Sigma Journal of Engineering and Natural Sciences, Vol. 42, No. 3, pp. XX-XX, June, 2024
Yildiz Technical University Press, Istanbul, Turkey

Vector autoregressive clustering for redundancy analysis in air pollution monitorıng networks at Türkiye

Aytaç PEKMEZCİ[1,*], Muhammet Oğuzhan YALÇIN[2]

ABSTRACT

This study proposes a new approach to reduce the information redundancy at Air Pollution Monitoring Networks (APMNs) and costs required for monitoring them. Proposed approach is based on Vector Autoregressive (VAR) model which describes the relationship between multivariate time series and consists of three main steps: In the first step, VAR model between two or more than two time series consisting of air pollutant observations is estimated. This step is repeated as the number of monitoring stations (n) under study and thus, n parameter vectors are obtained. In the second step, parameters vectors are divided into homogenous groups by using clustering analysis. The objective of this step is to identify the similar monitoring stations in terms of the relationship. Last step is to calculate the reduced information redundancy and the monitoring costs. To evaluate the efficiency of proposed approach, data sets consisting of PM10 and SO2 time series obtained from 116 APMNs at Turkey are used. Fuzzy K-Medoids (FKM) as clustering method Xie-Beni (XB) index as cluster validity index are preferred. Experimental results showed that information redundancy and monitoring cost in PM10 and SO2 stations can reduced at the rate of 63.36 by following proposed approach.

Keywords: *Air pollution, Information redundancy, Vector autoregressive models, Time series analysis.*

INTRODUCTION

Air pollution is presence of chemicals or compounds, in the atmosphere, at levels that effect negative on human and environment health. These chemicals or compounds are generally called as "air pollutant". Particulate Matter, Carbon Monoxide, Sulphur Dioxide (SO2), Carbon Monoxide (CO), Carbon Dioxide (CO2) and Nitrogen (N) are the most important air pollutants. Many studies have investigated the effects of the pollutants on human health and ecosystem (Ghorani -Azam et al., 2016; Kurt Kar et al., 2016; Liu et al., 2018; Landrigan et al., 2019). It is concluded that there is significant correlation between them. In order to minimize these effects, detecting of air pollution rapidly is considerable important. Air pollution monitoring network (APMN) is a main tool used for this objective. It provides an opportunity of giving correct information about air quality to public, evaluating the results of air pollution and taking precaution for protecting the environment and decreasing harmful effects of air pollution on creatures. But, APMNs require a lot of monitoring costs and need to expensive devices for monitoring. In this case, it becomes extremely important to decrease the costs required for APMNs. So far, many studies have been carried out for this objective. The most of these studies are based on detecting the stations having similar behavior in terms of an air pollutant via clustering analysis (Giri et al., 2006; Gramch et al., 2006; Lu et al., 2006; Morlini, 2007; Ignaccolo et al. 2008; Pires et al., 2008; D'Urso and Maharaj, 2009; D'Urso et al., 2015; Güler et al., 2016a, 2016b, Cotta et all., 2020). But in all of these studies, either one air pollutant is considered or analyses are carried out for each air pollutant separately and the relationship between air pollutants are not taken into account.

In this study, an approach is proposed for reducing monitoring cost in APMNs at Turkey for more than one air pollutant simultaneously. The proposed approach is based on clustering the parameters of the VAR model which indicates the relationship between air pollutants. In this way, it is aimed to get information about all air pollutants in

This paper was recommended for publication in revised form by Regional Editor Gülhayat Gölbaşı Şimşek
[1] Muğla Sıtkı Koçman University, Faculty of Science, Department of Statistics, Muğla, Türkiye
[2] Muğla Sıtkı Koçman University, Faculty of Science, Department of Statistics, Muğla, Türkiye
** E-mail address: aytac0803@yahoo.com*
Orcid id: https://orcid.org/0000-0003-4020-0069 Aytaç Pekmezci, 0000-0003-4017-5588 Muhammet Oğuzhan Yalçın
Manuscript Received 21 June 2022, Revised 20 September 2022, Accepted 07 November 2022

the model by only monitoring medoid (cluster centers) stations of air pollutant(s) selected as independent variable(s) and to decrease more the monitoring.

The organization scheme of this study can be given as follows. In Section 2, the material and methods used in this study are explained. Section 3 consists of and Section 4 concludes the study.

MATERIAL AND METHODS

This section briefly explains the proposed approach. In this study, the relationship between weekly PM10 and SO2 concentrations is considered. The data set are download from the website of http://laboratuvar.cevre.gov.tr/Default.ltr.aspx and each of which involves the period of between January 2018 and September 2021.

