TECHNICAL TRANSACTIONS 12/2017
CZASOPISMO TECHNICZNE 12/2017
ENVIROMENTAL ENGINEERING
DOI: 10.4467/2353737XCT.17.218.7761

Marek Kubala (qmq@vistula.pk.edu.pl)
Institute of Water Supply and Environmental Protection, Faculty of Environmental Engineering, Cracow University of Technology

The usefulness of cluster analysis in the analysis of data obtained in the monitoring of the water environment

Przydatność analizy skupień w analizie danych pozyskanych w monitoringu środowiska wodnego

Abstract
Data obtained through the monitoring of the water environment often includes a number of indicators, and is frequently collected from a large area or over a long period of time. Analysis of such data can be problematic. The division of elements which have a certain degree of similarity into subgroups may facilitate data analysis and provide indications as to the direction of the analysis. One tool for the separation of such groups of similar elements is cluster analysis.

This paper describes the two most commonly used cluster analysis algorithms and summarises the results of several applications of cluster analysis in water monitoring.

Keywords: water environment monitoring, cluster analysis, hierarchical clustering, k-means clustering

Streszczenie
Dane monitoringu środowiska wodnego zawierają często pomiary wielu wskaźników, a także bywają zbierane z dużego obszaru czy w długim okresie. Analiza takich danych może być utrudniona. Podział ich na podgrupy, których elementy wykazują pewne podobieństwo, może przyczynić się do łatwiejszej analizy oraz dostarczyć przesłanek co do jej kierunku. Jednym z narzędzi wydzielania takich grup podobieństwa jest analiza skupień. Praca przedstawia opis dwóch najczęściej stosowanych algorytmów analizy skupień oraz streszcza rezultaty kilku zastosowań analizy skupień w monitoringu środowiska wodnego.

Słowa kluczowe: monitoring środowiska wodnego, analiza skupień, metody hierarchiczne, algorytm k-średnich

1. Introduction

Monitoring the water environment is often a matter of the collection and analysis of data (indicators) obtained at many points across a large area or over a long period of time. In addition, in many cases monitoring is not limited to measuring only one indicator but operates in a multidimensional metric space. In such cases, the relationship between them may be difficult to establish.

Generally, complex environmental processes may depend on many factors that are not always recognised. Measuring only selected indicators can lead to revealing completely different relationships between them in one area than in another area. Similarly, measurements conducted over a long period of time may lead to situations in which one set of data is influenced by different processes occurring in the environment leading to completely different dependencies than in another set of data. If such disparate spatial or temporal areas are not properly isolated, the correlation analysis between the different indicators in the set will, at best, not produce any positive results; at worst, this leads to the drawing of erroneous conclusions. Even if data diversity is expected, a strict spatial or temporal criterion may not be available to separate the relevant data groups. In addition, in the case of measurements with too little spatial density, it may not be possible to separate zones with different characteristics, especially when the different zones are mutually intertwined and one measurement point lies in one zone.

The first approach in this situation may be the graphical analysis of data, such as the analysis of the distribution of a selected indicator. If its distribution is characterised by several distinctly separated modes, this gives the basis for exploring the reasons for such a distribution [1]. A similar analysis of the density of measurements can be made either in two-dimensional space (on the basis of two indicators) or three-dimensionally. For larger dimensions, this analysis is no longer possible. In addition, usually the groups are not clearly separated, or the data set contains too few measurement points to make the division clear. In this situation, appropriate tools are needed to support the decision-making process with regard to division. Such tools include methods of cluster analysis.

2. Cluster analysis

Cluster analysis, or clustering, is the process of dividing a set of objects into subsets (clusters) in such a way that objects belonging to the same cluster are somehow more similar to one another than to objects from other clusters. This task is not defined by a unique algorithm, and its choice may depend not only on the type of data we want to analyse but also on the choice of the cluster definition itself. This term is usually understood as referring to a group of objects that are not too far apart, or areas of data space with a particular density of measurement points or intervals of statistical distribution. The appropriate clustering algorithm depends on the type of data being analysed and the clustering goal. This also applies to parameters such as the distance function, the density threshold and the number of expected clusters. Therefore, cluster analysis is not only limited to the automatic application of an appropriate algorithm but also involves a gradual decision-making process in subsequent stages of its implementation, starting with the choice of variables in which the data set is divided (if the data set is multidimensional), and finally evaluating the real value of obtained division [1].

