Environmental Modeling & Assessment (2022) 27:45–54
https://doi.org/10.1007/s10666-021-09802-5

Clustering the Concentrations of PM_{10} and O_3: Application of Spatiotemporal Model–Based Clustering

Parisa Saeipourdizaj[1] · Saeed Musavi[1] · Akbar Gholampour[2] · Parvin Sarbakhsh[1]

Received: 27 February 2021 / Accepted: 28 October 2021 / Published online: 11 November 2021

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

Keywords Air quality · Spatial · Temporal · Meteorological factor · Mixture · BIC

1 Introduction

Nowadays, air pollution is one of the world's most serious environmental issues [1]. Every year, the presence of diverse sources of pollution and airborne particles kills millions of people around the world. According to the World Health Organization (WHO), about 92% of the world's population lives in polluted areas, and of the total deaths in the world, 11.6% of them are due to air pollution exposure [2].

The amount of air pollution is increased by several causes, including a lack of strategic planning in urbanization concerns, the use of private cars instead of public transportation, and the development of diverse industrial sectors near cities [3, 4]. Furthermore, meteorological parameters such as temperature, wind speed and direction, relative humidity, and rainfall have an impact on air quality [4–6]. As a citation, several studies have reported the potential impact of meteorological factors on the ambient air quality [4, 7–10].

Iran, as a developing country, is no exception to the world's present air pollution problems. As a result, hundreds of Iranians died of air pollution in 2017 [11]. According to the studies, Tabriz has been identified as one of Iran's most polluted cities [2, 12, 13].

Tabriz's Air Quality Monitoring Stations (AQMS), which is part of the Environmental Protection Organization, has established nine (spatial) stations in the city. These stations take hourly (temporal) measurements of air

✉ Parvin Sarbakhsh
p.sarbakhsh@gmail.com

Parisa Saeipourdizaj
s_parisa72@yahoo.com

Saeed Musavi
saeedmusavi@ymail.com

Akbar Gholampour
Gholampoura@tbzmed.ac.ir

[1] Health and Environment Research Center, Department of Statistics and Epidemiology, Faculty of Health, Tabriz University of Medical Sciences, Tabriz, Iran

[2] Health and Environment Research Center, Department of Environmental Health Engineering, Faculty of Health, Tabriz University of Medical Sciences, Tabriz, Iran

pollution concentrations, resulting in massive volumes of spatiotemporal data. This recorded information should be analyzed to monitor and control the pollution level and finally take appropriate actions. Traditional approaches are unable to process and interpret massive data with spatiotemporal dependencies due to their complexity. Appropriate data mining approaches are required to analyze this heterogeneous data and derive valuable information.

Clustering is one of the most important statistical approaches in data mining since it examines enormous datasets such as pollution levels in the air. Clustering is the process of grouping objects in datasets that have similar features together. In other words, it is a data function that divides observations into subgroups based on their similarity and identifies the data pattern. Clusters can be conveniently studied and interpreted due to clustering's ability to reduce data size. As a result, in air pollution data with a large amount, clustering can recognize the pattern and provide more clear interpretation by lowering the amount of the data. The effects of meteorological factors on pollutant concentrations inside each cluster can be studied using obtained clusters of air pollution data.

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

Recently, a new model-based clustering method for spatiotemporal data has been introduced by Cheam et al. [1], which considers both spatial and temporal dimensions for fitting a mixture model for spatiotemporal data. Changes in time and space were considered simultaneously in this method. Thus, we can cluster the days of the year according to the spatiotemporal information of pollutant concentrations.

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

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

2 Method

2.1 Data

Tabriz is the capital of East Azerbaijan in Iran, and it is one of the country's largest and most industrialized cities. Tabriz has a population of almost 1.8 million people and a land area of 320 square kilometers [24], according to the most recent census in 2017. Tabriz has four seasons and is semi-arid in terms of climate, so that it is rainy and moderate in spring, hot and dry in summer, rainy in autumn, and cold and snowy in winter. The ambient temperature reaches 30–35 °C on hot summer days and − 10 to − 20 °C on cold winter days [27]. Furthermore, the average yearly wind speed is 1.65 m per second [24, 28]. In 2017, the average temperature was around 13.32 °C, with the lowest and highest temperatures of (− 5, 39.40 °C), average relative humidity of 20.32%, and average wind speed of 7.02 km per hour, according to Tabriz Meteorological Organization reports.

Based on the information received from AQMS, the recorded hourly data on the pollutant concentrations in 2017 were more accurate and complete than in other recent years. As a result, the pollutant data collected in 2017 was selected to assess and analyze.

PM_{10} (particulate matter with an aerodynamic diameter less than 10 μm) and O_3 (surface ozone) concentrations were measured at each station using beta attenuation and UV spectrophotometry, respectively. Table 1 shows the geographical coordinates of four quality monitoring stations in Tabriz (see Table 1).