ESTIMATING VECTOR AUTOREGRESSIVE MODELS

VAR model is a statistical model used for investigating the relationships between two or more than two time series. VAR model between two number of time series can be defined as below:

$$y_t = \beta_0 + \sum_{i=1}^{p_1} \beta_{1i} y_{t-i} + \sum_{i=1}^{p_2} \beta_{2i} x_{t-i} + \varepsilon_{1t} \tag{1}$$

$$x_t = \alpha_0 + \sum_{i=1}^{p_1} \alpha_{1i} x_{t-i} + \sum_{i=1}^{p_2} \alpha_{2i} y_{t-i} + \varepsilon_{2t} \tag{2}$$

Where y_t and x_t are time series relating to different variables, p_1 and p_2 are lag length and ε_{1t} and ε_{2t} are error terms that follow normal distribution with zero mean and σ^2 variance.

The estimating of VAR model consists of several steps. These steps can be given as follows.

Step 1: Testing stationarity

VAR model assumes that all-time series (y_t, x_t) to be analyzed are stationary, i.e, statistical properties of time series such as mean, variance and covariance are all constant over time. In order to test stationary, unit root tests are used. These tests basically examine following hypothesis.

H_0: Unit root is present in time series
H_1: Unit root is not present in time series

Where hypothesis H_0 states that time series is nonstationary. In the literature, there exists many unit root tests. In this study, Augmented Dickey Fuller (ADF) (Dickey and Fuller, 1979) has been used. ADF test statistic is calculated as follows:

$$ADF_{statistic} = \frac{\beta_{11}}{SE(\beta_{11})} \tag{3}$$

Where $SE(\beta_{11})$ is standard error of β_{11}.

To decide whether time series is stationary or not, absolute value of ADF test statistic is compared with critical value (Dickey and Fuller, 1979). If this value is smaller than critical value, it is decided that time series is non-stationary. In that case, first-order difference of original time series is taken in order to make time series stationary. Unit root test is applied to differenced time series again and if differenced time series is still non-stationary, its second-order difference is taken. This process is repeated until time series become stationary. The number of taken difference indicates stationarity order of time series. For estimating VAR model, in the other words, for performing cointegration test, time series in the model must be stationary of same order.

Step 2: Determining lag length

The second step of estimating VAR model is the determination of lag length that refers to the number of previous values of time series ($y_{t-1}, y_{t-2}, \ldots, y_{t-p_1}, x_{t-1}, x_{t-2}, \ldots, x_{t-p_2}$). In this respect, many criteria have been employed in the literature. The most known criteria are Akaike Information Criterion (AIC), Schwarz Information Criterion (SIC) and Hannan-Quinn Criterion (HQC).

In this study, AIC and SIC criteria are preferred for selecting lag length. These criteria are calculated as follows:

$$AIC_p = -2n[\ln(\hat{\sigma}^2)] + 2p \tag{4}$$

$$SIC_p = nln(\hat{\sigma}^2) + n^{-1}pln(n) \tag{5}$$

Where n is length of time series and p is the lag length. $\hat{\sigma}^2$ in equations is computed as bellow

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} \varepsilon_t^2}{n - p - 1} \tag{6}$$

The working principle of lag length selection is as follows. VAR model is estimated for various lag length. The model errors (ε_t^2) and information criteria are calculated for these models. Finally lag length that provides smallest information criteria is selected.

Step 3: Cointegration test

Cointegration test investigates that whether a long-run relationship between the time series exists. Johansen (JH) and Engle-Granger (EG) are widely used cointegration tests. In JH test, $\pi = \sum_{i=1}^{p_1} \beta_i - I$ and $\tau_i = -\sum_{j=i+1}^{p_1} \beta_j$ firstly are calculated by using parameters of estimated VAR model.

The JH test includes examination of matrix π. Let rank of π be r. r gives the number of cointegrated time series. In here, three possible cases arise:

1. **r = n (the number of variables in the model)**: Time series are stationary at level.
2. **r = 0:** There are no cointegration between time series.
3. **r < n:** There exists r cointegrated time series.

With aim of detecting the number of cointegrated time series or whether cointegration exists or not, JH test uses two likelihood ratios known as trace test and maximum eigenvalue statistics. These statistics are calculated as follows:

$$\lambda_{trace}(r) = -n \sum_{i=r+1}^{p_1} \ln(1 - \lambda_i) \tag{7}$$

$$\lambda_{max}(r) = -nln(1 - \lambda_{r+1}) \tag{8}$$

Where λ_i is the estimate value of characteristics roots of the π matrix. These test statistics are compared to critical values tabulated by Osterwald-Lenum (1992). If the test statistics are larger than critical value, it is decided that cointegration exists. Besides, value of r gives the number of cointegrated time series.

