Determination of spatial and temporal changes in surface water quality of Filyos River (Turkey) using principal component analysis and cluster analysis

Elif Yağanoğlu1* • Aycan Mutlu Yağanoğlu2 • Gökhan Arslan3 • Adem Yavuz Sönmez4

1 Atatürk University, Agriculture Faculty, Department of Soil Science, Erzurum, Turkey
2 Atatürk University, Agriculture Faculty, Department of Animal Science, Erzurum, Turkey
3 Atatürk University, Fisheries Faculty, Department of Fisheries and Seafood Processing Technology, Erzurum, Turkey
4 Kastamonu University, Fisheries Faculty, Department of Basic Sciences, Kastamonu, Turkey

ARTICLE INFO

Article History:
Received: 24.08.2020
Received in revised form: 08.10.2020
Accepted: 09.10.2020
Available online: 13.10.2020

Keywords:
Filyos River
Surface water quality
Principal component analysis
Cluster analysis

ABSTRACT

Monitoring water quality is one of the high priorities for the protection of water resources. Many different approaches are used to analyse and interpret the variables that determine the variance of water quality observed in various sources. Statistical methods, especially multivariate statistical techniques, constitute an important part of these approaches. In this study, ten water quality parameters, which were measured for twelve months from seven stations determined on Filyos River, were evaluated by carrying out principal component analysis (PCA) and cluster analysis (CA) from multivariate statistical methods. In addition, dominant quality parameters designating the quality of the water source were determined. According to PCA results, 4 principal components contained the key variables and accounted for 69.49% of total variance of surface water quality from Filyos River. Dominant water quality parameters were observed to be temperature, EC, DO and pH. While the study revealed that the river is exposed to agricultural pollution alongside with the water quality character generated by the climatic conditions, it also suggested that multivariate statistical methods are useful tools in evaluating complex data sets such as water quality data, and monitoring the quality of water resources.

Please cite this paper as follows:

* Corresponding author
E-mail address: elifyaganoglu@atauni.edu.tr (E. Yağanoğlu)
Introduction

Monitoring water quality is one of the highest priorities of environmental conservation policy (Simeonov et al., 2002). Supplying water of high quality for purposes such as irrigation, drinking water, etc., and controlling and minimising problems caused by pollution are the principal objectives. Thus, nowadays, surface water quality constitutes one of the most significant determinants in resource management.

For this reason, monitoring water quality is a must in terms of water source management. Moreover, employing accurate methods in monitoring quality parameters is just as important.

Water quality can be defined as characterization of some parameters which represent a water composition in a specific place and time. Raw data are usually vast, meaning mostly they will not be distributed normally, they will be auto correlated or co-linear. Thus, multivariate analysis methods such as discriminant analysis, factor analysis, cluster analysis and principal component analysis are widely used in understanding spatial and temporal dissimilarities in water quality (Zeng and Rasmussen, 2005; Shrestha and Kazama, 2007).

Principal component analysis (PCA) is a data analysis method that is often used to decrease the number of variables of a large number of interrelated variables and also to keep as much variation (information) as possible. PCA is used to calculate an uncorrelated set of variables (pc’s or factors). These factors are put in an order so that most of the variations existing in original variables are retained.

Cluster analysis, on the other hand, is used to group objects within a class according to their similarities, and among different classes according to their dissimilarities (Panda et al. 2006). These similarities and dissimilarities are determined based on Euclidean and Manhattan distance measures (Kaufman and Rousseeuw, 1990).

Based on the fundamentals discussed above, in this study, we implemented principal component analysis (PCA) and cluster analysis for identifying practical pollution indicators in order to reveal agricultural and domestic pollution in Filyos River located in Western Black Sea Basin, Turkey. In addition, we also intended to provide a basis for future work on developing realistic tools that could help local decision-makers on the suitable management of the surface water quality in the river basin.

