Italy). *Environmental*
417 *Software*. 11, 73–80 (1996).

- 418 15. Saksena, S., Joshi, V., Patil, R.S.: Cluster analysis of Delhi's ambient air quality data. *Journal of*
419 *Environmental monitoring*. 5, 491–499 (2003).
- 420 16. Gramsch, E., Cereceda-Balic, F., Oyola, P., Von Baer, D.: Examination of pollution trends in
421 Santiago de Chile with cluster analysis of PM10 and ozone data. *Atmospheric environment*. 40,
422 5464–5475 (2006).
- 423 17. Molinari, N.: Free knot splines for supervised classification. *Journal of classification*. 24, 221–234
424 (2007).
- 425 18. Gabusi, V., Volta, M.: A methodology for seasonal photochemical model simulation assessment.
426 *International journal of environment and pollution*. 24, 11–21 (2005).
- 427 19. Morlini, I.: Searching for structure in measurements of air pollutant concentration.
428 *Environmetrics: The official journal of the International Environmetrics Society*. 18, 823–840
429 (2007).
- 430 20. Fraley, C., Raftery, A.E.: Model-based clustering , discriminant analysis , and density estimation.
431 (2002).
- 432 21. Vrbik, I., McNicholas, P.D.: Parsimonious skew mixture models for model-based clustering and
433 classification. *Computational Statistics & Data Analysis*. 71, 196–210 (2014).
- 434 22. Murphy, K., Murphy, T.B.: Parsimonious Model-Based Clustering with Covariates. arXiv preprint
435 arXiv:1711.05632. (2017).
- 436 23. Asghari, F.B., Mohammadi, A.A.: The effect of the decreasing level of Urmia Lake on particulate
437 matter trends and attributed health effects in Tabriz, Iran. *Microchemical Journal*. 104434 (2019).
438 <https://doi.org/10.1016/j.microc.2019.104434>.
- 439 24. Amini Parsa, V., Salehi, E., Yavari, A.R., van Bodegom, P.M.: Analyzing temporal changes in
440 urban forest structure and the effect on air quality improvement. *Sustainable Cities and Society*.
441 48, 101548 (2019). <https://doi.org/10.1016/j.scs.2019.101548>.
- 442 25. Barzeghar, V., Sarbakhsh, P., Hassanvand, M.S., Faridi, S., Gholampour, A.: Long-term trend of
443 ambient air PM10, PM2. 5, and O3 and their health effects in Tabriz city, Iran, during 2006–2017.

- 444 Sustainable Cities and Society. 54, 101988 (2020).
- 445 26. Yicun, G., Khorshiddoust, A.M., Mohammadi, G.H., Sadr, A.H.: The relationship between PM 2 .
446 5 concentrations and atmospheric conditions in severe and persistent urban pollution in Tabriz ,
447 northwest of Iran. (2020).
- 448 27. Azarafza, M., Ghazifard, A.: Urban geology of Tabriz City: Environmental and geological
449 constraints. *Advances in environmental research*. 5, 95–108 (2016).
450 <https://doi.org/10.12989/aer.2016.5.2.095>.
- 451 28. Kalajahi, M.J., Khazini, L., Rashidi, Y., Heris, S.Z.: Development of Reduction Scenarios Based
452 on Urban Emission Estimation and Dispersion of Exhaust Pollutants from Light Duty Public
453 Transport: Case of Tabriz, Iran. *Emission Control Science and Technology*. 1–19 (2019).
- 454 29. Barrero, M.A., G. Orza, J., Cabello, M., Cantón, L.: Categorisation of air quality monitoring
455 stations by evaluation of PM10 variability. *The Science of the total environment*. 524-525C, 225–
456 236 (2015). <https://doi.org/10.1016/j.scitotenv.2015.03.138>.
- 457 30. Song, C., He, J., Wu, L., Jin, T., Chen, X., Li, R., Ren, P., Zhang, L., Mao, H.: Health burden
458 attributable to ambient PM2.5 in China. *Environmental pollution (Barking, Essex : 1987)*. 223,
459 575–586 (2017). <https://doi.org/10.1016/j.envpol.2017.01.060>.
- 460 31. Norazian, M.N., Shukri, Y.A., Azam, R.N., Al Bakri, A.M.M.: Estimation of missing values in air
461 pollution data using single imputation techniques. *ScienceAsia*. 34, 341–345 (2008).
462 <https://doi.org/10.2306/scienceasia1513-1874.2008.34.341>.
- 463 32. Schwarz, G.: Estimating the dimension of a model. *The annals of statistics*. 6, 461–464 (1978).
- 464 33. Aike, H.A.I.: *A New Look at the Statistical Model Identification*. (1974).
- 465 34. Vuorenmaa, J., Augustaitis, A., Beudert, B., Bochenek, W., Clarke, N., de Wit, H.A., Dirnbock,
466 T., Frey, J., Hakola, H., Kleemola, S.: Long-term changes (1990-2015) in the atmospheric
467 deposition and runoff water chemistry of sulphate, inorganic nitrogen and acidity for forested
468 catchments in Europe in relation to changes in emissions and hydrometeorological conditions.
469 *Science of the total environment*. 625, 1129–1145 (2018).