Step 4: Granger causality test

Although cointegration indicates that time series have the long-run relationship, it does not give information about the direction of this relationship. Granger Causality (GC) test (Granger, 1969) is used for this kind of analysis. According to GC test, if previous value of X_t is useful in forecasting Y_t, the X_t series is Granger causes of Y_t or if previous values of Y_t is useful in forecasting X_t, the Y_t series is Granger causes of X_t. Following hypothesis are testing by using GC test:

$H_0: \beta_{21} = \beta_{22} = \cdots = \beta_{2p_2} = 0$ (**X_t series is not Granger causes of Y_t**)

$H_1: at\ least\ one\ of\ \beta_2\ coefficients \neq 0$ (**X_t series is Granger causes of Y_t**)

$H_0: \alpha_{21} = \alpha_{22} = \cdots = \alpha_{2p_2} = 0$ (**Y_t series is not Granger causes of X_t**)

$H_1: at\ least\ one\ of\ \alpha_2\ coefficients \neq 0$ (**$\mathbf{Y_t}$ series is Granger causes of $\mathbf{X_t}$**)

In order to test above hypothesis, F test should be calculated:

$$F = \frac{(RSS_R - RSS_u)/p_2}{RSS_u/(T-k)} \tag{9}$$

Where RSS_R is residual sum of squares relating to regression model consisting of only Y (Model 1), RSS_u is residual sum of squares for the regression model consisting of both Y variables and X variables (Model 2) and k is the number of parameters in Model 2.

The calculated F statistics is compared to critical value. If calculated F statistics is higher than critical value, it is decided that X_t series is Granger causes of Y_t. Then, VAR model defined in Eq. (1) is estimated**.** If second hypothesis is tested and if F statistics calculated for this hypothesis is higher than critical value, it is decided that Y_t series is Granger causes of X_t. Then, VAR model given in Eq. (2) is estimated.

Step 5: Estimating VAR model

VAR model is estimated by using Ordinary Least Squares (OLS) technique. If it is assumed that VAR model given in Eq. (1) is estimated, following equation is used:

$$\boldsymbol{\beta} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{Y} \tag{10}$$

Where $\boldsymbol{\beta}$ is parameter vector of VAR model. Matrix $\mathbf{Z}$ is defined as below:

$$\boldsymbol{Z} = [\boldsymbol{I}'\boldsymbol{Y}'_{t-1}\boldsymbol{Y}'_{t-2}\ \dots\ \boldsymbol{Y}'_{t-p_1}\boldsymbol{X}'_{t-1}\boldsymbol{X}'_{t-2}\ \dots\ \boldsymbol{X}'_{t-p_2}] \tag{11}$$

After estimating the parameter vector β, predicted values ($\widehat{\boldsymbol{Y}}$) are calculated by substituting values of β in Eq. (1)

Xie-Beni (XB) INDEX

Before applying clustering algorithm, optimal number of clusters should be determined. In this study, cluster validity indices proposed by XB (Xie and Beni, 1991) is preferred. XB index is based on two clustering criteria called as the compactness and separation. Let $\boldsymbol{\beta} = \{\boldsymbol{\beta_{11}}, \boldsymbol{\beta_{12}}, \dots, \boldsymbol{\beta_{1p_1}}, \boldsymbol{\beta_{21}}, \boldsymbol{\beta_{22}}, \dots, \boldsymbol{\beta_{2p_2}}\}$ be data set, where **$\boldsymbol{\beta}$s** are parameter vector of VAR model, $\bar{\beta}_{ij}s$, {i=1,2,…,c, j=1,2,…,$p_1 + p_2$} are j. component of i. cluster, $u_{ki}s$ {k=1,2,..,n, i=1,2,..,c} are fuzzy membership degree of k. monitoring station to i. cluster. The compactness and separation are calculated for XB index.

COMPACTNESS:

$$\boldsymbol{C} = \sum_{\mathrm{k=1}}^{\mathrm{n}}\sum_{\mathrm{i=1}}^{\mathrm{c}} \mathrm{u_{ki}}\left\|\beta_{ki} - \bar{\beta}_i\right\|^2 \tag{12}$$

Where n is number of monitoring stations and c is the number of cluster and ‖… ‖ is Euclidian distance.

SEPARATION:

$$\boldsymbol{S} = \mathrm{nx}\min_{i \neq j}\left\|\bar{\beta}_i - \bar{\beta}_j\right\|^2 \tag{13}$$

Based on compactness and separation criteria, XB index is given as follows:

$$\boldsymbol{XB} = \frac{\boldsymbol{C}}{\boldsymbol{S}} \tag{14}$$

XB index is calculated for all cluster numbers until predefined maximum number of clusters is reached and then the number of clusters, providing minimum XB index is set as the optimal number of clusters.