Air quality information usually contains inaccurate, missing, and outlier data. As a result, the available information was examined for outliers and missing data. The outlier and inaccurate data were removed by the z-score method before inputing the missing data, estimating the parameters, and data mining [2, 29, 30]. As a result, the hourly concentrations of monitoring stations were compared to time trend data from the same stations and data from nearby monitoring stations. First, the original data were converted into z-scores (with mean = 0 and SD = 1) based on the following settings: (1) having $|z| > 4(|z_t > 4|)$, (2) the difference from the prior value being greater than $9\left(z_t - z_{t-1} > 9\right)$, (3) the ratio of the z-score value to its centered moving average of order 3(MA3) being greater than $2(z_t/MA3\left(z_t\right) > 2)$, and (4) the difference between the singular monitoring air quality station and the prior value $(z_t - z_{t-1})$ being at least twice greater than the difference between the city's monitoring air quality station's averaged increment $(\text{city}(z_t - z_{t-1}))$ and the prior value $[(z_t - z_{t-1})/city(z_t - z_{t-1})] > 2$, and then the outlier data were deleted from the original data.

According to WHO guidelines, there must be at least 50% valid information of 1 year and the desired station for

Table 1 Geographical coordinates of air quality monitoring station

Latitude			Longitude			Station
Second	Minute	Degree	Second	Minute	Degree	
12	19	46	36	03	38	Abresan
47	13	46	36	03	38	Baghshomal
52	14	46	12	04	38	Rahahn
23	17	46	12	04	38	Rastekoche

statistical analysis, so the information for January to December 2017, which had at least 50% of the complete data, was selected (250 days). According to the instructions mentioned for the valid data in AQMS, PM_{10} and O_3 data of four monitoring stations were considered in this research (Abrasan, Baghshamal, Rahahan, and Rastekoche). It should be noted that only the hourly information of 4 out of 9 stations was available. Of the total data available for pollutants (PM_{10}, O_3), the percentage of missing data was (11.88%, 0.6%) for Abrasan, (3.3%, 3.22%) for Baghshamal, (2.22%, 1.68%) for Rahahan, and (2.38%, 1.58%) for Rastekoche. The missing data were inputed using the linear interpolation method [31] with R (4.0.2) software (package: imputeTS version 3.1) to improve the accuracy and validity of the results. Hourly concentrations of each pollutant (PM_{10}, O_3) are used to compute the diurnal concentration. The acquired diurnal data are then utilized to assess the variations in air pollution in each cluster. Meteorological data for Tabriz was gathered from https://en.tutiempo.net in 2017.

2.2 Statistical Analysis

The spatiotemporal air pollution data were clustered by fitting a mixture model–based that took into account the spatiotemporal dimensions (STMC). To fit the model, spatial (geographical coordinates of data recording stations) and temporal (day and hour of recorded data) information were taken into account, an optimal number of clusters was determined using the Bayesian Information Criterion (BIC) [32] and Integrated Completed Likelihood criterion (ICL) [33], and the model parameters were estimated using the EM algorithm. Finally, each observation was categorized into an appropriate cluster.

STMC was compared to temporal mixture model–based clustering that only considered the temporal dimension (TMC) and mixture model–based clustering that did not consider the spatiotemporal dimensions (MC) to see how well it performed. BIC and ICL are the criteria used to compare the fitted models. The TMC and MC models were fitted in the same manner as the STMC models, except for a difference in considering the dimensions.

The nonparametric Mann–Kendall test (MK) [2, 34–36] was calculated to investigate the relationship between meteorological factors and pollutant concentrations (PM_{10} and O_3) within the detected clusters.

R (4.0.2) software with "SpaTimeClus" (version 1.0) and "mclust" (version 5.4.6) packages were used for statistical analysis. The following is a brief explanation of STMC.

2.2.1 Mixture models for Spatiotemporal Data

Mixture model–based clustering is a broad family of algorithms designed for modeling an unknown distribution as a mixture of simpler distributions, sometimes called basis distribution [37, 38]. In this method, unlike other clustering methods that cluster the data based on some similarity measures, a Gaussian statistical distribution was considered for data. Thus, the purpose of model-based clustering is to estimate the parameters of the statistical distribution and hidden variables, considered as a cluster label of the data [37, 39]. Moreover, this model is applicable to a variety of data kinds such as continuous, ordinal, categorical, mixed, and functional.

Extension of mixture model–based clustering for spatiotemporal data was introduced in 2017 by Cheam et al. [1], which is a generalization for mixture model–based clustering considering only the temporal dimension [40]. Besides, the STMC method considers the variations in time and space simultaneously. Suppose each observation $x_i = \{x_{ijt}\}_{\substack{j=1,\dots,J \\ t=1,\dots,T}}$, as spatiotemporal data, including J × T observations on the predetermined temporal framework $m = (m_1, \dots, m_T)$ and the spatial framework $s = (s_1, \dots, s_J)$. Therefore, $x_{ijt} \in \mathbb{R}$ is the observation i at location j and time t. For spatial coordinates, location j is a two-dimensional vector $s_j = (v_j, w_j)$.

