Baseline

Chemometric evaluation of the heavy metals distribution in waters from the Dilovası region in Kocaeli, Turkey

Deniz Bingöl a,*, Ümit Ay a, Seda Karayünlü Bozbaş a, Nevin Uzgören b

Keywords: Heavy metals, Principal component analysis, Correlation analysis, Cluster analysis, Chemometrics

A R T I C L E   I N F O

ABSTRACT

The main objective of this study was to test water samples collected from 10 locations in the Dilovası area (a town in the Kocaeli region of Turkey) for heavy metal contamination and to classify the heavy metal (Cr, Mn, Co, Ni, Cu, Zn, As, Cd, Pb and Hg) contents in water samples using chemometric methods. The heavy metals in the water samples were identified using inductively coupled plasma-mass spectrometry (ICP-MS). To ascertain the relationship among the water samples and their possible sources, the correlation analysis, principal component analysis (PCA), and cluster analysis (CA) were used as classification techniques. About 10 water samples were classified into five groups using PCA. A very similar grouping was obtained using CA.

In Turkey, a developing country, environmental pollution problems have increased since 1960 due to the rapid growth of industry and population increase in the Marmara region, specifically in İzmit Bay. Since the 1960s, more than 250 large industrial plants have been built in the area surrounding the bay. Industrial activities in the region are mostly located along the northern coast of İzmit Bay and surrounded by residential neighborhoods (Demiray et al., 2012). With its 50,000 inhabitants, the Dilovası district of Kocaeli is a symbol of Turkey’s uncontrolled industrialization. There are 185 companies serving in 45 sectors mainly metal (iron–steel, aluminum), chemistry (e.g., paint), and energy (coal-fired electric power plant), in the Dilovası organized industrial zone, which is located at the center of a bowl-like topographic structure. This caused serious environmental problems in the region, for example, soil, air, and water pollution. Although serious measures were taken to reduce and control pollution since the beginning of 1990, pollution levels are still high in the region (Karademir, 2006). In Dilovası, deaths caused by cancer have surpassed those caused by cardiovascular diseases, becoming the leading cause of death (Tuncer, 2009). Dilovası’s sewer system is directly connected to Dil Creek. This waste is not subject to any purification. Dil Creek is located in the eastern Marmara region and discharges into İzmit Bay. This water source is used for irrigation and as drinking water for animals. One of the most important sources of pollution in İzmit Bay is Dil Creek, which flows into the western part of the bay. An estimated 60% of the total waste water directly enters İzmit Bay (Telli-Karakoç et al., 2002). Large industrial plants (mainly paint and metal industries) around Dil Creek to discharge their solid and liquid waste into Dil Creek after limited treatment.

In this area, İzocam, DYO, Lever, Wishes, Olmuksa, Yazıcıoğlu, Polisan and Çolakoğlu factories generate the most stream pollution. Factories such as the Porland porcelain factory are able to pollute the river and creek from a distance thanks to hundreds of meters of pipes. Fifteen years ago, Dilovası was home to cherry, apple, and peach orchards and vineyards; however, unfortunately, the Dil Creek pollution destroyed the orchards and vineyards. Waste is discharged directly into the stream without a water treatment system causing the extinction of many species. In the past, people swam and fished in Dil Creek’s clean water. Heavy metals are deemed serious pollutants because of toxicity, persistence, and non-degradability in the environment (Fang and Hong, 1999; Klavins et al., 2000; Tam and Wong, 2000; Yuan et al., 2004). Over the past century, heavy metals have been discharged into the world’s rivers and estuaries as a result of rapid industrialization (Chen et al., 2004; Cobelo-Garcia and Prego, 2003; Pekey, 2006; Tam and Wong, 2000).

Multivariate methods are being increasingly used because a large amount of information can be compared in graphical form, which is very difficult to do using number tables or univariate statistics. There are many well established multivariate methods for classification, of which the most commonly used are correlation analysis, principal component analysis (PCA), and cluster analysis (CA) (Brereton, 2007). Some researchers have determined similarities between samples and groups of samples using PCA and CA. For example, researchers can use multivariate methods to evaluate...
trace metal concentrations in some spices and herbs (Karadas and Kara, 2012); to the assessment of the level of some heavy metals in sediments (Idiris, 2008), to determine the levels of essential, trace and toxic elements in citrus honeys from different regions (Yücel and Sultanoglu, 2012); to determine trace elements in commonly consumed medicinal herbs (Tokaloğlu, 2012); to evaluate the mineral content of medicinal herbs (Kolasani et al., 2011); to evaluate trace metal concentration in some herbs and herbal teas (Kara, 2009); to evaluate heavy metals in street dust samples (Tokaloğlu and Kartal, 2006); to determine concentrations of key heavy metals in street dust and analyze their potential sources (Lu et al., 2010); to identify heavy metals in pastureland (Franco-Uría et al., 2009); to identify source of eight hazardous heavy metals in agricultural soils (Cai et al., 2012); to classify sea cucumber according to region of origin (Liu et al., 2012a); and to evaluate the heavy metal contamination of surface soil (Yaylalı-Abanuz, 2011).

