Atmospheric Pollution Research

www.atmospolres.com

Using principal component analysis and fuzzy c–means clustering for the assessment of air quality monitoring

Senay Cetin Dogruparmak [1], Gulsen Aydin Keskin [2], Selin Yaman [3], Atakan Alkan [2]

[1] Department of Environmental Engineering, Kocaeli University, 41380 Kocaeli, Turkey
[2] Department of Industrial Engineering, Kocaeli University, 41380 Kocaeli, Turkey
[3] Institute of Science, Environmental Sciences, Kocaeli University, 41380, Turkey

ABSTRACT

Determining whether a reduction can be made in the total number of monitoring stations within the Air Quality Monitoring Network is very important since in case of necessity, the devices at one group of stations having similar air pollution characteristics can be transferred to another zone. This would significantly decrease the capital investment and operational cost. Therefore, the objective of this study was grouping the monitoring stations that share similar air pollution characteristics by using the methods of principal component analysis (PCA) and fuzzy c–means (FCM). In addition, this study also enables determining the emission sources, evaluating the performances of the methods and examining the zone in terms of pollution. In the classification of monitoring stations, different groups were formed depending on both the method of analysis and the type of pollutants. As a result of PCA, 5 and 3 classes have been determined for SO_2 and PM_{10}, respectively. This shows that the number of monitoring stations can be decreased. When reduced classes were analyzed, it was observed that a clear distinction cannot be made considering the affected source type. During the implementation of the FCM method, in order to facilitate comparison with the PCA, the monitoring stations were classified into 5 and 3 groups for SO_2 and PM_{10}, respectively. When the results were analyzed, it was seen that the uncertainty in PCA was reduced. When the two methods are compared, FCM was found to provide more significant results than PCA. The evaluation in terms of pollution, the results of the study showed that PM_{10} exceeded the limit values at all the monitoring stations, and SO_2 exceeded the limit values at only 3 of the 22 stations.

Keywords: *Fuzzy c–means clustering, particulate matter, principal component analysis, sulfur dioxide*

Corresponding Author:
Senay Cetin Dogruparmak
☎ : *+90-262-3033194*
☎ : *+90-262-3033003*
✉ : *senayc@kocaeli.edu.tr*
senaycetin@hotmail.com

Article History:
Received: 29 January 2014
Revised: 14 April 2014
Accepted: 19 May 2014

doi: 10.5094/APR.2014.075

1. Introduction

Marmara Region is one of the major residential areas of Turkey, where industrialization resulted in an increase in population and road links. Due to its facilities, geographical situation, and ecological characteristics, it has been the focus of constant attention regarding industrialization, transportation and residential development. One of the environmental problems in the region is the pollutant emissions from industry, residential areas and traffic into the atmosphere. How the quantities and properties of these emissions are vary according to time, distance, and the influence of meteorological conditions must be followed significantly. Therefore a total of 39 air quality monitoring stations have been established in 11 provinces in the Marmara Region by the Ministry of Environment and Urbanization creating an air quality monitoring network (MEU, 2013). These stations have been established in 4 different categories, urban, traffic, industrial, and rural. There are differences in the measured parameters between different categories of monitoring stations. At the stations, the measured pollutants are: PM_{10}, SO_2, NO, NO_2, NO_X, O_3. In this study, SO_2 and PM_{10} were considered because these two pollutants are measured concurrently at most of the stations.

Major natural sources of SO_2 are volcanoes and oceans. On the other hand, anthropogenic emissions of SO_2 are produced by fossil fuel combustion (mainly coal and heavy oils), biomass burning and the smelting of sulfur containing ores. SO_2 and its oxidation by–products are removed from the atmosphere by wet and dry deposition (Pires et al., 2008). This results in the acidification of soils and surface waters with serious consequences for plant life and water fauna. Besides, buildings and cultural monuments are also damaged by acidification. Sulfate particles in the atmosphere are the largest source of haze and impaired visibility in many locations (Kone and Buke, 2012). SO_2 can be transported over large distances, causing transboundary pollution (Pires et al., 2008). Being an irritant, it causes human organ damages. It can affect the respiratory system and the functioning of the lungs, and causes irritation in the eyes (Ozbay, 2012). This pollutant also affects plants. Depending on its mass concentration levels, it can cause chlorophyll degradation; reduction of photo-synthesis; increased respiration rates; and changes in protein metabolism. On the other hand, PM is consisted of solid and liquid particles suspended in the atmosphere. They are emitted by both natural (volcanic eruptions, and forest fires) and anthropogenic sources (all types of man–made combustion and some industrial processes) (Pires et al., 2008). Similar to SO_2, the deposition of PM onto soils and surface waters can change their nutrient composition which has an effect on the diversity of ecosystems. PM is a significant contributor to reduced visibility (Kone and Buke, 2012). PM also has an adverse effect on human health. Extended exposures to PM_{10} and to $PM_{2.5}$ (particles with an aerodynamic diameter smaller than 2.5 mm) have been associated with respiratory and cardiovascular diseases (Pires et al., 2008). Consistent estimates of the relationship between daily variations in particulate matter and health effects have been provided by epidemiological studies. Inhalation of particulate matter is directly correlated with bronchitis symptoms and reduced lung function. An increase in PM_{10} mass concentration by 10 µg/m^3 results in a 5% increase in premature total mortality in case of lifelong exposure (Byrd et al., 2010).