Based on several typical cluster models, the following main algorithms can be distinguished: hierarchical clustering algorithms, centroid algorithms, and algorithms based on the spatial density of objects and statistical distribution functions [2]. We will discuss the first two types of algorithms.

3. Hierarchical clustering

Hierarchical cluster analysis is designed to build a hierarchy of clusters with different levels of fragmentation. Following the method of building the hierarchy, the algorithms are divided into agglomerative and divisive types. Agglomerative algorithms start from the level where each observation has its own cluster and gradually, progressing up the hierarchical ladder, the individual clusters are joined together. By contrast, divisive algorithms take the lead from one cluster which is the set of all observations and the next steps of the hierarchy depend on successive divisions; however, they are not universally available and rarely applied. This is necessary to determine which clusters should be connected in the next step (if the agglomeration method is used). To accomplish this, the clustering criterion and the metric that will calculate the distance between clusters must be selected. There are three main criteria: complete linkage, single linkage, and average linkage.

The complete-linkage criterion (also called the diameter or the maximum method) interprets the distance between clusters as the maximum distance from any element of one cluster to any element of the other cluster. By contrast, the single-linkage criterion (also called the connectedness or minimum method) interprets the distance between clusters as the minimum distance from any element of one cluster to any element of the other cluster. Finally, the average linkage criterion takes on the average value of distances between elements of clusters. A variant of this criterion is the median criterion, which is more proof to outliers.

To calculate the distance, one has to measure it. The most commonly chosen measurements are the Euclidean distance, the square Euclidean distance, the taxicab distance and the maximum distance. The Euclidean distance is simply the straight-line distance between two points. This value can be squared, as the square Euclidean distance, to place greater weight on objects that are further apart. The taxicab distance is the sum of absolute differences between the coordinates of two points. The maximum distance is defined as the maximum value of the absolute differences between the coordinates of two points.

All these distance measures are sensitive to the scale of the values of the coordinates of the points. In order to eliminate the dominance of a parameter with high values relative to others, the parameters involved in calculating the distance should be normalised. Another way to solve this problem is to replace the Euclidean distance with the Mahalanobis distance [3].

Having already selected the linking criterion and the metric, we use the following algorithm [4]:

1) Assign a separate cluster to each object. In the beginning, there are as many clusters as there are points. In this case, the distance between clusters is equal to the distance between points.
2) Identify the two nearest clusters and combine them into one. The number of clusters is reduced by one.
3) Calculate the distance between the new cluster and all others using the clustering criterion.
4) Repeat steps 2 and 3 until you get a single cluster containing all of the points.

The algorithm is very simple and intuitive; its drawback is that its computational complexity is of the order of $O(n^3)$.

4. Centroid clustering

In this method, the cluster is represented by a point of gravity (centroid), which is not necessarily a component of the data set. One of the simplest algorithms of this method is the k-means algorithm. It leads to the simple and easy division of the data set into a given number of clusters [5]. The first step is to set the centroid for each cluster. It is very important to deploy them appropriately because this algorithm finds only the local minimum of the target function, so the clustering depends on the initial centroid spacing. The best choice is to place them as far apart as possible. In this case, an objective function is a squared error function.

The next step is to assign each point belonging to the data set to the closest centroid. After the first grouping, the centroids' positions are recalculated as the gravity points of the clusters formed in the previous step. After creating a new set of centroids, we introduce new relationships between data points and the closest new centroid. We repeat this operation as long as the gravity points change their position.

Formally, this algorithm aims to minimise the objective function, which in this case is a square error function. Of course, the behaviour of the algorithm depends on the metric we assume and on the initial distribution of gravity points, as has already been said. Therefore, it is recommended to repeat it several times with different starting points to avoid stopping at the local minimum.

5. Examples of applications

The following cases are examples in which cluster analysis gave interesting results in water monitoring.

The first example is the application of cluster analysis to investigate the impact of floods on the quality of waters of the Goczałkowice Reservoir [6]. This study analysed the values of 22 water quality indicators measured 42–48 times in 2010; during this year, the reservoir has received three floods. This occurred 17/05/10, 02/06/10 and in the first days of September. For all measurements from 2010, cluster analysis was performed with both methods described previously for physicochemical indicators. Both methods separated a group of three measurements, namely those taken on 19/05/10, 25/05/10 and 07/09/10, the remaining measurements were attributed to the second group. These dates correspond to the aforementioned floods. To investigate whether this division is real, the significance of the differences between the groups for these indicators was checked by the Mann-Whitney U test. In 16 out of 22 cases, the differences were statistically significant. Similar results were obtained for bacteriological water quality indicators.