The density function of the cth element (or cluster) for the spatiotemporal model is as follows:

$$f_c(x_i|\vartheta_c) = \prod_{j=1}^{J} \prod_{t=1}^{T} \sum_{h=1}^{H} \Upsilon_{cjth}(a_c) \times \varphi\left(x_{ijt}|\left[M_t - M_{t-1}\right]' \beta_{ch} + x_{ij(t-1)}, \sigma^2_{ch}\right) \quad (1)$$

so that $\vartheta_c = \left(a_c, \beta_{ch}, \varrho^2_{ch}, h = 1, \dots, H\right)$ is a set of parameters of the element cth, $M_t = \left(1, m_t, \dots, m_t^{\varrho}\right)$ the polynomial vector with Q-degree of M_t, $\beta_{ch} = \left(\beta_{ch0}, \dots, \beta_{ch\varrho}\right)'$ is the

coefficient vector of the hth regression model for the cth element, and $\phi\left(.|\mu,\varrho^2\right)$ is the univariate Gaussian distribution density with mean μ and variance σ^2. The weight of this mixture, γ_{cjth}, depends on both dimensions (time and space) via a logistic function. In the expectation and maximization steps of EM algorithm, the parameters of the model, which includes $\vartheta_c = \left(a_c, \beta_{ch}, \sigma^2_{ch}, h = 1, \ldots, 5\right)$, were estimated. In this study, each x_{ijt} represents the hourly concentration of the desired pollutant, i.e.,, observation i in time t (T=24 h per day for n=250 days) and location j (J=4 stations).

2.2.2 Model Selection

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

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

so, M is obtained by applying the maximum number of each element. Thus, assuming $C_{\min} = H_{\min} = \mathcal{Q}_{\min} = 1$, the number of selectable models is as follows:

$$\mathrm{Models}(M) = C_{\max} \times H_{\max} \times \mathcal{Q}_{\max} \tag{3}$$

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

3 Results

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

3.1 Description of Clusters Obtained from STMC

Figure 1 shows the scatter plot of the diurnal average concentrations of PM_{10} (right) and O_3 (left), where each representative dot corresponds to the day profile of pollutant measurements for a cluster (see Fig. 1). Also, Table 3 presents descriptive information on the detected clusters and meteorological factors (see Table 3). It is important to note that the number of available days for analyzing the concentration variations of pollutants in each season was spring 93 days, summer 31 days, autumn 90 days, and winter 36 days.

3.1.1 Description of Clusters Obtained for O_3 Pollutant

In clustering O_3 concentrations based on BIC and ICL criteria, STMC with three clusters was selected as the best model. O_3 contains three clusters with average concentrations of low (cluster2), moderate (cluster1), and high (cluster3). O_3-cluster1 encompasses 102 days which is the most member days.

Spring forms approximately 60% of this cluster, and the rest of the members are in O_3-cluster3. The average concentration for spring in O_3-cluster3 (65.51 $\mu g/m^3$) is higher than O_3-cluster1 (55.87 $\mu g/m^3$). O_3-cluster2 includes approximately 85% of autumn days (59 days), and the rest of the members are in O_3-cluster1. Although O_3-cluster1 encompasses less number of autumn days (22 days), the average concentration of O_3-cluster1 (48.39 $\mu g/m^3$) members is higher than O_3-cluster2 (18.68 $\mu g/m^3$) members. Additionally, most winter days (22 days) are in O_3-cluster1.

3.1.2 Description of Clusters Obtained for PM_{10} Pollutant

In clustering PM10 concentrations based on BIC and ICL criteria, STMC with four clusters was selected as the best model. PM10-cluster4 has the highest number of days (90 days) and the highest average concentration (78.77

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

Number of clusters	Number of free parameters	ICL	BIC	pollutant	Clustering model
9	134	−197,991.8	−195,164.4	O_3	MC
9	134	−210,927.6	−207,898.5	PM_{10}	
4	139	−91,726.87	−91,722.5	O_3	TMC
5	114	−91,625.41	−91,618.1	PM_{10}	
3	122	−91,677.84	−91,670.71	O_3	STMC
4	107	−90,530.45	−90,525.04	PM_{10}	

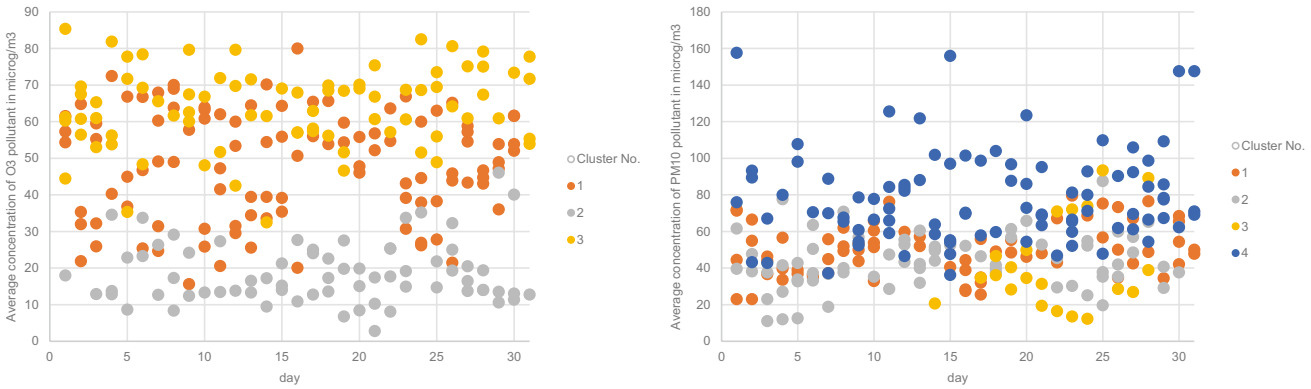

Fig. 1 Scatter plot of the diurnal average concentration of PM_{10} (right) and O_3 (left) clusters in STMC model

$\mu g/m^3$). PM10-cluster1 and PM10-cluster2 were almost the same in member days (70 and 69 days, respectively), whereas the average concentration of PM10-cluster1 (49.69 $\mu g/m^3$) was higher than the average concentration of PM10-cluster2 (43.81 $\mu g/m^3$). PM10-cluster3 (41.87 $\mu g/m^3$) contained the lowest member days (21 days), and its average concentration was close to PM10-cluster2.