SO_2 and PM_{10} concentrations should be monitored in the ambient air and the results should be interpreted in order to prevent their adverse effects. However, the number of monitoring stations in a zone should depend on the air quality of that zone. If it exceeds the requirements, the expenditures will increase. In order to determine whether a reduction can be made in the number of monitoring stations within the Air Quality Monitoring Network, this study focuses on grouping the monitoring stations sharing similar air pollution characteristics by using principal component analysis (PCA) and fuzzy c–means (FCM) methods. Multivariate statistical methods have been widely used in the studies conducted in recent years. As the number of analysis methods used in any study increases, the accuracy of the obtained results will be higher. Therefore, in this study, two different analysis methods were used. If studies on air quality monitoring stations are examined in detail, it is seen that PCA and CA methods are widely used (Abdalmogith and Harrison, 2005; Pires et al., 2008; Ibarra–Berastegi et al., 2009; Davis et al., 2009; Lau et al., 2009; Byrd et al., 2010; Lu et al., 2011). The difference between these two methods is, in PCA, each monitoring station is directly incorporated to a certain class, while in FCM, the extent to which a monitoring station should be included in both its own class and in other classes is determined. A literature search have shown that a comparative study like the present study has not been conducted previously. In addition, this study also enables determining the emission sources, evaluating the performances of the methods and examining the zone in terms of pollution.

2. Materials and Methods

2.1. Air quality monitoring network in Marmara Region

The Marmara Region, the selected research area, has an area of approximately 67 000 km^2. The study area has 11 provinces: Istanbul, Edirne, Kirklareli, Tekirdag, Canakkale, Kocaeli, Yalova, Sakarya, Bilecik, Bursa and Balikesir (MEF, 2010). In this study, data from air quality monitoring stations present in these cities were used. There are 22 monitoring stations in the study area: Istanbul (Aksaray, Alibeykoy, Besiktas, Esenler, Kadikoy, Kartal, Sariyer, Umraniye, Uskudar, Yenibosna), Kocaeli (City Center, Dilovasi, Organized Industrial Site–OSB), Sakarya, Yalova, Balikesir, Bilecik, Bursa, Canakkale, Edirne, Kirklareli and Tekirdag. Although there are 39 monitoring stations within the region, only 22 of them are measuring SO_2 and PM_{10} concurrently. The SO_2 and PM_{10} concentrations used in this paper are daily data obtained from 22 stations between 2008–2011 While applying PCA and FCM methods, daily data are used. On the other hand, annual averages are used while assessing the pollution at the zone of the air quality monitoring stations. The method used for the measurement of SO_2 concentrations is based on the principle of UV fluorescence; and the method used for the measurement of PM_{10} concentrations is ß–ray attenuation. The study area and the monitoring stations are shown in Figure 1 and the characteristics of the monitoring stations are given in Table 1.

2.2. Clustering methods

Clustering methods have been used in a variety of fields such as geology, business, engineering systems, medicine and chemistry (Linusson et al., 1998; Narayan et al., 2011; Ferraretti et al., 2012; Kannan et al., 2012; Yan et al., 2013). Clustering can be described as the optimal partitioning of *n* data into *c* subgroups, such that data that belong to the same group are as similar to each other as possible (Li and Shen, 2010). The objective of clustering is to find the data structure and also to partition the data set into groups with similar individuals. These clustering methods may be statistical, hierarchical, or heuristic (Pedrycz et al., 2004; Yang et al., 2004). In this study, one statistical analysis method, PCA, and one heuristic method, FCM, were used to evaluate the monitoring stations.

PCA. PCA, proposed by Pearson (1901), is a multivariate, statistical and exploratory analysis method. In this method, so–called principal components (PCs) are used to transform a set of interrelated variables into a set of uncorrelated variables (Pires et al., 2008; Lau et al., 2009; Moreno et al., 2009; Lu et al., 2011; Ozbay, 2012; Hu et al., 2013). These PCs are linear combinations of the original variables and are obtained in such a way that the first PC explains the largest fraction of the original data variability. The second PC explains a lesser fraction of the data variance than the first PC and so forth (Pires et al., 2008; Lau et al., 2009).