Material and Methods

Study Area

Filyos River is located on the southern west coast of Black Sea, Turkey. As shown in Figure 1, it flows through West Black Sea River Basin and disembogues into Black Sea at Filyos district of Zonguldak province. It has a drainage area of 13300 km², has two main branches: Yenice and Devrek streams, and has a length of 312 km (Kucukali, 2008; Sönmez et al., 2018). Gökcèbey and Çaycuma districts are located by the river and their populations are on the rise. Araç and Gerede streams are among its main branches. Recent industrial investments in the region such as paper and cement factories has changed economic structure while, economy of the local community was depending on forestry and agriculture formerly (Seker et al., 2005). Sönmez and Kale (2020) reported that the annual streamflow of the river tended to decrease particularly caused by climatic changes.

Figure 1. Sampling stations and study area

Delta of the river is wilderness and terrestrial ecosystem is quite rich and diverse as a consequence of topographical features of the region in which the river is located (Kucukali, 2014). Locations of the stations are indicated on the map (Figure 1) and a handheld GPS device was used in obtaining coordinates (Table 1).

Sampling Design and Collection

Samplings were carried out in duplicate and monthly from seven stations on Filyos River, Turkey between 2014 December and 2015 December. Samples were collected by using Nansen bottle and filtered through membrane filter with a 0.45 µm pore size, and stored in polyethylene bottles. Both polyethylene and Nansen bottles were rinsed with ambience water beforehand (Alam et al., 2001). Dissolved oxygen, turbidity, conductivity, pH and temperature parameters were measured with multiparameter in situ during sampling. On the other hand, spectrophotometric techniques were used while determining COD, BOD, phosphate, nitrite, nitrate and ammonium parameters in the laboratory (APHA, 2012).
Principal Component Analysis (PCA)

Principal component analysis (PCA) method explains the variance structure expressed by variables via correlations with new variables that are the components of original variables and are not intercorrelated. The number of base components is either equal to the number of original variables or less. When principal components are present, correlation matrix or variance-covariance matrix of the original variables is used. The basic components help analysing the dimension and deducing. Basic components are calculated based on the given formula in Equation 1 (Kuo et al. 2008; Mehat et al., 2014).

\[
Y_{mk} = \sum_{i=1}^{n} X_m(i).V_{ik}
\]  

Contribution of chemical and physical characteristics factor related to the base components are calculated by the square of the eigenvalue vectors. Weight values (wk) of the chemical characteristic variables are expressed as mean of the contributions of chemical characteristic factors acquired by squaring each eigenvalue vector (Mehat et al., 2014).

Cluster Analysis

Cluster analysis (CA) is used for exposition of multidimensional and large datasets as it is often seen in environmental data (Cieszynska et al., 2012). CA is a useful tool for grouping water samples resulting in high external (between clusters) heterogeneity and high interval (within clusters) homogeneity (Shrestha and Kazama, 2007). A widely used approach in analysing similarity between dataset and the sample is hierarchical agglomerative clustering (McKenna, 2003). Euclidean distance is a distance coefficient that is used to determine the similarity between two samples and a distance which can be represented by the difference between analytical values from both of the samples (Otto, 1998). Cluster analysis result is often presented as a tree like diagram (dendogram), that demonstrates the summary of clustering procedure with a considerable reduction in dimensionality of original data (Shrestha and Kazama, 2007).

In our study, hierarchical agglomerative CA was used on normalized data set through Ward’s method, and measure of similarity was Euclidean distances.