CLUSTERING AND FUZZT K-MEDOIDS ALGORITHM

Clustering analysis is a data mining technique used for dividing data set into groups such that data points within the same group are as similar as possible, whereas data points from different groups are as dissimilar as possible. These groups are called as cluster. Many clustering algorithms exist in the literature. This study uses FKM (Joshi and

Krishnapuram, 1999) clustering algorithm based on fuzzy clustering. In fact, the objective function in fuzzy clustering can be defined as follows:

$$\boldsymbol{J(U,\beta_k,\bar{\beta})=\sum_{k=1}^{n}\sum_{i=1}^{c}u_{ij}^{m}\left\|\beta_k-\bar{\beta}_i\right\|^2} \tag{15}$$

In FKM, cluster centers are called as the medoid ($\bar{\beta}_i$) which corresponds to data point of a cluster whose sum of distance to all the data points in the cluster makes minimal. The reason of choosing the FKM is to find a data point in the data set as cluster center. In FKM, medoid is calculated by the following equation:

$$\bar{\boldsymbol{\beta}}_{\mathbf{i}} = argmin_{1\le z\le n}\sum\nolimits_{t=1}^{n} u_{ti}^{m}\|\boldsymbol{\beta_z}-\boldsymbol{\beta_t}\|^2 \tag{16}$$

The update equation for membership degree (u_{ij}) is obtained as follows:

$$u_{ij}=\sum_{k=1}^{c}\left(\frac{\|\boldsymbol{\beta_i}-\bar{\boldsymbol{\beta}}_{\mathrm{i}}\|}{\|\boldsymbol{\beta_i}-\bar{\boldsymbol{\beta}}_{\mathrm{j}}\|}\right)^{\frac{-2}{m-1}} \tag{17}$$

The steps of FKM are as in Table 1.

Table 1. FKM algorithm.

Step 1: Entering initial values Number of clusters (c), initial medoids ($\bar{\boldsymbol{\beta}}_i$ i=1,2,…,c), fuzziness index (m), termination criteria (ε), iteration number iter = 1
Step 2 : Calculating membership degrees (u_{ij} i=1,2,..,c k=1,2,…,n) by using Eq. (17)
Step 3 : Increase iter by one iter = iter+1
Step 4 : Calculating new values of medoids ($\bar{\beta}_i$) by using Eq. (16)
Step 5 : If $\left|\bar{\boldsymbol{\beta}}^t-\bar{\boldsymbol{\beta}}^{t-1}\right|<\varepsilon$ iteration is terminated, otherwise go to Step 2.
Step 6 : Assign data points ($\boldsymbol{\beta_k}$) to clusters according to maximum membership degrees.

REDUNDANCY ANALYSIS

For redundancy analysis, below steps are followed.

- Stationarity test is separately performed for each of all PM_{10} and SO_2 time series. The first difference of non-stationary PM_{10} and SO_2 series are taken. Unit root test is re-applied and if the differenced time series is still non-stationary, its second-order difference is taken. This process is repeated until PM_{10} and SO_2 series become stationary. The number of taken difference indicates stationarity order of the time series. In order to estimate VAR model, PM_{10} and SO_2 series obtained from the same station must be stationary of same order. The next steps of the analysis are continued with same-order stationary stations. PM_{10} and SO_2 stations which are not same-order stationary are continued to be monitored. When the number of these stations is considered as n_1, the number of stations which are continued to be monitored is determined as $2xn_1$ (n_1 number of PM_{10} and SO_2) in this step.
- In the second step, PM_{10} and SO_2 time series having long-time relationship are determined by using Cointegration test. If the number of non-cointegrated series is equal to n_2, $2*n_2$ stations are continued to monitor at this step.
- Dependent and independent variables are determined by using GC test. $2*n_3$ stations with no causality relationship are continued to monitor. The parameters of VAR model are estimated for the stations remained.
- The number of cluster c is determined by using these parameters and FKM clustering algorithm is applied.

Lastly, below equation is used to calculate the percentage of the decreased monitoring cost (PDR-MC).

$$PDR-MC=\left(1-\frac{(2xn_1+2xn_2+2xn_3+c)}{2xN}\right)x100 \tag{18}$$

Where N is total number of monitoring stations, c is the number medoid stations to be monitored from air pollutant determined as independent variable.

RESULTS AND DISCUSSION

To identify stations that do not require to be monitored and thus reduce the monitoring cost, the procedure given in Fig1. is followed.