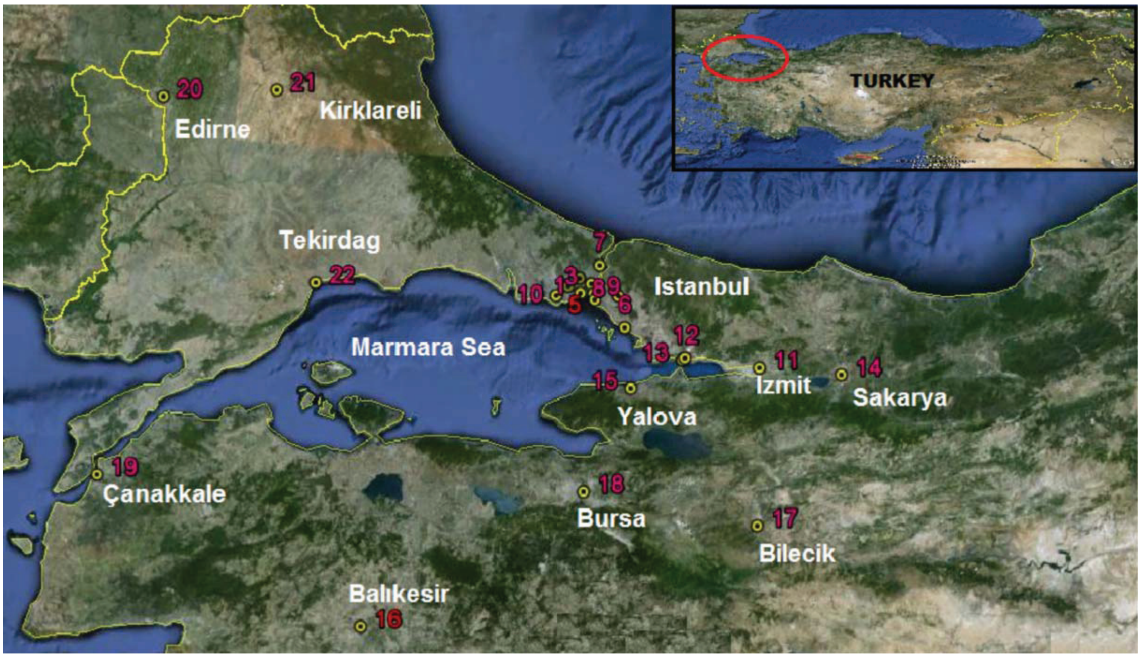

Figure 1. *Study area and sampling sites (satellite image by Google Earth).*

Table 1. *Characteristics of air quality monitoring stations in Marmara Region*

Station	Height above sea level (m)	Approximate distance to major roadways (m)	Approximate distance to residential areas (m)	Approximate distance to industries (m)
Istanbul–Aksaray	41	40	190	
Istanbul–Alibeykoy	6	30	100	
Istanbul–Besiktas	98	10	120	
Istanbul–Esenler	55	30	210	
Istanbul–Kadikoy	13	100	10	
Istanbul–Kartal	31	25	150	276
Istanbul–Sariyer	105	42	75	
Istanbul–Umraniye	154	170	250	
Istanbul–Uskudar	70	45	50	
Istanbul–Yenibosna	30	47	70	
Kocaeli–City Center	4	135	252	3 631
Kocaeli–Dilovasi	47	336	30	552
Kocaeli–OSB	30	135	100	421
Sakarya	42	10	30	806
Yalova	5	112	140	12 751
Balikesir	142	115	126	2 705
Bilecik	534	156	25	3 334
Bursa	91	158	425	484
Canakkale	9	17	25	
Edirne	41	20	35	
Kirklareli	204	102	76	
Tekirdag	26	25	25	800

In this method, first, a set of factors are derived from a data set by considering eigenvalues. In order to make the interpretation of the factors that are considered relevant, the first selection step is generally followed by a rotation of the factors that were retained. Varimax, developed by Kaiser, is the most popular rotation method. Obtained factor loads represent the contribution of each variable in a specific principal component. Principle components are computed by multiplying standardized data matrix with previously calculated weights (Ozbay, 2012). In this study, PCA was evaluated using Bartlett's sphericity test. These calculations were performed on original data by using SPSS 18 statistics program.

FCM algorithm. The classical clustering methods assign data to exactly one cluster. Since Zadeh proposed fuzzy sets described by a membership function, fuzzy clustering has been widely studied and applied in various areas (Yang and Wu, 2006). FCM, which is one of the most well–known and popular methodologies in clustering analysis, was introduced by Bezdek (1981), the origins of the algorithm tracing back to Dunn (Tsekouras and Sarimveis, 2004; Yang et al., 2004). Basically FCM clustering is dependent of the measure of distance between samples in a multi dimensional space. Mostly, FCM uses the common Euclidean distance which supposes that each feature has equal importance in the algorithm (Wang et al., 2004; Corsini et al., 2005). FCM algorithm aims to minimize the variance of the data within each cluster (Liao et al., 2003). Compared to the other clustering methods, FCM is more flexible because it shows those objects that have some interface with more than one cluster in the partition (Mingoti and Lima, 2006).

At FCM, the clusters are determined with respect to cluster numbers (c) that are defined by users and initial membership values for the input vector. The memberships of the clusters are defined with corresponding membership values. Also within the algorithm, clusters are described by prototypes which represent the cluster centers. It is an iteratively optimal algorithm based on the iterative minimization of the objective function in Equation (1).

$$J_m(U,V)=\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ki}^{m}\|x_k-v_i\|^2 \tag{1}$$

In Equation (1), n is the total number of data vectors in a given data set and c is the number of clusters; $X=(x_1, x_2,..., x_n)\subset R^S$ and $V=(v_1, v_2,..., v_c)\subset R^S$ are the feature data and cluster centres; and $U=(u_{ki})_{n*c}$ is a fuzzy partition matrix that is composed of the membership of each feature vector x_k in each cluster i. Here, u_{ki} should satisfy $\sum_{i=1}^{c}u_{ki}=1$ for k=1, 2,..., n and $u_{ki}\geq 0$ for all i=1, 2,...,c and $k=1$, 2,..., n. The exponent $m>1$ in Equation (1) is a parameter called fuzzifier. To minimize Equation (1), the cluster centers v_i and membership matrix U need to be calculated with regard to the following iterative formula:

$$u_{ki}=\left\{\begin{matrix}\left(\sum_{j=1}^{c}\left(\frac{\|x_k-v_i\|}{\|x_k-v_j\|}\right)^{\frac{2}{m-1}}\right)^{-1} \quad if\ \|x_k-v_j\|>0,\\ 1, if\ \|x_k-v_i\|=0,\\ 0, if\,\exists j\neq i\|x_k-v_j\|=0\end{matrix}\right\} \tag{2}$$
$$(\text{For } k=1,\ldots,n \text{ and } i=1,\ldots,c)$$

$$v_i=\frac{\sum_{k=1}^{n}u_{ki}^{m}\cdot x_k}{\sum_{i=1}^{N}u_{ki}^{m}}, i=1,2,\ldots,c \tag{3}$$

The procedure of the FCM algorithm is given below:

Step 1: Input the number of clusters c, the fuzzifier m and the distance function ‖ ‖.
Step 2: Initialize the cluster centers v_i^0 (i=1, 2,..., c).
Step 3: Compute u_{ki} (k=1, 2,..., n; i=1, 2,..., c) by using Equation (2).
Step 4: Compute v_i^1 (i=1, 2,..., c) by using Equation (3).
Step 5: If $max_{1\leq i\leq c}$ ($\|v_i^0-v_i^1\|/\|v_i^1\|)\leq\varepsilon$ then go to Step 6; else let $v_i^0=v_i^1$ (i=1, 2,..., c) and go to Step 3.
Step 6: Output the clustering results: cluster centers, v_i^1 (i=1, 2,..., c) membership matrix U and, in some applications, the

elements of each cluster *i*, i.e., all the x_k such that $u_{ki}>u_{kJ}$ for all $j≠k$.