Results and Discussion

Descriptive Statistics and Correlations

Basic statistics were performed in order to give initial information related to water quality data. Details of the descriptive statistics of the water quality variables measured in twelve months are presented in Table 1.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Mean</th>
<th>Median</th>
<th>Mode</th>
<th>Std. Deviation</th>
<th>Minimum</th>
<th>Maximum</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>8.54</td>
<td>8.5</td>
<td>8.44</td>
<td>0.17</td>
<td>7.88</td>
<td>9</td>
</tr>
<tr>
<td>DO</td>
<td>9.57</td>
<td>9.59</td>
<td>8.21</td>
<td>1.17</td>
<td>7.73</td>
<td>11.83</td>
</tr>
<tr>
<td>EC</td>
<td>452.8</td>
<td>439.12</td>
<td>412.5</td>
<td>110.99</td>
<td>203.12</td>
<td>761</td>
</tr>
<tr>
<td>Temperature</td>
<td>14.71</td>
<td>13.38</td>
<td>6.97</td>
<td>6.19</td>
<td>4.85</td>
<td>26.02</td>
</tr>
<tr>
<td>Nitrite</td>
<td>0.06</td>
<td>0.02</td>
<td>0.01</td>
<td>0.1</td>
<td>0.01</td>
<td>0.75</td>
</tr>
<tr>
<td>Nitrate</td>
<td>0.72</td>
<td>0.6</td>
<td>0.3</td>
<td>0.62</td>
<td>0</td>
<td>4.95</td>
</tr>
<tr>
<td>Phosphate</td>
<td>0.8</td>
<td>0.43</td>
<td>0.26</td>
<td>0.8</td>
<td>0.04</td>
<td>2.92</td>
</tr>
<tr>
<td>Ammonium</td>
<td>0.38</td>
<td>0.17</td>
<td>0.02</td>
<td>0.62</td>
<td>0.02</td>
<td>3.2</td>
</tr>
<tr>
<td>COD</td>
<td>9.14</td>
<td>8.21</td>
<td>1.79</td>
<td>5.45</td>
<td>1.79</td>
<td>25.31</td>
</tr>
<tr>
<td>BOD</td>
<td>10.69</td>
<td>10.21</td>
<td>8.89</td>
<td>5.15</td>
<td>0.5</td>
<td>21.24</td>
</tr>
</tbody>
</table>

Descriptive statistics indicate that most of the parameters have high standard deviation and high change interval. Therefore, it can say that water quality in Filyos River has temporal and spatial dependence due to ongoing natural and anthropogenic processes in the basin (Gonzalez et al., 2014). When average values of water quality parameters compared with Water Pollution Control Regulation of Turkey, it is found that Filyos River partially suffers from oPO4^3-, NH4^+ and NO2^- pollution. This situation indicates that Filyos River has high organic pollution since nitrogen compounds in surface waters are usually related to organic pollution (Yang et al., 2007). Also, severe oPO4^3- pollution shows the impact of agricultural and domestic effluents (Wu, 2005).

According to correlation coefficients (Table 2), there were direct and statistically significant correlations between DO and BOD (0.515), EC and Ammonium (0.471), Nitrite and Nitrate (0.358), Phosphate and BOD (0.351), and Temperature and Nitrite (0.282). We can say that an increase in any of these variables positively affects the other. On the other hand, negative significant correlations were found between temperature and DO (-0.915), Temperature and BOD (-0.600), Nitrite and DO (-0.328), Ph and BOD (-0.283), and...
Ammonium and BOD (-0.218). It was shown in many studies that the DO level in water is inversely proportional to the temperature (Sönmez et al., 2008; Wang et al., 2013). There are also various studies which reported similar results regarding correlations with other parameters (Özgüler, 2001; Boyacıoğlu et al., 2005; Ustaoğlu and Tepe, 2018).

**Table 2. Correlations between dependent variables**

<table>
<thead>
<tr>
<th>Variables</th>
<th>pH</th>
<th>DO</th>
<th>EC</th>
<th>Temperature</th>
<th>Nitrite</th>
<th>Nitrate</th>
<th>Phosphate</th>
<th>Ammonium</th>
<th>COD</th>
<th>BOD</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>1</td>
<td>-.082</td>
<td>-.048</td>
<td>.141</td>
<td>-.114</td>
<td>-.105</td>
<td>-.153</td>
<td>.091</td>
<td>-.163</td>
<td>-.283**</td>
</tr>
<tr>
<td>DO</td>
<td>1</td>
<td>-.073</td>
<td>-.915**</td>
<td>-.328**</td>
<td>-.039</td>
<td>.013</td>
<td>-.073</td>
<td>-.049</td>
<td>.515**</td>
<td></td>
</tr>
<tr>
<td>EC</td>
<td>1</td>
<td>.070</td>
<td>.214</td>
<td>.176</td>
<td>-.052</td>
<td>.471**</td>
<td>-.021</td>
<td>-.108</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Temperature</td>
<td>1</td>
<td>.282**</td>
<td>.012</td>
<td>-.062</td>
<td>.056</td>
<td>.073</td>
<td>-.600**</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nitrite</td>
<td>1</td>
<td>.358**</td>
<td>-.052</td>
<td>.209</td>
<td>-.146</td>
<td>-.257**</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nitrate</td>
<td>1</td>
<td>.137</td>
<td>.005</td>
<td>-.063</td>
<td>.040</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phosphate</td>
<td>1</td>
<td>-.185</td>
<td>-.111</td>
<td>.351**</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ammonium</td>
<td>1</td>
<td>-.166</td>
<td>-.218'</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COD</td>
<td>1</td>
<td>.108</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BOD</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Table 3. Principal component analysis results**