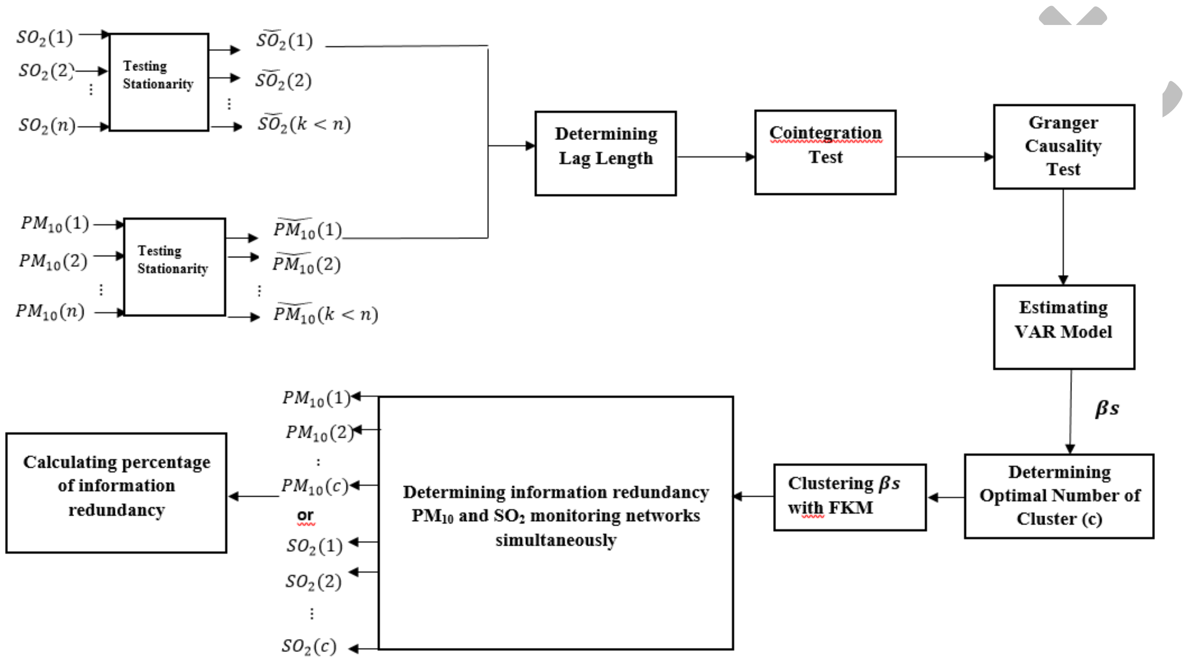

Fig. 1. The procedure followed in this study.

When the procedure given in Fig. 1 is followed, below results are found.

- According to ADF test, it is found that all PM_{10} and SO_2 time series are stationary of same order. Thus, the value of n_1 is determined as zero and no monitoring station has been eliminated at this step.
- The lag lengths are selected between 2 and 12.
- Since all time series are stationary at the level, no station is removed from the analysis ($(n_2 = 0)$).
- GC test arise that PM_{10} series are Granger cause of SO_2 series for 75 stations. According to this result, SO_2 series are selected as dependent variables and PM_{10} series are selected as independent variables in the VAR model. This means that PM_{10} concentrations can be used to estimate SO_2 concentrations.

Besides, 41 PM_{10} and SO_2 stations with no causality relationship are removed from the analysis and thus it is concluded that these stations are required to monitor. Table 2 shows these stations.

Table 2. The stations which continue to be monitored.

Name of Stations
Adana Meteoroloji, Aksaray, Amasya Suluova, Ankara Kayaş, Ankara Keçiören Sanatoryum, Ardahan, Aydın, Balıkesir Bandırma, Bilecik, Bilecik Bozüyük, Bingöl, Bursa Beyazıt, Çankırı, Çorum Mimar Sinan, Denizli Bayramyeri, Elazığ, Gaziantep, Isparta, İstanbul Başakşehir, İstanbul Esenyurt, İzmir Bornova, İzmir Güzelyalı, Karabük Kardemir2, Kayseri OSB Sanayi, Kırklareli, Kocaeli Körfez, Kocaeli Yeniköy, Konya Karatay, Mardin, Muş, Niğde, Sakarya, Samsun Tekkeköy, Şanlıurfa, Tekirdağ, Tekirdağ Çerkezköy, Tekirdağ Çorlu, Tokat Turhal, Tunceli, Yalova Armutlu, Zonguldak Çatalağzı

According to this, analyses are continued with 75 number of PM_{10} and SO_2 stations. VAR model denoted the relationship between SO_2 and PM_{10} is estimated for each of these stations. Table 3 shows the parameters of these models.

Table 3. The parameters of VAR models.