Step 7: Stop (Sun et al., 2004).

MATLAB R2010b was used to cluster the monitoring stations by FCM that are mentioned above on original data.

3. Results and Discussion

3.1. Variations of SO_2 and PM_{10} in Marmara Region

The data derived from the annual averages of SO_2 and PM_{10} daily variations obtained from 22 air quality monitoring stations in the Marmara Region is shown in Figures 2 and 3.

When the SO_2 and PM_{10} pollution was evaluated for the period 2008–2011 (Figures 2 and 3), a reduction was observed in SO_2 concentrations at the Kocaeli–City Center station However, it was found that concentrations have increased at the Uskudar station (Figure 2). While PM_{10} concentrations have decreased at Alibeykoy, Besiktas, Canakkale stations, they have increased at the Sakarya station (Figure 3). Periods of both increasing and decreasing concentrations were observed at other stations from 2008 to 2011. The reason for the reduction in pollutant concentrations may be the strict controls on the coal entering to the cities. Increased SO_2 and PM_{10} concentrations, especially during the winter months, related to the high ash and sulfur content of coal, may have been prevented in this way. Additionally, increased use of natural gas as a fuel for residential heating and by the industry may be considered as another cause for the reduction in the concentrations of SO_2 and PM_{10}. On the other hand, the increased concentrations of some of the pollutants, may be due to adverse meteorological conditions or due to local sources.

When the measurement results are compared with the limit values in national and international regulations (Figure 2), SO_2 concentrations at all stations except those in Kocaeli-Dilovası, Çanakkale, Edirne, Kırklareli and Tekirdağ, are below USEPA's (U.S. Environmental Protection Agency) limit of 80 µg/m^3 and EU's (European Union) and the AQAMR's (Air Quality Assessment and Management Regulation) limit of 20 µg/m^3 (USEPA, 1996; EU, 2006; AQAM, 2008). When the stations exceeding the limit values are considered, Tekirdag station is found to be the one in most critical condition. The average values computed for all years of measurements, are above the limit values given in national and international regulations. Hence it is required to take immediate precautions in this area.

When the PM_{10} concentrations are compared with the limit values, the results are found to be different than those for SO_2 (Figure 3). PM_{10} concentrations in all stations exceed the WHO's (World Health Organization) limit value of 20 µg/m^3. And when the average of the years 2008–2011 for each station is considered, the PM_{10} limit value of 40 µg/m^3 set by EU and AQEMR was found to be exceeded at all stations. The limit value of 50 µg/m^3 set by U.S. EPA, was also exceeded at most of the stations. Hence when an overall assessment for the Marmara Region regarding to the PM_{10} pollution is conducted, it can be said that the limit values for PM_{10} concentrations are exceeded. Thus, the necessary precautions must be taken to reduce emissions.

In Turkey, for the evaluation of the air quality data within the scope of the EU accession process, the procedures given in Air Quality Assessment and Management Regulation (AQEMR), which was published in the Official Gazette No. 26898 dated 06.06.2008, are in effect. In this regulation, it is aimed to progressively reduce the national air pollution until 2014, and to ensure full compliance with EU limit values by then. So, when comparing the results with the national limit values, the target limit values given in AQEMR were taken into account instead of the limit values that are valid during the current transition period.

3.2. Clustering analysis results

PCA. The rotated factors obtained by PCA clustering of 22 different air quality monitoring stations are given in Table 2.

Table 2. *Results of PCA for SO_2 and PM_{10}*

Monitoring Stations	SO_2					PM_{10}		
	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 1	Factor 2	Factor 3
Aksaray	**0.478**	0.021	0.426	0.059	0.092	**0.681**	0.326	0.102
Alibeykoy	0.293	0.431	0.371	0.212	**0.545**	**0.764**	0.239	0.308
Besiktas	**0.806**	0.255	0.166	–0.081	0.109	**0.726**	0.153	0.188
Esenler	**0.863**	0.128	0.110	0.231	0.036	**0.432**	0.154	0.405
Kadikoy	**0.615**	0.161	0.067	0.231	0.067	**0.780**	0.387	0.053
Kartal	**0.768**	0.114	0.267	0.340	–0.090	**0.682**	0.182	0.285
Sariyer	**0.644**	0.262	0.097	0.189	0.400	**0.736**	0.177	0.296
Umraniye	0.332	0.205	0.242	**0.439**	0.243	**0.745**	0.431	0.087
Uskudar	0.603	0.192	0.229	0.084	**0.605**	**0.852**	0.361	0.072
Yenibosna	**0.817**	0.199	0.029	0.304	–0.182	**0.740**	0.304	0.156
Kocaeli–City Center	0.245	–0.008	**0.818**	–0.177	0.111	**0.676**	0.087	0.511
Kocaeli–Dilovasi	0.087	0.001	**0.882**	0.211	0.188	**0.736**	0.209	0.420
Kocaeli–OSB	–0.029	–0.202	0.424	0.073	**0.713**	**0.750**	0.138	0.394
Sakarya	0.357	**0.704**	0.324	0.163	0.175	0.296	**0.776**	–0.177
Yalova	0.234	**0.781**	–0.001	0.212	0.092	**0.489**	0.312	0.432
Balikesir	0.329	**0.604**	0.504	0.224	–0.003	**0.638**	0.251	0.447
Bilecik	0.098	0.346	**0.700**	0.297	0.104	0.345	**0.624**	0.185
Bursa	0.224	–0.103	0.155	**0.439**	–0.436	**0.605**	0.311	0.367
Canakkale	0.106	**0.846**	–0.080	0.247	–0.090	0.137	0.046	**0.831**
Edirne	0.205	0.339	0.175	**0.759**	0.137	0.200	**0.671**	0.483
Kirklareli	0.238	0.296	–0.039	**0.814**	–0.214	0.176	**0.738**	0.424
Tekirdag	0.225	0.267	0.080	**0.712**	0.347	0.455	0.434	**0.552**

The coefficient for each station in which factor loading has reached the highest value, are indicated in bold.