Spring days were present in all clusters, and this season's highest and lowest average concentrations belonged to PM10-cluster4 and PM10-cluster2, respectively. It is important to note that regarding the spring, PM10-cluster1 and PM10-cluster3 had moderate average concentrations and were close in values. However, PM10-cluster3 had a slightly lower average concentration than PM10-cluster1. Summer days were present in PM10-clusters 1, 2, and 4, but PM10-cluster1 included only 2 days which can be disregarded.

Additionally, PM10-cluster4 had a higher average concentration than PM10-cluster2. Autumn days were seen in PM10-cluster2 (approximately 60%) and PM10-cluster4 (approximately 37%). PM10-cluster4 containing low member days exhibited a higher average concentration than PM10-cluster2. Regarding winter, PM10-cluster1 included 50% of days and PM10-cluster4 30% of them. Furthermore, PM10-cluster4 had a higher average concentration than PM10-cluster1.

Table 3 Descriptive statistics for PM_{10} and O_3 clusters obtained from STMC

Descriptive statistics		Pollutant						
		Cluster of PM_{10}				Cluster of O_3		
		C1	C2	C3	C4	C1	C2	C3
Average concentration of all stations ($\mu g/m^3$	Spring	46.19	37.53	41.23	75.78	55.87	25.34	65.51
	Summer	38.05	55.28	–	94.45	47.21	–	68.64
	Autumn	52.28	41.39	–	77.03	32.92	18.51	47.19
	Winter	56.45	48.67	45.67	69.19	25.34	18.98	50.81
Average concentration of all stations ($\mu g/m^3$)		49.69	43.81	41.87	78.77	48.39	18.68	63.85
Standard Deviation ($\mu g/m^3$)		14.06	15	24.37	25.21	16.64	8.36	10.82
Range ($\mu g/m^3$)		127.82	103.89	124.20	168.40	64.36	43.27	52.9
Minimum ($\mu g/m^3$)		9.86	4.99	5.63	26.71	15.64	2.81	32.51
Maximum ($\mu g/m^3$)		137.68	108.88	129.83	195.11	80	46.09	85.41
Number of days		70	69	21	90	102	70	78
Average temperature (°C)		12.44	13.67	11.26	15.33	13.67	6.68	20.21
Average relative humidity (%)		47.80	49.54	50.81	44.29	46.94	58.19	37.79
Average total rainfall/snowmelt (mm)		0.94	1	1.73	0.27	1.45	0.22	0.41
Average visibility (km)		9.19	9.61	9.60	10.08	9.48	8	11.08
Average wind speed (km/h)		12.40	11.67	14.84	13.87	12.43	10.64	15.69

3.2 Analysis of the Correlation Between Pollutant Concentrations and Meteorological Factors Within Clusters in STMC

The MK test to assess the correlation between meteorological factors and the concentrations of PM_{10} and O_3 was calculated, shown in Table 4 (see Table 4).

According to the MK's test results, all clusters presented a statistically significant correlation between the average concentration of O3 and meteorological factors, except O3-cluster2 for rainfall. The relationship between the average concentration of O3 and temperature in all clusters was positive and weak to moderate with a very high statistical significance (p-value < 0.05); in respect of relative humidity, it was negative and moderate with a very high statistical significance (p-value < 0.05). Moreover, all clusters showed a positive and weak relationship between the average concentration of O3 and wind speed.

The MK's test results for PM10 revealed a statistically significant correlation between average concentration and meteorological factors (temperature and relative humidity) in PM10-cluster2 and PM10-cluster3. The relationship between temperature and average concentration in PM10-cluster2 was positive and weak to moderate, with a very high statistical significance (p-value < 0.05), but it was negative and weak about relative humidity. The relationship between temperature and average concentration in PM10-cluster3 was positive and moderate to strong whereas the relationship between temperature and relative humidity was negative and moderate to strong. Furthermore, in PM10-cluster1, PM10-cluster2, and PM10-cluster3, the relationship between rainfall and average concentration was negative and weak. It is important to note that there was no statistically significant correlation between PM10 and wind speed.

4 Discussion

The goal of our study was to apply a mixture model–based clustering method for clustering the days according to hourly concentrations of PM_{10} and O_3 over the 1 year in Tabriz. This method does not consider the type of data and can be used for various types such as continuous, ordinal, categorical, mixed, and functional. Spatial and temporal information is always available in the air pollution datasets, which is a type of functional data. Therefore, the mixture model–based clustering method is a proper choice for analyzing this functional data with spatial and temporal dependencies.