Stations	β_0	β_1	**Stations**	β_0	β_1
Adana Çatalan	0.0000	0.1349	**İzmir Çiğli İBB**	7.8732	0.1731
Adana Doğankent	8.9639	0.1267	**İzmir Gaziemir**	11.8157	0.0069
Adana Valilik	0.0000	0.1957	**İzmir Şirinyer İBB**	12.4022	-0.0429
Afyon	-2.3923	0.2966	**Kahramanmaraş Elbistan**	5.7135	0.1110
Ağrı Doğubeyazıt	0.0000	0.1748	**Karabük Kardemir 1**	18.0263	0.0913
Ağrı Patnos	-6.7359	0.3371	**Karabük Tören Alanı**	6.2590	0.3190
Amasya	5.6906	0.0549	**Karaman**	4.8504	0.0901
Amasya Merzifon	6.0510	0.0634	**Kars**	-3.9624	0.3660
Ankara Bahçelievler	0.0000	0.1017	**Kırıkkale**	3.2623	0.2634
Ankara Sıhhıye	8.1307	-0.0206	**Kırklareli Lüleburgaz**	0.0000	0.3811
Ankara Sincan	4.8046	0.0157	**Kırşehir**	5.5050	0.2001
Ankara Siteler	6.4285	0.0467	**Kocaeli**	0.0000	0.1141
Antalya	0.0000	0.0950	**Kocaeli Alikahya**	-1.3012	0.1603
Balıkesir	0.0000	0.2956	**Kocaeli Gebze**	5.9187	0.1010
Bartın	2.8879	0.1330	**Konya Karkent Sanayi**	3.9651	0.1016
Batman	4.9448	0.0236	**Konya Meram**	5.1902	0.1862
Bayburt	5.4535	0.0409	**Malatya**	9.0593	0.0343

Burdur	-2.0906	0.2611	**Muğla**	8.1977	0.2244
Bursa İnegöl	3.8774	0.1754	**Nevşehir**	0.0000	0.2433
Çanakkale	4.8993	0.0858	**Ordu Stadyum**	0.0000	0.4023
Çanakkale Biga İçdaş	0.0000	0.2449	**Ordu Ünye**	5.8899	0.0761
Çanakkale Can	0.0000	0.4177	**Osmaniye**	-2.7179	0.1931
Çorum	0.0000	0.2522	**Rize**	0.0000	0.0934
Denizli Merkezefendi	0.0000	0.2266	**Samsun Atakum**	0.0000	0.2677
Diyarbakır	2.2547	0.1108	**Samsun Bafra**	0.0000	0.2052
Düzce	0.0000	0.0588	**Samsun Canik**	0.0000	0.2708
Edirne	0.0000	0.1364	**Siirt**	9.2586	0.1588
Edirne Keşan	-130.4499	4.3573	**Sinop Boyabat**	7.1267	0.1688
Erzincan	0.0000	0.1293	**Sivas Meteoroloji**	11.3122	0.0941
Erzincan Trafik	0.0000	0.2073	**Tekirdağ Merkez**	-14.011	0.7658
Erzurum Aziziye	-3.8869	0.3773	**Trabzon Akçaabat**	0.0000	0.2154
Erzurum Palandöken	0.0000	0.1761	**Trabzon Fatih**	-0.7293	0.1392
Gümüşhane	0.0000	0.1375	**Van**	0.0000	0.4680
Hakkari	0.0000	5.7226	**Yalova**	0.0000	0.2985
Iğdır Aralık	4.8471	0.0160	**Yozgat**	0.0000	0.4033
İstanbul Kandilli	7.4947	0.1211	**Zonguldak Çatalağzı Kuzyaka**	-6.1723	0.3910
İstanbul Şirinevler	0.0000	0.1999	**Zonguldak Trafik**	0.0000	0.1533
İstanbul Ümraniye	0.0000	0.3959			

Table 4 provides the results of XB index.

Table 4. The results of XB index.

Number of Cluster	2	3	4	5	6	7	8	9	10
XB	9.98	**3.67**	4.83	8.54	15.38	177626.7	155900.7	138645.6	123923.8

According to Table 4, the optimal number of clusters is found as 3 since it has the smallest XB index. This states that there are three different groups in terms of the relationship between PM_{10} and SO_2 in Turkey. When FKM clustering algorithm with 3 number of clusters is applied to the parameters given in Table 3, the clusters given in Table 5 are constituted.

Table 5. Stations for each cluster and medoid stations.

