Atmospheric Pollution Research 5 (2014) 656-663

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).*