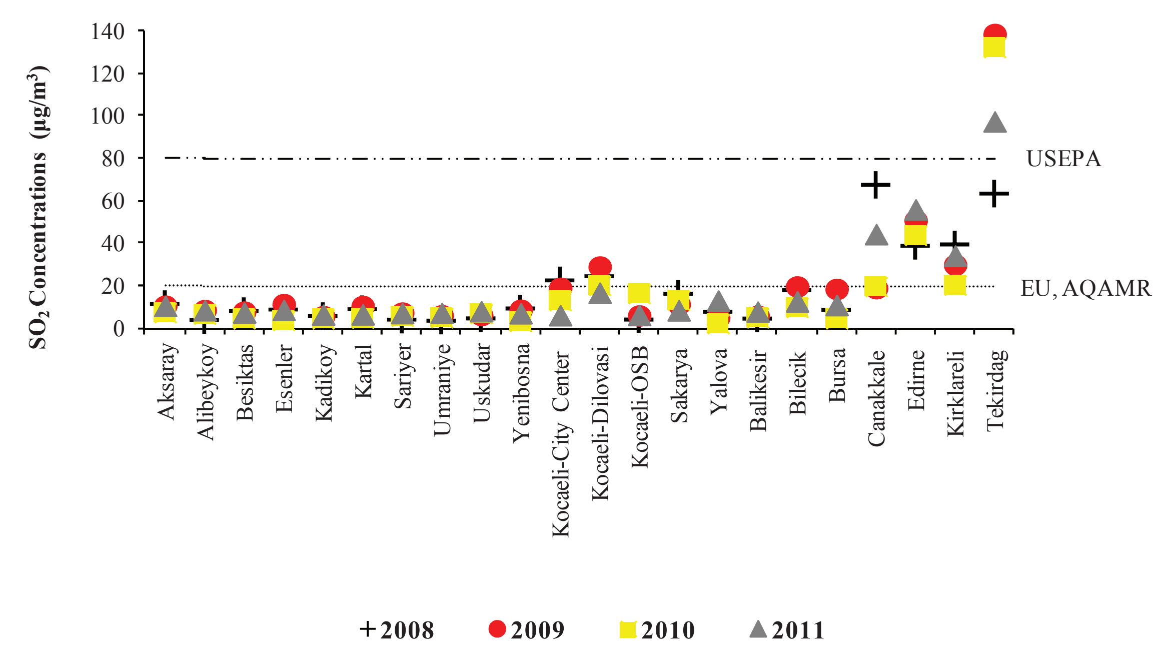

Figure 2. *Annual average SO$_2$ concentrations in Marmara Region for the period 2008–2011.*

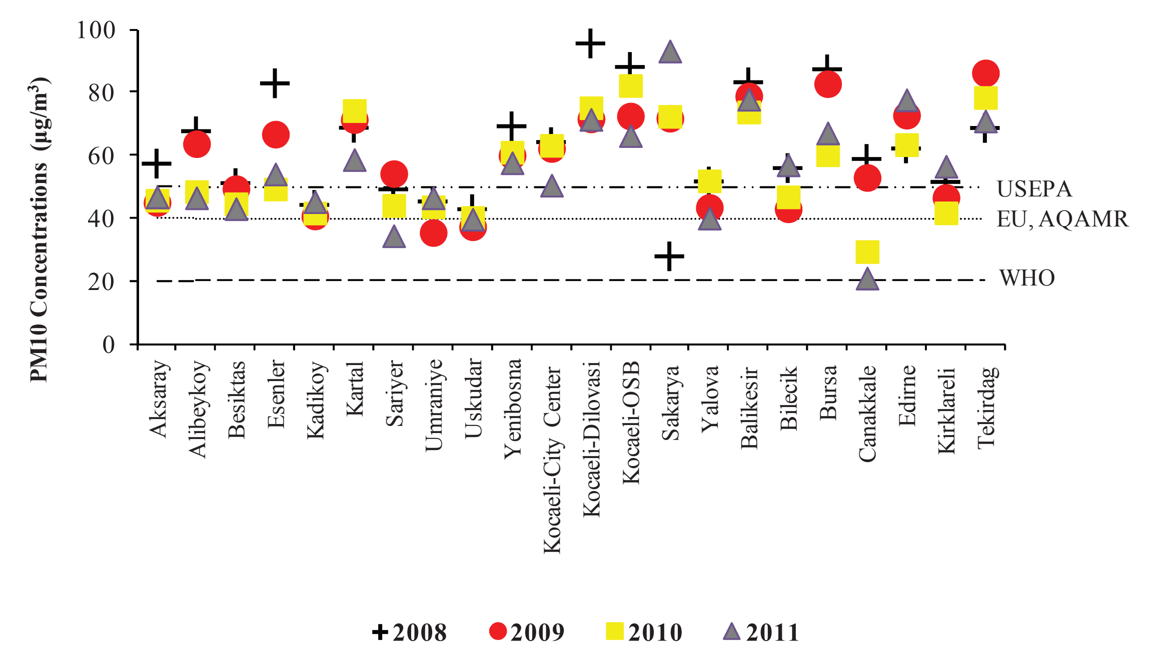

Figure 3. *Annual average PM$_{10}$ concentrations in Marmara Region for the period 2008–2011.*

After the classification of monitoring stations in the Marmara Region that exhibit similar behavior, 5 and 3 factor groups were obtained for SO$_2$ and PM$_{10}$, respectively. This result shows that the number of monitoring stations can be decreased. However, when reducing the number of monitoring stations, the factor loadings at the same cluster must certainly be considered. For example, when the factor loadings of Esenler and Yenibosna; and Esenler and Aksaray are compared to each other, the loadings of Esenler and Yenibosna are closer (Table 2). It means that these two stations have more similar air pollution characteristics. This shows that in case of necessity, the devices at one of the stations with closer factor loadings at the same cluster can be transferred to another zone for which there is a plan to establish a new station. By this way there may be a significant reduction in the investment and operational cost.