Cluster Number	Stations
1	Adana Doğankent, Amasya, Amasya Merzifon, Ankara Sıhhıye, Ankara Sincan, Ankara Siteler, Bartın, Batman, Bayburt, Bursa İnegöl, Çanakkale, Diyarbakır, Iğdır Aralık, İstanbul Kandilli, İzmir Çiğli İBB, İzmir Gaziemir, İzmir Şirinyer, Kahramanmaraş Elbistan, Karabük Kardemir 1, Karabük Tören Alanı, Karaman, Kırşehir, Kocaeli Gebze, Konya Karkent Sanayi, Konya Meram, Malatya, Muğla, Ordu Ünye, Siirt, Sinop Boyabat, Sivas Meteroloji
2	Adana Çatalan, Adana Valilik, Afyon, Ağrı Doğubeyazıt, Ağrı Patnos, Ankara Bahçelievler, Antalya, Balıkesir, Burdur, Çanakkale Biga İçdaş, Çanakkale Çan, Çorum, Denizli Merkezefendi, Düzce, Edirne, Erzincan, Erzincan Trafik, Erzurum Aziziye, Erzurum Palandöken, Gümüşhane, İstanbul Şirinevler, İstanbul Ümraniye, Kars, Kırıkkale, Kırklareli Lüleburgaz, Kocaeli, Kocaeli Alikahya, Nevşehir, Ordu Stadyum, Osmaniye, Rize, Samsun Atakum, Samsun Bafra, Samsun Canik, Trabzon Akçaabat, Trabzon Fatih, Van, Yalova, Yozgat, Zonguldalk Çatalağzı Kuzyaka, Zonguldak Trafik
3	Edirne Keşan, Hakkari, Tekirdağ Merkez

The results obtained from Table 5 can be interpreted as follows:

- Clusters consist of 31, 41 and 3 stations respectively.
- When the results are interpreted according to cluster 3, it can be said that the stations Edirne Keşan, Hakkari and Tekirdağ Merkez have similar behavior in terms of the relationship between PM_{10} and SO_2. SO_2 and PM_{10} values of all these stations can be estimated by only monitoring Tekirdağ Merkez PM_{10} station. It is possible to interpret the other clusters similarly.
- Medoids are Kahramanmaraş Elbistan, Samsun Canik and Tekirdağ Merkez respectively and PM_{10} stations relating to the medoids should be continued to monitor.

From this, the percentage of the decreased monitoring cost can be determined as follows.

$$PDR-MC=\left(1-\frac{(2x0+2x0+2x41+3)}{2x116}\right)x100=63.36\%$$

As a consequence of analyses, the results obtained can be summarized as follows.

- The analyses are started with 116 number of PM_{10} and SO_2 monitoring stations at Turkey.
- In the first step of building VAR models, it is observed that all pairs of PM_{10} and SO_2 are stationary of same order. Thus, no monitoring station is eliminated in this step.
- In the second step of building VAR models, it is concluded that there exists long-term relationship between all pairs of PM_{10} and SO_2.
- According to the results of Granger Causality test, causality relationship between 41 pairs of PM_{10} and SO_2 are not found. It is decided that the monitoring should be continued for these stations. These stations are determined as Adana_Meteroloji, Aksaray, Amasya_Suluova, Ankara_Kayaş, Ankara_Keçiören, Ardahan,

Aydın, Balıkesir_Bandırma, Bilecik, Bilecik_Bozüyük, Bingöl, Bursa_Beyazıt, Çankırı, Çorum_Mimar Sinan, Denizli_Bayramyeri, Elazığ, Gaziantep, Isparta, İstanbul_Başakşehir, Kırklareli, İstanbul_Esenyurt, İzmir_Bornova, İzmir_Güzelyalı, Karabük_Kardemir 2, Kayseri_Sanayi, Kocaeli_Körfez, Kocaeli_Yeniköy, Mardin, Muş, Niğde, Sakarya, Samsun_Tekkeköy, Şanlıurfa, Tekirdağ, Tekirdağ_Çorlu, Tekirdağ_Çerkezköy, Tokat_Turhal, Tunceli, Yalova_Armutlu, Zonguldak_Çatalağzı Cumayanı.

- Granger Causality test revealed that PM_{10} concentrations are Granger cause of SO_2 concentrations. Therefore, VAR models are estimated such that dependent variables are SO_2 and independent variables PM_{10}.
- Xie-Beni index found that optimal number of clusters is equal to three. This means that there exist three groups which have different behavior in terms of the relationship between PM_{10} and SO_2 concentrations at Turkey.
- The parameters of VAR models estimated for 75 monitoring stations are clustered by using FKM algorithm. As a result of clustering, the stations that represent the clusters are found as Kahramanmaraş Elbistan, Samsun Canik, Tekirdağ Merkez.
- The number of stations to be monitored is found as 85 (41 PM_{10} +3 PM_{10}+41 SO_2). Thus, it is concluded that the monitoring cost and information redundancy at Turkey are decreased at rate of 63.36% by only monitoring 85 of 232 stations. This reduction in air monitoring cost means that the total cost (manpower, money, time, etc.) required for the future prediction values of PM10 and SO2 variables at each station is reduced.