The aim of this study was to apply the chemometric techniques of correlation analysis, principal component analysis (PCA), and cluster analysis (CA) to results obtained from inductively coupled plasma-mass spectrometry (ICP-MS) of water samples, and to identify similarities in heavy metal content.

An ICP-MS inductively coupled plasma-mass spectrometry instrument (Perkin Elmer DRC-e/Cetax ADX-500) was used to determine Cr, Mn, Co, Ni, Cu, Zn, As, Cd, Pb, and Hg content in each region. A Hanna pH 211 Microprocessor pH-meter was used to measure the pH values of the solutions. The pH-meter was standardized with NBS buffers prior to each measurement. The pH-meter was calibrated according to region of origin (Liu et al., 2012a); and to evaluate the heavy metal contamination of surface soil (Yaylalı-Abanuz, 2011). The aim of a correlation analysis is to measure the relationship between variables. Pearson’s correlation coefficient ($r$ for sample) is the most common correlation coefficient. The correlation coefficients can range from $-1$ to $+1$ and are independent of the units of measurement. Usually, $|r| > 0.75$ indicates that there is a significant relationship between the variables. In this study, first the relationships between variables are examined using Pearson’s correlation coefficient. In the statistical analysis, a significant correlation among the variables is not required. In such cases, the correlations should be removed from the data set. The $p$ number of relevant variables can be expressed as the $k$ number ($k < p$) of new artificial variables, which are linear components of these variables do not correlate within them. This function performs the PCA (Özdamar, 2002).

PCA is probably the most widely used multivariate statistical technique used in chemometrics, and because of the importance of multivariate measurements in chemistry, it is regarded by many as the technique that most significantly changes chemist’s view of data analysis. Exploratory data analysis such as PCA is primarily used to determine general relationships between data. Sometimes, more complex questions need to be answered; for example, do the samples fall into groups? The aims of PCA are to determine underlying information from multivariate raw data (Brereton, 2007). Additional interpretations between heavy metals and water samples may be obtained using more powerful chemometric techniques such as PCA. PCA is a projection method that allows easy visualization of all the information contained in a data set. PCA helps determine differences between samples and identifies which variables contribute most to this difference (Liu et al., 2012b). PCA is a process that transforms components of data matrix $A$ ($A_{n \times p}$), including $n$ samples and $p$, as shown in the following equation:
where $T$ is $n \times q$ score matrix, and $B$ is $q \times p$ PCA loading matrices. $q$ gives the minimum number of principal components needed for a PCA analysis of matrix $A$. Each column vector of matrix $T$ and each row vector of matrix $B$ is considered one principal component of matrix $A$ (Diraman et al., 2009). In PCA, the information carried by the original variables is projected onto a smaller number of underlying (“latent”) variables called principal components. The first principal component covers as much of the variation in the data as possible, the second principal component is orthogonal to the first and covers as much of the remaining variation as possible, and so on (Kara, 2009). Because the first and second principal component usually covers a large portion of the total, a clustering of samples, according to the effect of all variables within the two-dimensional plane, is possible by plotting against each of the first two column vectors (the first two principle components: PC1 and PC2) of matrix $T$. For a grouping that depends on the distribution of variables in the system, the first two rows of matrix $B$ is plotted against each other (Diraman et al., 2009). If the first m principal component describes a large portion of the total variance, the rest of the $p-m$ principal component can be neglected. In this case, there is a small variance (information) loss, and the work-space size is reduced to $m$ from $p$ ($m < p$) (reduction of dimension) (Tatlıdil, 1992). Principal components ($Y_i$) are independent, and their variances are equal to the eigenvalue ($\lambda_i$) of correlation matrixes. The total variance of the original system is equal to the total variance of principal components:

$$A = T \times B + E_A$$

$$\lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \text{Var}(Y_i).$$

The total variability of the data matrix is equal to the total variability of principal components:

$$\lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \text{Var}(Y_i).$$

The variability ratio explained by the kth principal component

$$= \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \quad (k = 1, 2, \ldots, p).$$

In applications, a few principal components describes a proportion larger than 80% of the total variance, without causing a loss of information that can substitute for the original variable (Ersungur et al., 2007). The number of eigenvalues greater than one value when using standardized data matrix gives the value of $m$ (Tatlıdil, 1992).