When analyizing factors, it is seen that factor 1 for SO$_2$ is Esenler, Yenibosna, Besiktas, Kartal, Sariyer, Kadikoy, Aksaray; factor 2 is Canakkale, Yalova, Sakarya, Balikesir; factor 3 is Kocaeli–Dilovasi, Kocaeli–City Center, Bilecik; factor 4 is Kirklareli, Edirne, Tekirdag, Bursa, Umraniye; and, factor 5 is Kocaeli–OSB, Uskudar, Alibeykoy. When these factors are analyzed separately, although it is difficult to determine the emission source type (point, areal, linear) that affects each group, a general assessment can be made. Most of the monitoring stations included in factors 1, 4, and 5, are influenced by the sources of areal+linear type. Kartal, Tekirdag, Bursa and Kocaeli–OSB stations included in these three groups are also exposed to point sources, besides areal+linear sources. When these factors are analyzed further, all the stations other than Bilecik and Bursa stations, are found to be within or close to the province of Istanbul. Factor 2 is the group with monitoring stations that are under the influence of point+areal+linear sources. Also, the stations in this group are located in the south and east of Marmara Region. Factor 3 is similar to factor 2, with stations effected by point+areal+linear sources. However, the difference between these two factors is that for all of the stations that are included in factor 3, all the sources are dominant. However, at the Canakkale station, which is one of the four stations within factor 2, areal+linear source are dominant, with no observed effect of the point sources. Therefore, when all the results are analyzed, it is difficult to make a clear distinction regarding the type of emission sources because there are ambiguities in the classification. There are studies in the literature that have also such uncertainties (Lau et al., 2009).

As a result of the PCA analysis for PM_{10}, factor 1 includes Uskudar, Kadikoy, Alibeykoy, Kocaeli–OSB, Umraniye, Yenibosna, Sariyer, Kocaeli–Dilovasi, Besiktas, Kartal, Aksaray, Kocaeli–City Center, Balikesir, Bursa, Yalova, Esenler; factor 2 includes Sakarya, Kirklareli, Edirne, Bilecik; and factor 3 includes Canakkale, Tekirdag. When these factors are examined separately, similar to the SO_2 monitoring stations, it is difficult to determine the type of emission source that affects the stations within each factor. However, a general evaluation can also be made. When factors 1, 2 and 3 are analyzed, it is not possible to make a distinction in terms of the type of emission sources. In each factor group, there are stations that are under the influence of both linear and point+ areal+linear sources. Therefore, in this respect, there is an uncertainty.

FCM algorithm. The monitoring stations that are classified for SO_2 using FCM are given in Table 3. Since 5 factors were obtained by the PCA analysis for SO_2, the monitoring stations were also classified in 5 clusters in FCM in order to be able to do a comparison. Although it is difficult to determine the type of affecting emission source for each cluster, when the results are compared with the PCA results, the uncertainty decreased slightly. When the clusters are examined separately, all of the monitoring stations in cluster 2 and cluster 5 are under the influence of stations with point+areal+line sources. clusters 1 and 4 are influenced by areal+linear sources. Although there are stations in cluster 3 that are under the influence of point+areal+linear sources, the number of stations under the influence of areal+linear sources is greater. Furthermore, all the monitoring stations in the province of Istanbul are found in this cluster.

At Table 3, Aksaray monitoring station belongs to cluster 1, cluster 2, cluster 3, cluster 4, and cluster 5 with membership degrees 0.0131, 0.0028, 0.5229, 0.0218, and 0.4395 respectively. In this case, this station is the member of cluster 3 with the highest membership degree of 0.5229. The membership degrees of other monitoring stations are determined similarly.

Since 3 factors were obtained by PCA analysis for PM_{10}, monitoring stations are classified in 3 clusters in the FCM as well. Results are shown in Table 4. It is relatively easier to determine the type of emission source affecting the monitoring stations here. When the clusters are examined separately, all monitoring stations that are in cluster 3 are under the effect of point+areal+linear sources. In cluster 1, Esenler, Yenibosna and Edirne stations are influenced by areal+linear sources. In this cluster, Sakarya and Bursa stations are influenced by point+areal+linear sources. In this analysis, the superiority of FCM compared to other clustering methods is evident. While the conventional classification methods indicate whether a certain set of data belongs to a certain class or not, the FCM method shows the membership of the data to each of the clusters, with the total membership being equal to 1. The membership values of the Sakarya and Bursa stations for cluster 1 are, as shown in Table 4, 0.3670 and 0.3929, respectively; their membership values for cluster 3 are 0.3470 and 0.3863, respectively. At this point, the decision makers may include these stations in cluster 3 since the membership values for the two clusters are very close. From this perspective, cluster 1 is completely under the influence of areal+linear sources, whereas cluster 3 is completely under the influence of point+areal+linear sources. Hence, when compared with PCA, the emission source type is more pronounced in the FCM clustering method. In cluster 2, all monitoring stations, except Yalova and Bilecik are influenced by linear+areal sources.