STMC with spatial and temporal dependencies, TMC with temporal dependency, and MC lacking in both dependencies were fitted to data to assess the applicability of statistical models for our data. BIC, ICL, and number of free parameters were calculated for these models. The results showed that the STMC compared to TMC and MC had a better fitting on spatiotemporal data of PM_{10} and O_3. The following is a review of various pollutant clustering studies in which none of the dependencies (spatial and temporal) are simultaneously considered.

Jin et al. [41] demonstrated the application of the k-means clustering method to identify O_3 spatial regimes (or clusters) over California's San Joaquin Valley. Clusters show the days having the same O_3 geographical distribution. In terms of concentration, of a total of six recognized regimes, two corresponded to low-O_3-cluster, three to moderate-O_3-cluster, and one to high-O_3-cluster. Moreover, meteorological measurements were used to describe O_3 spatial distributions and their correlation to those in San Francisco Basin.

Pandey et al. [42] demonstrated the spatial and temporal variability of $PM_{1.0}$, $PM_{2.5}$, PM_{10}, NO_2, and SO_2 in India using the average linkage clustering approach. Clusters including monitoring sites represent similar behavior in

Table 4 Mann–Kendall correlation coefficient between meteorological parameters and PM_{10} and O_3 clusters

Meteorological parameters	Cluster of PM_{10}				Cluster of O_3		
	C1	C2	C3	C4	C1	C2	C3
Average temperature (°C)	−0.101 0.218	0.288** 0.000	0.536* 0.001	0.085 0.235	0.322** 0.000	0.299** 0.000	0.318** 0.000
Average relative humidity(%)	−0.004 0.964	−0.233** 0.005	-0.404* 0.011	−0.055 0.447	−0.345** 0.000	−0.401** 0.000	−0.357** 0.000
Average total rainfall and/or snowmelt (mm)	−0.196* 0.039	−0.219* 0.022	−0.183 0.260	−0.220** 0.003	−0.224** 0.003	0.057 0.558	−0.251** 0.006
Average visibility (km)	−0.253** 0.003	0.086 0.312	−0.135 0.446	−0.208* 0.014	0.301** 0.000	0.276** 0.001	0.211* 0.010
Average wind speed (km/h)	−0.230** 0.005	−0.037 0.652	0.091 0.566	−0.001 0.992	0.280** 0.000	0.281** 0.001	0.213** 0.001

**Correlation significant is 0.01 level; *correlation significant is 0.05 level

terms of pollutant dispersions and spatial variations. According to the results, concentrations of all varieties of PM_{10} were highest in the winter and lowest in the wet season at all sites.

Huang et al. [43] investigated the characteristics of $PM_{2.5}$ in China using a hierarchical clustering approach. According to the results, $PM_{2.5}$ information, collected from 13 monitoring sites, was arranged in 3 clusters. One of the most important findings was that the temporal distribution of $PM_{2.5}$ demonstrated that winter had the highest concentration, and fall has a higher concentration value than spring, and the lowest concentration belongs to summer. In terms of spatial distribution, three out of 13 monitoring sites exhibited the highest concentration of $PM_{2.5}$.

In a study conducted in northern China [44], cluster analysis was used to reveal the spatial and temporal distribution of $PM_{2.5}$, SO_2, NO_2, CO, and O_3 pollutants. Pearson's correlation coefficient was used to investigate the relationship between pollutant concentrations with each other. Hierarchical cluster analysis was used in this study to categorize nine cities into various groups based on a monthly average of the above pollutants in each city. Using cluster analysis, nine cities were divided into four clusters based on the monthly average of pollutants. According to the clustering analysis, air pollution was mainly associated with industrial city structures and geographical and socioeconomic factors.

Many studies on the relationship between pollutant concentrations and meteorological factors have been conducted. They indicate that the hourly, daily, monthly, and seasonal variations of air pollutants in a residential area can be caused by meteorological factors such as atmospheric temperature, relative humidity, wind speed, solar radiation intensity, and so on. The amount of pollutant concentration will be affected by changing the above elements during the seasons [2, 45, 46].

Among the meteorological factors, atmospheric stability and wind speed have the most influence on atmospheric dispersion and decreasing air pollution [30, 34]. The intensification of air pollution in Tabriz can be attributed to the temperature inversions and calm conditions.

Geographical characteristics, meteorological variables, and residential buildings in the direction of the wind entering the city all contribute to the city's current situation. It should be mentioned that the data show a fluctuating variation in PM_{10} and O_3 concentrations in Tabriz from 2006 to 2017 [25].

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

Based on most conducted studies, there was a clear and logical trend in the monthly and seasonal variations of O_3 concentration. The highest and lowest values of O_3 concentration were reported in summer and winter, respectively, according to the variation pattern. This highest value may be due to increased atmospheric temperature, increased intensity of sunlight, long days, and long sunny hours in hot seasons, all of which increase photochemical reactions and O_3 production, whereas the lowest value is related to reduced daylight (sunlight time), lower temperature, and sunshine duration [2, 45, 47]. O_3 generation is also influenced by several processes and activities such as transport, deposition, and NOx titration. The increase in O_3 concentration in the warm seasons is directly related to the increment degree of temperature that is one of the main parameters in controlling the O_3 formation [48, 49].