<table>
<thead>
<tr>
<th>Component</th>
<th>Initial Eigenvalues</th>
<th>Extraction Sums of Squared Loadings</th>
<th>Rotation Sums of Squared Loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Total</td>
<td>% of Variance</td>
<td>Cumulative %</td>
</tr>
<tr>
<td>1</td>
<td>2.717</td>
<td>27.174</td>
<td>27.174</td>
</tr>
<tr>
<td>2</td>
<td>1.627</td>
<td>16.271</td>
<td>43.445</td>
</tr>
<tr>
<td>3</td>
<td>1.461</td>
<td>14.609</td>
<td>58.054</td>
</tr>
<tr>
<td>4</td>
<td>1.144</td>
<td>11.439</td>
<td>69.493</td>
</tr>
<tr>
<td>5</td>
<td>.899</td>
<td>8.991</td>
<td>78.484</td>
</tr>
<tr>
<td>6</td>
<td>.765</td>
<td>7.653</td>
<td>86.173</td>
</tr>
<tr>
<td>7</td>
<td>.502</td>
<td>5.022</td>
<td>91.159</td>
</tr>
<tr>
<td>8</td>
<td>.447</td>
<td>4.470</td>
<td>95.629</td>
</tr>
<tr>
<td>9</td>
<td>.361</td>
<td>3.612</td>
<td>99.242</td>
</tr>
<tr>
<td>10</td>
<td>.076</td>
<td>.758</td>
<td>100.000</td>
</tr>
</tbody>
</table>

**Principal Component Analysis**

Eigenvalues of each principal component are given in Table 3. The components with an eigenvalue greater than 1 were considered as significant. Eigenvalues of principal components were found to be 2.717, 1.627, 1.461, 1.144, 0.899, 0.765, 0.502, 0.447, 0.361 and 0.076, respectively. First four components explained 69.49% of total variation in the data set while, 27.174% of total variance was explained by the first factor; 16.271% by the second factor; 14.609% by the third factor; and 11.439% by the fourth factor. Remaining six components were found to explain 30.51% of total variation, while these were found to be insignificant.

Contribution of the chemical characteristics corresponding to the principal components is shown by the eigenvalue vectors. Contributions of the chemical characters calculated by squaring each eigenvalue vector are given in Table 4. For instance, the contribution weight of water parameters was 0.973 for the first, -0.950 for the second, 0.786 for the third, and -0.632 for the fourth component. In our study, first four main components explained 69.49% of the total variance. An eigenvalue provides information on measure of the significance of a factor. In addition, eigenvalues greater than 1 are accepted as significant (Kim and Mueller, 1987; Muangthong and Shrestha, 2015). According to Liu et al. (2003), loadings greater than 0.75 are considered as strong, loadings between 0.50-0.75 as moderate and loadings smaller than 0.30-0.50 as weak.
Table 4. Eigenvectors of the principal component analysis and contribution of each individual chemical characteristic for the principal component

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Component 1</th>
<th>Component 2</th>
<th>Component 3</th>
<th>Component 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>-.373</td>
<td>-.632</td>
<td></td>
<td></td>
</tr>
<tr>
<td>DO</td>
<td>-.950</td>
<td>.786</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EC</td>
<td>.973</td>
<td>.711</td>
<td>.543</td>
<td>.611</td>
</tr>
<tr>
<td>Temperature</td>
<td>.372</td>
<td>.321</td>
<td>.543</td>
<td>.813</td>
</tr>
<tr>
<td>Nitrite</td>
<td>-.442</td>
<td>.611</td>
<td></td>
<td>.850</td>
</tr>
<tr>
<td>Nitrate</td>
<td>.813</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Phosphate</td>
<td>-.311</td>
<td>.850</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Ammonium</td>
<td>-.702</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In the first component, the negative correlation of temperature with DO and BOD points out the natural process (Kükrer and Mutlu, 2019). Temperature values are inversely proportional to the solubility of oxygen in the water (Shrestha and Kazama, 2007; Atea, et al., 2017; Abdelali et al., 2018).