PCA allows users to see the results by themselves rather than as a result of a property provider because the principal components are capable of an intermediate step for more extensive investigations. In particular, CA uses principal component scores, which are fairly common conditions (Özdamar, 2002).

The purpose of CA is to organize observations of a number of groups/variables and determine if they share observed properties. PCA is used primarily to determine general relationships between data. CA is often coupled with PCA to check results and to group individual parameters and variables. A dendrogram is the most commonly used method of summarizing hierarchical clustering (Lu et al., 2010). This technique is an unsupervised classification procedure that involves a measurement of the similarity between objects to be clustered.

The average of three results and standard deviations of analyses obtained using ICP-MS are shown in Table 1. Using the data in Table 1, the heavy metals and water samples were classified using correlation analysis, principal component analysis, and cluster analysis.

In the first stage of this study, the findings obtained by calculating descriptive statistics of selected variables were interpreted (Table 2). In the second stage, correlation analysis was applied to determine whether the principal component analysis is appropriate to standardize data sets (Table 3). The results of the analysis showed that there are significant correlations between the variables (heavy metals). Therefore, the principal component analysis is appropriate. In the third stage of study, the principal component analysis was applied and the eigenvalues, which belong to the first four components, were found to be greater than one and explained 91% of the total variance (Table 4). The water samples were classified in accordance with the findings, graphics were drawn, and variables (heavy metals) that are effective in this classification were determined. In the final stage of the study, water samples were classified using the Ward algorithm with a hierarchical clustering analysis (HCA). HCA was used to assess the spatial similarity or dissimilarity in water samples according to heavy metals (Malik and Nadeem, 2011). HCA was carried out using the first four principal components, explaining 91% of the total variance.

The findings obtained by calculating the descriptive statistics of the selected variables were given in Table 2. Descriptive statistics for the data set of 10 variables were analyzed. For all the variables, positive and strong asymmetry (skewness > 0.5) was found (i.e., the observation values collected had relatively smaller values than the average). The kurtosis values show that the series, except for Cu and As, is sharper than normal, which means that the collection of smaller than average values is higher than normal. These results show that the series is not normally distributed. In addition, the standard deviation values were very close to the average values for many variables (Mn, Ni, Cu, Zn, and As), and some were even found to be higher (Co, Cd, Pb, and Hg). This situation indicated that the observation values were significantly different than the mean (i.e., the variability was very high). This means that the water samples taken from different locations were differently characterized by their heavy metal content.

### Table 1

<table>
<thead>
<tr>
<th>Water samples</th>
<th>Average (µg/L ± StDev*)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cr</td>
</tr>
<tr>
<td>d1</td>
<td>0.027 ± 0.07</td>
</tr>
<tr>
<td>d2</td>
<td>0.023 ± 0.08</td>
</tr>
<tr>
<td>d3</td>
<td>0.020 ± 0.09</td>
</tr>
<tr>
<td>d4</td>
<td>0.263 ± 0.27</td>
</tr>
<tr>
<td>d5</td>
<td>0.071 ± 0.04</td>
</tr>
<tr>
<td>d6</td>
<td>0.081 ± 0.05</td>
</tr>
<tr>
<td>d7</td>
<td>0.058 ± 0.04</td>
</tr>
<tr>
<td>d8</td>
<td>0.027 ± 0.05</td>
</tr>
<tr>
<td>d9</td>
<td>0.010 ± 0.04</td>
</tr>
<tr>
<td>d10</td>
<td>0.015 ± 0.03</td>
</tr>
</tbody>
</table>

* StDev: Standard deviation.
In multivariate statistical analyzes, the statistical analysis must be used to standardize data instead of the original data when the measurement units and variability of variables investigated were different. However, the standard deviation values for the study were quite different. The original data matrix was standardized using the following equation: 

\[ z = \frac{X - \mu}{\sigma} \]

where \( X \) is the original value, \( \mu \) is the mean, and \( \sigma \) is the standard deviation. In the PCA analysis, the correlation matrix was used for the analysis. The Person’s correlation coefficients for 10 heavy metals are presented in Table 4. The positive and negative correlation coefficients indicate positive and negative correlations respectively, between the two metals. A significantly positive correlation at \( p < 0.01 \) was found between the heavy metal pairs Ni–Zn (0.907), Co–As (0.875), Cr–Pb (0.841), and Mn–Hg (0.91). In addition, Mn is positively correlated with Cu at \( p < 0.05 \), and Cd is not correlated to the other elements. There are statistically significant and high correlations between variables. Therefore, the application of PCA to the data set is significant in eliminating the dependence structure and/or reducing size.