According to the results of this study, PM_{10}-cluster4 had the highest temperature. PM_{10}-cluster3 and PM_{10}-cluster2 had lower temperatures, higher rainfall, and higher wind speed. It is important to note that PM_{10}-cluster3 and PM_{10}-cluster2 were close in values with slight differences. In the following, the effect of meteorological factors on PM_{10} concentration in different seasons is summarized.

Because of the Asian dust phenomenon, PM_{10} concentrations may rise in the spring and summer [36, 50]. Furthermore, the concentration of PM_{10} rises during warmer seasons due to high ambient temperatures and low relative humidity. Moreover, the correlation between PM_{10} concentration with temperature and relative humidity is positive and negative, respectively [8].

The studies on the relationship between PM_{10} and rainfall show that the washout effects of rainfall reduce PM_{10} concentrations in the atmosphere. Rainfall and relative humidity eventually have a negative correlation with PM_{10} [51].

According to the results of the analysis of the average concentrations of O_3 and PM_{10} in each cluster, the relationship between meteorological factors and average concentration in each cluster, and the member days of seasons in individual clusters, it is possible to conclude that the clustering method used in this study has good fitness, and the results were interpretable. Finally, model-based clustering provides a strong basis and shows excellent efficiency in analyzing data with spatiotemporal dimensions. Using the above approach on datasets with spatial and temporal information produces more reliable and accurate results [1].

5 Conclusion

Data mining methods are very applicable in evaluating air pollution data, and the results play a significant role in managing and preventing the accumulation of pollutants. Clustering, one of the data mining approaches, reduces the amount of data examined while revealing hidden information. As a result, analyzing individual clusters with a small amount of data decreases the errors in the results. In this study, the STMC using spatiotemporal dimensions classify the days in Tabriz, 2017, based on their PM_{10} and O_3 concentrations. The results demonstrated that considering the dimensions of the data in the analysis could increase the efficacy of the clustering in our spatiotemporal air pollution data. The findings of STMC's examination of the obtained clusters in terms of available meteorological parameters suggested that the clustering was acceptable and meaningful.

6 Limitation

This study may have some limitations:

- Over the year 2017, there were insufficient measurements of pollutant concentrations in the winter and summer. This insufficiency was evident in the data collected in January, February, July, and September.
- Due to the different sources for pollutant production in each region, geographical conditions and meteorological factors that influence on the pollutant level variations, it was impossible to directly compare with the results of the other studies.

Author Contribution Parisa Saeipourdizaj (first author): formulation and evaluation of overarching research goals and aims; setting the data in software package format; application of statistical, computational, and other formal techniques to analyze; application of available software codes; preparation (drafting, reviewing, translating, and revising the paper), and presentation of the manuscript. Saeed Musavi: statistical analysis, manuscript preparation and reviewing the paper. Akbar Gholampour: reviewing the paper. Parvin Sarbakhsh (corresponding author): formulation and evaluation of overarching research goals and aims; statistical analysis; preparation (drafting, reviewing, translating, and revising the paper), and presentation of the manuscript. All the authors have read and approved the final manuscript.

Funding This work was supported by the Tabriz University of Medical Sciences. The Health and Environment Research Center provided financial support.

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

Code Availability Not applicable.

Declarations

Ethics Approval This article has been extracted from the thesis submitted for MSc degree in Biostatistics which has been approved by the ethics committee of Tabriz University of Medical Sciences (Ethic number: IR.TBZMED.REC.1398.352).

Competing Interests The authors declare no competing interests.

References

1. Cheam, A. S. M., Marbac, M., & McNicholas, P. D. (2017). Model-based clustering for spatiotemporal data on air quality monitoring. *Environmetrics, 28*(3), e2437. https://doi.org/10.1002/env.2437
2. Faridi, S., Shamsipour, M., Krzyzanowski, M., Künzli, N., Amini, H., Azimi, F., Malkawi, M., Momeniha, F., Gholampour, A., Hassanvand, M. S., & Naddafi, K. (2018). Long-term trends and health impact of PM2.5 and O3 in Tehran, Iran, 2006–2015. *Environmental International, 114*, 37–49. https://doi.org/10.1016/j.envint.2018.02.026.
3. Manju, A., Kalaiselvi, K., Dhananjayan, V., Palanivel, M., Banupriya, G. S., Vidhya, M. H., Panjakumar, K., & Ravichandran, B. (2018). Spatio-seasonal variation in ambient air pollutants and influence of meteorological factors in Coimbatore, Southern India. *Air Quality, Atmosphere and Health., 11*, 1179–1189. https://doi.org/10.1007/s11869-018-0617-x
4. Zhang, H., Wang, Y., Hu, J., Ying, Q., & Hu, X. M. (2015). Relationships between meteorological parameters and criteria air pollutants in three megacities in China. *Environmental Research., 140*, 242–254. https://doi.org/10.1016/j.envres.2015.04.004
5. Shukla, J. B., Misra, A. K., Sundar, S., & Naresh, R. (2008). Effect of rain on removal of a gaseous pollutant and two different particulate matters from the atmosphere of a city. *Mathematical and Computer Modelling., 48*, 832–844.
6. Goyal, S. K., & Rao, C. V. C. (2007). Assessment of atmospheric assimilation potential for industrial development in an urban environment: Kochi (India). *Science of the total environment., 376*, 27–39.
7. Owoade, O. K., Olise, F. S., Ogundele, L. T., Fawole, O. G., & Olaniyi, H. B. (2012). Correlation between particulate matter concentrations and meteorological parameters at a site in Ile-Ife, Nigeria. *Ife Journal of Science no, 1*(14), 83–93.
8. Dominick, D., Latif, M. T., Juahir, H., Aris, A. Z., & Zain, S. M. (2012). An assessment of influence of meteorological factors on PM sub (10) and NO sub (2) at selected stations in Malaysia. *Sustainable Environment Research., 22*, 305–315.
9. Islam, M. M., Afrin, S., Ahmed, T., & Ali, M. A. (2015). Meteorological and seasonal influences in ambient air quality parameters of Dhaka city. *Journal of Civil Engineering, 43*, 67–77.
10. Galindo, N., Yubero, E., Nicola, J. F., & Crespo, J. (2015). Chemical characterization of PM1 at a regional background site in the western Mediterranean. *Aerosol and Air Quality Research., 16*, 530–541.
11. Number of deaths from air pollution, 1990 to 2017. https://ourworldindata.org/. Accessed 10 November 2021.
12. Naddafi, K., Hassanvand, M. S., Yunesian, M., Momeniha, F., Nabizadeh, R., Faridi, S., & Gholampour, A. (2012). Health impact assessment of air pollution in megacity of Tehran, Iran. *Iranian journal of environmental health science & engineering, 9*, 28.
13. Hassanvand, M. S., Naddafi, K., Faridi, S., Arhami, M., Nabizadeh, R., Sowlat, M. H., Pourpak, Z., Rastkari, N., Momeniha, F., & Kashani, H. (2014). Indoor/outdoor relationships of PM10, PM2.5, and PM1 mass concentrations and their water-soluble ions in a