- 470 35. Cerro, J.C., Cerda, V., Pey, J.: Trends of air pollution in the Western Mediterranean Basin from a
471 13-year database: A research considering regional, suburban and urban environments in Mallorca
472 (Balearic Islands). *Atmospheric Environment*. 103, 138–146 (2015).
- 473 36. Ahmed, E., Kim, K.-H., Shon, Z.-H., Song, S.-K.: Long-term trend of airborne particulate matter
474 in Seoul, Korea from 2004 to 2013. *Atmospheric Environment*. 101, 125–133 (2015).
- 475 37. McLachlan, G., Peel, D.: *Finite mixture models*, Wiley series in probability and statistics, (2000).
- 476 38. McNicholas, P.D.: *Mixture model-based classification*. Chapman and Hall/CRC (2016).
- 477 39. Mcnicholas, P.D.: *Model-Based Clustering*. 373, 331–373 (2016). <https://doi.org/10.1007/s0035>.
- 478 40. Same, A., Chamroukhi, F., Govaert, G., Aknin, P.: Model-based clustering and segmentation of
479 time series with changes in regime. *Advances in Data Analysis and Classification*. 5, 301–321
480 (2011).
- 481 41. Jin, L., Harley, R.A., Brown, N.J.: Ozone pollution regimes modeled for a summer season in
482 California’s San Joaquin Valley: A cluster analysis. *Atmospheric environment*. 45, 4707–4718
483 (2011).
- 484 42. Pandey, B., Agrawal, M., Singh, S.: Assessment of air pollution around coal mining area:
485 emphasizing on spatial distributions, seasonal variations and heavy metals, using cluster and
486 principal component analysis. *Atmospheric pollution research*. 5, 79–86 (2014).
- 487 43. Huang, P., Zhang, J., Tang, Y., Liu, L.: Spatial and temporal distribution of PM_{2.5} pollution in
488 Xi’an City, China. *International journal of environmental research and public health*. 12, 6608–
489 6625 (2015).
- 490 44. Tian, D., Fan, J., Jin, H., Mao, H., Geng, D., Hou, S., Zhang, P., Zhang, Y.: Characteristic and
491 spatiotemporal variation of air pollution in Northern China based on correlation analysis and
492 clustering analysis of five air pollutants. *Journal of Geophysical Research: Atmospheres*. 125,
493 e2019JD031931 (2020).
- 494 45. Sicard, P., Serra, R., Rossello, P.: Spatiotemporal trends in ground-level ozone concentrations and
495 metrics in France over the time period 1999-2012. *Environmental research*. 149, 122–144 (2016).

496 46. Zhao, S., Yu, Y., Yin, D., He, J., Liu, N., Qu, J., Xiao, J.: Annual and diurnal variations of gaseous
497 and particulate pollutants in 31 provincial capital cities based on in situ air quality monitoring data
498 from China National Environmental Monitoring Center. *Environment international*. 86, 92–106
499 (2016).

500 47. Carvalho, V.S.B., Freitas, E.D., Martins, L.D., Martins, J.A., Mazzoli, C.R., de Fatima Andrade,
501 M.: Air quality status and trends over the Metropolitan Area of Sao Paulo, Brazil as a result of
502 emission control policies. *Environmental Science & Policy*. 47, 68–79 (2015).

503 48. Lacressonniere, G., Foret, G., Beekmann, M., Siour, G., Engardt, M., Gauss, M., Watson, L.,
504 Andersson, C., Colette, A., Josse, B.: Impacts of regional climate change on air quality projections
505 and associated uncertainties. *Climatic Change*. 136, 309–324 (2016).

506 49. Pawlak, I., Jaros awski, J.: The influence of selected meteorological parameters on the
507 concentration of surface ozone in the central region of Poland. *Atmosphere-Ocean*. 53, 126–139
508 (2015).

509 50. Jang, E., Do, W., Park, G., Kim, M., Yoo, E.: Spatial and temporal variation of urban air pollutants
510 and their concentrations in relation to meteorological conditions at four sites in Busan, South
511 Korea. *Atmospheric Pollution Research*. 8, 89–100 (2017).

512 51. Giri, D., ADHIKARY, P.R., MURTHY, V.K.: The influence of meteorological conditions on
513 PM10 concentrations in Kathmandu Valley. (2008).

514
515
516
517
518
519
520
521

522 **Figure and Table legends**

523 Figure 1 Scatter plot of the diurnal average concentration of PM₁₀ (right) and O₃ (left) clusters in STMC

524 model

525 Table 1 Geographical coordinates of air quality monitoring station

526 Table 2 Goodness-of-fit criteria, number of free parameters and number of clusters for the analyzed
527 pollutants in each clustering model

528 Table 3 Descriptive statistics for PM₁₀ and O₃ clusters obtained from STMC

529 Table 4 Mann-Kendall correlation coefficient between meteorological parameters and PM₁₀ and O₃ clusters