Two pieces of information are connected, namely, geography and concentration. So, in many areas of multivariate analysis, one aim may be to connect the samples (e.g., geographical location/sampling site), which are represented by the scores to the variables (e.g., chemical measurements), which are represented by loadings (Brereton, 2007).

PCA was applied to the entire data set (Table 4). Due to the standardized data, the correlation matrix was used for the analysis. The PCA results are summarized in Table 4. The principal components that have eigenvalues higher than one were extracted. The results indicate that there were four eigenvalues higher than one. The first component explains 29.7% of the total variance and loads heavily on Ni and Zn. The second component, dominated Mn, Cu and Hg, accounts for 25.5% of the total variance. The third component is loaded by Co, Cu, and As, accounting for 22.4% of the total variance. The fourth component is dominated by Cr, Ni, Zn, and Pb, accounting for 11.4% of the total variance.
PCA scores according to the first principal component as determined by Ni and Zn content, and the second principal component, as determined by Mn, Cu, and Hg, are listed in Table 5. The water samples listed in the top formed a cluster, and the water sample was listed in the bottom when listing by the first main component. The d1 water sample was listed at the top with a large score difference over the second component in Mn, Cu, and Hg content. According to the analysis of the d1 water sample, these three heavy metals are highly effective in this location. By plotting the principal component, the inter-relationships between different variables can be viewed, and interpreted for sample patterns, groupings, similarities, or differences (Kara, 2009). The first and second principal component usually includes a large portion of the total variance; therefore, the first two principle components (PC1 and PC2) are plotted against each other, and clustering of samples is possible in the effects of all variables within the two-dimensional plane. The PC1 and PC2 score vectors and the PC1 and PC2 loading vectors from PCA are plotted against each other in Figs. 2 and 3, respectively. The first two components explain 57.2% of the variation in the data set.

When Figs. 1 and 2 are evaluated, d9 and d10 samples are shown to be characterized by Ni and Zn metals; the d1 sample is only characterized by Cu, Hg, and Mn metals; and the d4 sample is only characterized a Cr metal a cluster. The d2 and d3 water samples were mainly characterized by Co metal and Cu, Co, Mn, and Hg metals in low levels. The d6 sample was situated near the center and was slightly affected by Cd, As and Pb metals. Similarly, the d5, d7, and d8 samples were situated close to the center and were not characterized by any metal.

A biplot involves a superimposition of scores and a loadings plot, with the variables and samples represented on the same diagram. Fig. 4 shows a graph of the PC1, PC2, and PC3 score vectors from PCA against each other. The first three components explain 79.6% of the total variation in the data set.

Fig. 4 was plotted according to the first three main components, which showed that water samples taken from different locations gave similar results to the results of a two-dimensional graph and usually consisted of five groups.

Group 1: d1.
Group 2: d2 d3.
Group 3: d9 d10.
Group 4: d5 d6 d7 d8.
Group 5: d4.

At this stage of the study, the aim is to make HCA and the classification of the water samples from different locations, and to compare the findings obtained with the results obtained from the PCA. HCA was applied to the score vectors obtained from PCA. The score vectors of the first four principal components, explaining 91% of the total change, were used in the analysis. The measurement is based on squared Euclidean distance. In this study, the clustering method used was the Ward linkage method. Dendrogram obtained from the Ward linkage method is shown in Fig. 5.
Fig. 5. The Dendrogram results show that the appropriate number of clusters was five, which was similar to the PCA results.

Chemometric methods were applied to classify water according to heavy metal contents. The chemometric evaluation showed that a relationship exists between their heavy metal contents and the water samples from the Dilovası area. The water samples and heavy metals were classified into five groups by PCA and a cluster analysis. From the chemometric evaluation of heavy metal content, the first group contains only the d1 sample, the second group contains the d2 and d3 samples, the third group consists of the d9 and d10 samples, the fourth group is composed of the d5, d6, d7 and d8 samples, and the fifth group only contains the d4 water sample.

References