retirement home and a school dormitory. *Atmospheric Environment*, *82*, 375–382.
14. Lavecchia, C., Angelino, E., Bedogni, M., Bravetti, E., Gualdi, R., Lanzani, G., Musitelli, A., & Valentini, M. (1996). The ozone patterns in the aerological basin of Milan (Italy). *Environmental Software., 11*, 73–80.
15. Saksena, S., Joshi, V., & Patil, R. S. (2003). Cluster analysis of Delhi's ambient air quality data. *Journal of Environmental monitoring., 5*, 491–499.
16. Gramsch, E., Cereceda-Balic, F., Oyola, P., & Von Baer, D. (2006). Examination of pollution trends in Santiago de Chile with cluster analysis of PM10 and ozone data. *Atmospheric environment., 40*, 5464–5475.
17. Molinari, N. (2007). Free knot splines for supervised classification. *Journal of classification., 24*, 221–234.
18. Gabusi, V., & Volta, M. (2005). A methodology for seasonal photochemical model simulation assessment. *International journal of environment and pollution., 24*, 11–21.
19. Morlini, I. (2007). Searching for structure in measurements of air pollutant concentration. *Environmetrics: The official journal of the International Environmetrics Society*, *18*, 823–840.
20. Fraley, C., & Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American statistical Association*, 97(458), 611-631.
21. Vrbik, I., & Mcnicholas, P. D. (2014). Parsimonious skew mixture models for model-based clustering and classification. *Computational Statistics & Data Analysis., 71*, 196–210.
22. Murphy, K., & Murphy, T. B. (2017). Parsimonious model-based clustering with covariates. arXiv preprint arXiv:1711.05632.
23. Asghari, F. B., & Mohammadi, A. A. (2019). The effect of the decreasing level of Urmia Lake on particulate matter trends and attributed health effects in Tabriz, Iran. *Microchemical Journal,* 104434 https://doi.org/10.1016/j.microc.2019.104434.
24. Amini Parsa, V., Salehi, E., Yavari, A. R., & van Bodegom, P. M. (2019). Analyzing temporal changes in urban forest structure and the effect on air quality improvement. *Sustainable Cities and Society., 48*, 101548. https://doi.org/10.1016/j.scs.2019.101548
25. Barzeghar, V., Sarbakhsh, P., Hassanvand, M. S., Faridi, S., & Gholampour, A. (2020). Long-term trend of ambient air PM10, PM2. 5, and O3 and their health effects in Tabriz city, Iran, during 2006–2017. *Sustainable Cities and Society, 54*, 101988.
26. Yicun, G., Khorshiddoust, A. M., Mohammadi, G. H., & Sadr, A. H. (2020). The relationship between PM 2.5 concentrations and atmospheric conditions in severe and persistent urban pollution in Tabriz , Northwest of Iran.
27. Azarafza, M., & Ghazifard, A. (2016). Urban geology of Tabriz City: environmental and geological constraints. *Advances in environmental research. 5*, 95–108. https://doi.org/10.12989/aer.2016.5.2.095.
28. Kalajahi, M.J., Khazini, L., Rashidi, Y., & Heris, S.Z. (2019). Development of reduction scenarios based on urban emission estimation and dispersion of exhaust pollutants from light duty public transport: case of Tabriz, Iran. *Emission Control Science and Technology,* 1–19.
29. Barrero, M. A., G. Orza, J., Cabello, M., & Cantón, L. (2015). Categorisation of air quality monitoring stations by evaluation of PM10 variability. *The Science of the total environment,* 524–525C, 225–236. https://doi.org/10.1016/j.scitotenv.2015.03.138.
30. Song, C., He, J., Wu, L., Jin, T., Chen, X., Li, R., Ren, P., Zhang, L., & Mao, H. (2017). Health burden attributable to ambient PM2.5 in China. *Environmental pollution (Barking, Essex : 1987), 223*, 575–586. https://doi.org/10.1016/j.envpol.2017.01.060.
31. Norazian, M. N., Shukri, Y. A., Azam, R. N., & Al Bakri, A. M. M. (2008). Estimation of missing values in air pollution data using single imputation techniques. *ScienceAsia, 34*, 341–345 https://doi.org/10.2306/scienceasia1513-1874.2008.34.341.
32. Schwarz, G. (1978). Estimating the dimension of a model. *The annals of statistics., 6*, 461–464.
33. Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, 19(6), 716-723.