Table 3. *Clustered monitoring stations based on SO_2 by FCM*

Monitoring Stations	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster Membership
Aksaray	0.0131	0.0028	0.5229	0.0218	0.4395	3
Alibeykoy	0.0025	0.0006	0.9059	0.0041	0.0870	3
Besiktas	0.0022	0.0005	0.9236	0.0037	0.0700	3
Esenler	0.0026	0.0006	0.9114	0.0044	0.0810	3
Kadikoy	0.0040	0.0009	0.8756	0.0068	0.1127	3
Kartal	0.0042	0.0009	0.8452	0.0072	0.1426	3
Sariyer	0.0022	0.0005	0.9211	0.0037	0.0726	3
Umraniye	0.0038	0.0008	0.8785	0.0065	0.1104	3
Uskudar	0.0024	0.0005	0.9060	0.0040	0.0870	3
Yenibosna	0.0053	0.0012	0.8427	0.0093	0.1415	3
Kocaeli–City Center	0.0121	0.0025	0.2999	0.0185	0.6670	5
Kocaeli–Dilovasi	0.1337	0.0264	0.2581	0.1485	0.4334	5
Kocaeli–OSB	0.0175	0.0038	0.4608	0.0259	0.4919	5
Sakarya	0.0066	0.0014	0.7438	0.0111	0.2371	3
Yalova	0.0044	0.0010	0.8740	0.0074	0.1132	3
Balikesir	0.0032	0.0007	0.8711	0.0054	0.1195	3
Bilecik	0.0098	0.0020	0.2706	0.0164	0.7012	5
Bursa	0.0784	0.0186	0.3668	0.1427	0.3935	5
Canakkale	0.1935	0.0346	0.2147	0.3204	0.2367	4
Edirne	0.9890	0.0010	0.0019	0.0059	0.0022	1
Kirklareli	0.0215	0.0025	0.0102	0.9543	0.0115	4
Tekirdag	$8.70x10^{-6}$	0.9998	$3.89x10^{-6}$	$6.43x10^{-6}$	$4.21x10^{-6}$	2

Table 4. *Clustered monitoring stations based on PM_{10} by FCM*

Monitoring Stations	Cluster 1	Cluster 2	Cluster 3	Cluster Membership
Aksaray	0.1857	0.6882	0.1260	2
Alibeykoy	0.3340	0.4367	0.2293	2
Besiktas	0.2259	0.6150	0.1591	2
Esenler	0.3584	0.2940	0.3476	1
Kadikoy	0.2835	0.5114	0.2051	2
Kartal	0.3686	0.2066	0.4247	3
Sariyer	0.2299	0.6038	0.1663	2
Umraniye	0.2035	0.6571	0.1394	2
Uskudar	0.1426	0.7609	0.0966	2
Yenibosna	0.4026	0.2711	0.3263	1
Kocaeli–City Center	0.3599	0.1793	0.4608	3
Kocaeli–Dilovasi	0.3368	0.1544	0.5088	3
Kocaeli–OSB	0.3423	0.1605	0.4973	3
Sakarya	0.3670	0.2860	0.3470	1
Yalova	0.2445	0.5833	0.1722	2
Balikesir	0.3531	0.2049	0.4419	3
Bilecik	0.2759	0.5328	0.1913	2
Bursa	0.3929	0.2208	0.3863	1
Canakkale	0.2727	0.5165	0.2108	2
Edirne	0.3942	0.2257	0.3801	1
Kirklareli	0.3031	0.4881	0.2088	2
Tekirdag	0.3882	0.1768	0.4349	3

In Table 4, Aksaray monitoring station belongs to cluster 1, cluster 2, and cluster 3 with membership degrees 0.1857, 0.6882, and 0.1260 respectively. In this case this station is the member of cluster 2 with the highest membership degree of 0.6882. The membership degrees of other monitoring stations are determined similarly.

4. Conclusions

In this study, as a result of a PCA application, which is one of the methods used for determining the emission sources (point, areal, linear), the grouping of monitoring stations that show similar air pollution behavior within the Marmara Region, yielded 5 clusters for SO_2 and 3 cluster for PM_{10} from a total of 22 monitoring stations. This results show that the number of monitoring stations can be decreased. It is thought that reducing the 22 stations to 5 for SO_2 and 3 for PM_{10}, can affect determining the level of pollution in the zone. Therefore, to decrease the number of monitoring stations, the factor loadings in the same cluster must certainly be considered. Closer factor loadings show that the stations have similar air pollution characteristics. This shows that in case of necessity, the devices at one of the stations in the same cluster that have close factor loadings can be transferred to another zone where there is a plan to establish a new station. By this way there may be a significant reduction in the investment and operational costs.

When the clusters are analyzed for SO_2, it is difficult to make a clear distinction in terms of the dominant source type, because an evaluation shows that there are uncertainties in the classification. In addition, when the classes are analyzed for PM_{10}, it is seen that, in each factor group, there are monitoring stations that are influenced by areal+linear and point+areal+linear sources. Thus, there is uncertainty in the classification of PM_{10} sources as there is in SO_2 source classification.

The other method of analysis, the FCM algorithm, was run by reducing the total 22 monitoring stations to 5 classes for SO_2 and to 3 classes for PM_{10}, in order to facilitate comparison with the PCA. When the FCM results were compared with the PCA results, pollutant emission sources were more clearly identified in the FCM clustering method. So, when the performances of these two methods are evaluated, it can be said that FCM is superior to PCA.

When SO_2 and PM_{10} concentrations obtained from 22 monitoring stations in the Marmara Region are compared to the national and international limit values, only 3 stations exceeded the limit values for SO_2, whereas, PM_{10} concentrations are above the limit at all monitoring stations. Therefore, measures are needed to reduce the emissions at those residential areas that have monitoring stations with pollutant concentrations above the limit values.

References

Abdalmogith, S.S., Harrison, R.M., 2005. The use of trajectory cluster analysis to examine the long–range transport of secondary inorganic aerosol in the UK. *Atmospheric Environment* 39, 6686–6695.

AQAMR, 2008. The Ministry Of Environment and Urban Planning, R.G.S. 26898 R.G.T. 06.06.2008.

Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum, New York.

Byrd, T., Stack, M., Furey, A., 2010. The assessment of the presence and main constituents of particulate matter ten microns (PM_{10}) in Irish, rural and urban air. *Atmospheric Environment* 44, 75–87.