The second component constitutes 16.27% of the total variance. In this component, EC and Ammonium were found to have a strong positive correlation. EC values reveals the presence of electrolytic contaminants and dissolved salts. However, it does not give information regarding the specific ion composition (Adekunle et al., 2007). Positive strong ammonium value revealed the presence of nutrient pollution caused by agricultural effluent (Sing et al., 2005). Positive values of Nitrate, Nitrite and Phosphate in the third component also showed the effect of agricultural activities on the water source.

In the fourth component, pH exhibited negative results while, COD exhibited positive results. Mostly, biological and chemical reactions depend on pH value. In addition, it determines the metal ion solubility thus, effecting aquatic natural life (Hamed, 2019).

Figure 2 shows the graphical spatial representation of provided factors for chemical parameters. In this graph, the grouping of parameters and their correlation with the maintained factors can be observed.

**Cluster Analysis**

K-means algorithm is implemented in order to procure generalized cluster characteristics by using dominant parameters according to optimum number of clusters that is determined by FCM in the previous step. First, medians of clusters are found, thereafter the clusters are formed by assigning each object from the dataset to the nearest cluster medians. Dissimilarities of each object in the dataset from these centers of the clusters are evaluated by Euclidean distance. Cluster centers are chosen according to minimum distance. For validation and interpretation of clusters, silhouette is used (Kaufman and Rousseeuw, 1990).

![Figure 2. Component plot in rotated space](image)

![Figure 3. Lower-dimensional projection of the predictor space, which contains a total of 10 predictors](image)

Since the similarities between DO and BOD, and EC and Ammonium were very strong, these variables formed a group at a distance of 1 and 2 units, respectively. Nitrite and Nitrate were also very similar parameters. However, their distance was 6 units in the dendogram. While similarity distance of oxygen and phosphate was 13 units; ammonium was included to the group of nitrite, nitrate and pH at 14 units. At a distance of 14 units, pH and Temperature were similar, whereas Phosphate is merged with the group. BOD and COD came together at a distance of 20 units and phosphate was similar to nitrite at a distance of 25 units. The findings obtained in the cluster analysis results are in line with various similar studies (Kilic and Yucel, 2019; Hamed, 2019; Kükrer and Mutlu, 2019).
results characterized the pollution parameters in Filyos River and their relationship with each other.

**Figure 4.** Dendrogram using average linkage (between groups)

**Conclusion**

In this study, multivariate statistical techniques were used to identify spatial and temporal changes in water quality of Filyos River. Principal component analysis revealed that four principal components were able to explain 69.49% of the variability. The dominant water quality parameters were found to be temperature, DO, pH and EC. These indicators have shown that the river is under climatic and environmental pressure, especially when the flow rate is low. In addition, Ammonium, Nitrate, Nitrite and Phosphate parameters, which are in positive interaction under these dominant components, showed that Filyos River suffers from agricultural pollutant sources. Results of basic correlation, factor and cluster analyses applied to 10 physico-chemical water parameters measured from Filyos River for 12 months have substantially supported each other and have emerged to reveal both spatial and temporal pollution characteristics of the river. Moreover, it was shown that multivariate statistical techniques are effective in the investigation of water quality datasets. To be able to succeed effective water resources management, similar works should be conducted frequently in large water resources.

**Acknowledgements**

This study was supported by The Scientific and Technological Research Council of Turkey (TÜBİTAK Project No: 114O145).

**Compliance with Ethical Standards**

**Authors’ Contributions**

AYS: Collection of data, analysis and writing of the article. EY: Analysis and article writing MY: Statistical analysis. GA: Analysis of samples and article writing

**Conflict of Interest**

The authors declare that there is no conflict of interest.

**Ethical Approval**

For this type of study, formal consent is not required.

**References**