34. Vuorenmaa, J., Augustaitis, A., Beudert, B., Bochenek, W., Clarke, N., de Wit, H. A., Dirnbock, T., Frey, J., Hakola, H., & Kleemola, S. (2018). Long-term changes (1990–2015) in the atmospheric deposition and runoff water chemistry of sulphate, inorganic nitrogen and acidity for forested catchments in Europe in relation to changes in emissions and hydrometeorological conditions. *Science of the total environment., 625*, 1129–1145.
35. Cerro, J. C., Cerda, V., & Pey, J. (2015). Trends of air pollution in the Western Mediterranean Basin from a 13-year database: A research considering regional, suburban and urban environments in Mallorca (Balearic Islands). *Atmospheric Environment., 103*, 138–146.
36. Ahmed, E., Kim, K.-H., Shon, Z.-H., & Song, S.-K. (2015). Long-term trend of airborne particulate matter in Seoul, Korea from 2004 to 2013. *Atmospheric Environment., 101*, 125–133.
37. McLachlan, G., & Peel, D. (2000). Finite mixture models, willey series in probability and statistics.
38. McNicholas, P. D. (2016). Mixture model-based classification. *Chapman and Hall/CRC,*
39. Mcnicholas, P. D. (2016). *Model-based clustering., 373*, 331–373. https://doi.org/10.1007/s0035
40. Same, A., Chamroukhi, F., Govaert, G., & Aknin, P. (2011). Model-based clustering and segmentation of time series with changes in regime. *Advances in Data Analysis and Classification., 5*, 301–321.
41. Jin, L., Harley, R. A., & Brown, N. J. (2011). Ozone pollution regimes modeled for a summer season in California's San Joaquin Valley: A cluster analysis. *Atmospheric environment., 45*, 4707–4718.
42. Pandey, B., Agrawal, M., & Singh, S. (2014). Assessment of air pollution around coal mining area: Emphasizing on spatial distributions, seasonal variations and heavy metals, using cluster and principal component analysis. *Atmospheric pollution research., 5*, 79–86.
43. Huang, P., Zhang, J., Tang, Y., & Liu, L. (2015). Spatial and temporal distribution of PM2. 5 pollution in Xi'an City, China. *International journal of environmental research and public health, 12*, 6608–6625.
44. Tian, D., Fan, J., Jin, H., Mao, H., Geng, D., Hou, S., Zhang, P., & Zhang, Y. (2020). Characteristic and spatiotemporal variation of air pollution in Northern China based on correlation analysis and clustering analysis of five air pollutants. *Journal of Geophysical Research: Atmospheres, 125*, e2019JD031931.
45. Sicard, P., Serra, R., & Rossello, P. (2016). Spatiotemporal trends in ground-level ozone concentrations and metrics in France over the time period 1999–2012. *Environmental research., 149*, 122–144.
46. Zhao, S., Yu, Y., Yin, D., He, J., Liu, N., Qu, J., & Xiao, J. (2016). Annual and diurnal variations of gaseous and particulate pollutants in 31 provincial capital cities based on in situ air quality monitoring data from China National Environmental Monitoring Center. *Environment international., 86*, 92–106.
47. Carvalho, V. S. B., Freitas, E. D., Martins, L. D., Martins, J. A., Mazzoli, C. R., & de Fatima Andrade, M. (2015). Air quality status and trends over the Metropolitan Area of Sao Paulo, Brazil as a result of emission control policies. *Environmental Science & Policy, 47*, 68–79.
48. Lacressonniere, G., Foret, G., Beekmann, M., Siour, G., Engardt, M., Gauss, M., Watson, L., Andersson, C., Colette, A., & Josse, B. (2016). Impacts of regional climate change on air quality projections and associated uncertainties. *Climatic Change., 136*, 309–324.
49. Pawlak, I., & Jarosawski, J. (2015). The influence of selected meteorological parameters on the concentration of surface ozone in the central region of Poland. *Atmosphere-Ocean, 53*, 126–139.
50. Jang, E., Do, W., Park, G., Kim, M., & Yoo, E. (2017). Spatial and temporal variation of urban air pollutants and their concentrations

in relation to meteorological conditions at four sites in Busan. *South Korea. Atmospheric Pollution Research., 8*, 89–100.
51. Giri, D., ADHIKARY, P. R., MURTHY, V. K. (2008). The influence of meteorological conditions on PM10 concentrations in Kathmandu Valley.