Corsini, P., Lazzerini, B., Marcelloni, F., 2005. A new fuzzy relational clustering algorithm based on the fuzzy C–means algorithm. *Soft Computing* 9, 439–447.

Davis, H.T., Aelion, C.M., McDermott, S., Lawson, A.B., 2009. Identifying natural and anthropogenic sources of metals in urban and rural soils using GIS–based data, PCA, and spatial interpolation. *Environmental Pollution* 157, 2378–2385.

Ferraretti, D., Gamberoni, G., Lamma, E., 2012. Unsupervised and supervised learning in cascade for petroleum geology. *Expert Systems with Applications* 39, 9504–9514.

Hu, S., Luo, T., Jing, C.Y., 2013. Principal component analysis of fluoride geochemistry of groundwater in Shanxi and inner Mongolia, China. *Journal of Geochemical Exploration* 135, 124–129.

Ibarra–Berastegi, G., Saenz, J., Ezcurra, A., Ganzedo, U., de Argandona, J.D., Errasti, I., Fernandez–Ferrero, A., Polanco–Martinez, J., 2009. Assessing spatial variability of SO_2 field as detected by an air quality network using self–organizing maps, cluster, and principal component analysis. *Atmospheric Environment* 43, 3829–3836.

Kannan, S.R., Ramathilagam, S., Devi, R., Hines, E., 2012. Strong fuzzy c–means in medical image data analysis. *Journal of Systems and Software* 85, 2425–2438.

Kone, A.C., Buke, T., 2012. A comparison for Turkish provinces' performance of urban air pollution. *Renewable & Sustainable Energy Reviews* 16, 1300–1310.

Lau, J., Hung, W.T., Cheung, C.S., 2009. Interpretation of air quality in relation to monitoring station's surroundings. *Atmospheric Environment* 43, 769–777.

Li, Y.L., Shen, Y., 2010. An automatic fuzzy c–means algorithm for image segmentation. *Soft Computing* 14, 123–128.

Liao, T.W., Celmins, A.K., Hammell, R.J., 2003. A fuzzy c–means variant for the generation of fuzzy term sets. *Fuzzy Sets and Systems* 135, 241–257.

Linusson, A., Wold, S., Norden, B., 1998. Fuzzy clustering of 627 alcohols, guided by a strategy for cluster analysis of chemical compounds for combinatorial chemistry. *Chemometrics and Intelligent Laboratory Systems* 44, 213–227.

Lu, W.Z., He, H.D., Dong, L.Y., 2011. Performance assessment of air quality monitoring networks using principal component analysis and cluster analysis. *Building and Environment* 46, 577–583.

MEF (Ministry of Environment and Forestry), 2010. Clean Air Action Plan (2010–2013), General Directorate of Environmental Management, Ankara.

MEU (Ministry of Environment and Urbanization), 2013. Press Bulletin About Marmara Region Clean Air Center Manager, Ankara, 1–4.

Mingoti, S.A., Lima, J.O., 2006. Comparing SOM neural network with Fuzzy *c*–means, *K*–means and traditional hierarchical clustering algorithms. *European Journal of Operational Research* 174, 1742–1759.

Moreno, N., Viana, M., Pandolfi, M., Alastuey, A., Querol, X., Chinchon, S., Pinto, J.F., Torres, F., Diez, J.M., Saez, J., 2009. Determination of direct and fugitive PM emissions in a Mediterranean harbour by means of classic and novel tracer methods. *Journal of Environmental Management* 91, 133–141.

Narayan, P. K., Narayan, S., Popp, S., D'Rosario, M., 2011. Share price clustering in Mexico. *International Review of Financial Analysis* 20, 113–119.

Ozbay, B., 2012. Modeling the effects of meteorological factors on SO_2 and PM_{10} concentrations with statistical approaches. *Clean–Soil Air Water* 40, 571–577.

Pearson, K., 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine* 2, 559–572.

Pedrycz, W., Loia, V., Senatore, S., 2004. P–FCM: A proximity – based fuzzy clustering. *Fuzzy Sets and Systems* 148, 21–41.

Pires, J.C.M., Sousa, S.I.V., Pereira, M.C., Alvim–Ferraz, M.C.M., Martins, F.G., 2008. Management of air quality monitoring using principal component and cluster analysis – part I: SO_2 and PM_{10}. *Atmospheric Environment* 42, 1249–1260.

Sun, H.J., Wang, S.R., Jiang, Q.S., 2004. FCM–based model selection algorithms for determining the number of clusters. *Pattern Recognition* 37, 2027–2037.

Tsekouras, G.E., Sarimveis, H., 2004. A new approach for measuring the validity of the fuzzy c–means algorithm. *Advances in Engineering Software* 35, 567–575.

USEPA, 1997. National Ambient Air Quality Standards for PM10 , http://www.epa.gov/ttn/naaqs/.

USEPA, 1996. National Ambient Air Quality Standards for SO2 , http://www.epa.gov/ttn/naaqs/.

EU, 2006. Air Quality Directives, http://www2.dmu.dk/atmosphericenvironment/aq_besk/eudir.pdf.

Wang, X.Z., Wang, Y.D., Wang, L.J., 2004. Improving fuzzy c–means clustering based on feature–weight learning. *Pattern Recognition Letters* 25, 1123–1132.

WHO, 2006. Air quality guidelines: global update 2005, WHO Regional Office for Europe, Copenhagen.

Yan, Y., Chen, L.H., Tjhi, W.C., 2013. Fuzzy semi–supervised co–clustering for text documents. *Fuzzy Sets and Systems* 215, 74–89.

Yang, M.S., Hwang, P.Y., Chen, D.H., 2004. Fuzzy clustering algorithms for mixed feature variables. *Fuzzy Sets and Systems* 141, 301–317.

Yang, M.S., Wu, K.L., 2006. Unsupervised possibilistic clustering. *Pattern Recognition* 39, 5–21.