Another example is the study of the differentiation of heavy metal content in bottom sediments of dam reservoirs [7–9]. For example [8], cluster analysis enabled the separation of data due to similar proportions of granulometric fractions – for the Czorsztyński Reservoir – 5, and for the Goczałkowice Reservoir – 4. It has been found that for such a division, there is also a statistically significant difference in the content of lead in the sediment. This makes it clear that cluster analysis helps the isolation of sediment groups, the parameters of which differ significantly.

It was also possible to isolate similarity groups of photosynthetic dyes extracted from *Chlorella vulgaris* algae grown in waters from individual points of the Goczałkowice Reservoir collected at regular intervals [10, 11]. The similarity cluster identified three groups of measurements [11]. The first of these corresponded to the water quality during the occurrence of flooding and just after the flood. The second group corresponded to the water quality from the pre-flood period and from the measuring points located near the Vistula tributary from the whole period, with the exception of the first group. The third group concerned the measurements from the autumn period, when the vegetation clearly disappears.

Such a division seems to be very intuitive and obvious after its discovery, but no other standard method was able to obtain it.

6. Conclusions

The above examples show that cluster analysis is capable of separating data areas with different environmental quality indicators. This is possible both for data differentiated over time [6] and for spatially differentiated data [8]. This is also possible when differentiation of indicators simultaneously takes place both in time and in space [11].

The above examples clearly show that cluster analysis is a useful tool for data analysis in water monitoring.

I am deeply grateful to Ph.D. D.Sc. Anna Czaplicka and Ph.D. Zbigniew Ślusarczyk *for making their work available to me and giving me valuable explanations. I am also grateful to Prof. Agnieszka Generowicz, whose persistent encouragement and help contributed to the publication of this work.*

References

[1] Han J., Kamber M., Pei J., *Data Mining. Concepts and Techniques*, Elsevier, 2012.

[2] Jain A.K., Murty M.N., Flynn P.J., *Data clustering: a review*, ACM Computing Surveys, Vol. 31, No. 3, September 1999, 264–323.

[3] Mahalanobis P.C., *On the generalized distance in statistics*, Proceedings of the National Institute of Sciences of India, Vol. 2, No. 1, 1936, 49–55.

[4] Johnson S.C., *Hierarchical Clustering Schemes*, Psychometrika, Vol. 32, issue 3, 1067, 241–254.

[5] MacQueen J.B., Some *Methods for Classification and Analysis of Multivariate Observations*, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability, Vol. 1, University of California Press, Berkeley 1967, 281–297.

[6] Ślusarczyk Z., Czaplicka-Kotas A., *Wpływ powodzi w roku 2010 na jakość wód Zbiornika Goczałkowice*, Czasopismo Techniczne, 2-Ś/2012, 261–270.

[7] Czaplicka A., Bazan S., Szarek-Gwiazda E., Ślusarczyk Z., *Spatial distribution of manganese and iron in sediments of the Czorsztyn Reservoir*, Environment Protection Engineering, Vol. 42, No. 4, 2016, 179–188.

[8] Czaplicka A., Szarek-Gwiazda E., Ślusarczyk Z., *Factors influencing the accumulation of Pb in sediments of deep and shallow dam reservoirs*, Oceanological and Hydrobiological Studies, Vol. 46, issue 2, 2017, 174–185.

[9] Czaplicka A., Ślusarczyk Z., Szarek-Gwiazda E., Bazan S., *Rozkład przestrzenny żelaza i manganu w osadach dennych Zbiornika Goczałkowice*, Ochrona Środowiska 3/2017, 47–54.

[10] Czaplicka-Kotas A., *Badania wpływu jakości wody na wytwarzanie barwników fotosyntetycznych w komórkach glonów* Chlorella vulgaris *na potrzeby biomonitoringu wód powierzchniowych*, Ochrona Środowiska 1/2007, 27–33.

[11] Czaplicka-Kotas A., Lodowska J., Wilczok A., Ślusarczyk Z., *Changes of photosynthetic pigments concentration in the synchronous culture of* Chlorella vulgaris *as an indicator of water quality in Goczałkowice Reservoir*, Archives of Environmental Protection, Vol. 35, No.1, 2009, 65–73.