So far, many studies have been carried out with aim of optimizing the number of APMNs and reducing the monitoring cost. But, in all of these studies, either only one air pollutant is considered, or analyses are carried out for each air pollutant separately. There is no study that takes into the relationship between air pollutants. This study proposes an approach based on the relationship between air pollutants for decreasing monitoring cost.

Thus, it is aimed to get information about the other air pollutants by monitoring medoid stations relating to only one air pollutant and aimed to reduce monitoring cost more. In this study, relationship between SO_2 and PM_{10} air pollutants are taken into account. It is possible that redundancy analysis approach proposed in this study is carried out for the other pollutants and the other regions.

The originality of this study is that while most of the studies in the literature were on a single variable, this study used more than one variable (PM10 and SO2). The contribution of this study to the literature is to propose a new approach based on the relationship between multiple air pollutants to reduce the information redundancy and the cost of monitoring at air pollution monitoring stations.

In future studies, it can be compared using other clustering algorithms other than Fuzzy K-Medoids clustering algorithm.

REFERENCES

[1] Cotta H, Reisen V, Bondon P, Prezotti P (2020) Identification of redundant air quality monitoring stations using robust principal component analysis, Environmental Modeling & Assessment, 25: 521-530.

[2] D’Urso P, Maharaj EA (2009) Autocorrelation-based fuzzy clustering of time series. Fuzzy Sets Syst, 160:3565–3589.

[3] D’Urso P, Giovanni LD, Massari R (2015) Time series clustering by a robust autoregressive metric with application to air pollution. Chemon Intell Lab Syst, 141:107-124.

[4] Dickey DA, Fuller W.A., 1979. Distribution of the estimators for autoregressive time series with a unit root, J am Stat Assoc, 74:427-431.

[5] Ghorani-Azam A, Riahi-Zanjani B, Balali-Mood M (2016) Effects of air pollution on human health and practical measures for prevention in Iran. J Res Med Sci, 21-65.

[6] Giri D, Murthy VK, Adhikary PR, Khanal SN (2006) Cluster analysis applied to atmospheric PM10 concentration data for determination of sources and spatial patterns in ambient air-quality of Katmandu Valley. Res Commun, 93(5):684-688.

[7] Gramch 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. Atmos Environ, 40:5464-5475.

[8] Granger CWJ (1969) Investigating causal relations by econometric models and cross-spectral methods. Econometrica, 37:424-438.

[9] Güler Dincer N, Yalçın MO (2016a) Revealing information and equipment redundancies in air pollution monitoring networks in Turkey. Int. J. Environ. Sci. Technol., 13:2927-2938.

[10] Güler Dincer N, İşçi Güneri Ö, Yalçın MO (2016b) Time series clustering's application to identifying information redundancy at air pollution monitoring stations in Turkey. Sakarya University Journal of Science, 20:605-616.

[11] Ignaccolo R, Ghigo S, Giovenali E (2008) Analysis of monitoring networks by functional clustering. Environmetrics, 62:672-686.

[12] Joshi A, Krishnapuram LYR (1999) A fuzzy relative of k-medoids algorithm with application to web document and snippet clustering. IEEE International Fuzzy Systems, Conference Proceedings, 1281-1286.

[13] Kurt Kar Ö, Zhang J, Pinkerton KE (2016) Pulmonary health effects of air pollution. Curr Opin Pulm Med, 22(2):138-143.

[14] Landrigan PJ, Fuller R, Fisher S, Suk WA, Sly P, Chiles TC, Bose-O'Reilly S (2019) Pollution and children's health. Sci Total Environ, 650(2):2389-2394.

[15] Liu H, Liu S, Xue B, Lv Z, Meng Z, Yang X, Xue T, Yu Q, He K (2018) Gorund-level ozone pollution and its health impacts in China. Atmos Environ, 173:223-230.

[16] Lu HC, Chang CL, Hsieh JC (2006) Classification of PM10 distributions in Taiwan. Atmos Environ, 40:1453-1463.

[17] Morlini I (2007) Searching for structure in measurements for air pollutant concentration. Environmetrics, 18:823-840.

[18] Osterwald-Lenum M (1992) A note with quantiles of the asymptotic distribution of the maximum likelihood cointegration rank test statistics. Oxf. Bull Econ Stat, 54:461-472.

[19] Pires JCM, Sousa SIV, Pereira MC, Alvim-Ferraz MCM, Martins FG (2008) Management of air quality monitoring using principal component and cluster analysis-Part I: SO2 and PM10. Atmos Environ, 42:1249-1260.

[20] Xie XL, Beni GA (1991) A validity measure for fuzzy clustering. IEEE Trans Pattern Anal Mach Intell, 13(8):841-847.