3

Analysis of Water Quality Data for Scientists

József Kovács[1], Péter Tanos[1], János Korponai[2],
Ilona Kovácsné Székely[3], Károly Gondár[4],
Katalin Gondár-Sőregi[4] and István Gábor Hatvani[1]*
[1]*Eötvös Loránd University,*
Department of Physical and Applied Geology, Budapest,
[2]*West Transdanubian Water Authority,*
Department Kis-Balaton, Keszthely,
[3]*Budapest Business School, Institute of Methodology, Budapest,*
[4]*Smaragd GSH Ltd., Budapest,*
Hungary

1. Introduction

In the last few decades the need for stochastic models and the use of time series and data analysis methods in surface and groundwater research has increased greatly. The reason behind this phenomenon is the increase in the size and number of datasets in which it has become necessary to investigate the connection between random variables, and in the case of time series, their characteristics.

The most often used models are deterministic, although they are prepared from one sampling event. It must be stated that the statistics and model results obtained from this sampling event can significantly change if the sampling is to be reproduced because their results are probability variables (Kovács & Székely, 2006). In the case of deterministic models this problem is solved by means of sensitivity analyses, thus the uncertainty in the applied model remains. This may be the reason why the following can be found in the international literature regarding this question: "The future is stochastic modeling" (Kovács & Szanyi, 2005; Wilkinson, 2006).

This chapter is intended to introduce the application of a few exploratory data analysis techniques, primarily through examples. Exploratory data analysis methods are useful and important tools for obtaining an overview of systems which can be described by many different parameters, for determining the latent and explicit connections between the parameters and for sorting and grouping the data obtained based on mathematics.

The greatest value of this chapter lies in its interdisciplinary character; it casts light on environmental problems originating from a wetland ecosystem a river and a groundwater system as well.

*Corresponding Author

2. Materials and methods

2.1 Data bases and data handling

2.1.1 Data base in four dimensions

Before describing in detail the methods applied, the general properties of the environmental datasets and their handling should be discussed. In most aspects of water research the events analyzed are described by different parameter samples obtained from more than one spatial sampling location. The location of a planar sampling site is determined by two spatial coordinates (x; y) forming the two dimensions, while the parameter types (e.g.: runoff, water level, calcium, height etc.) take place in the third dimension. A hydrogeological process is often described by status parameters sampled only once. If, however, samples are analyzed over time, one is dealing with time series and the three dimensions are extended to four, with time as the fourth axis (Fig. 1).

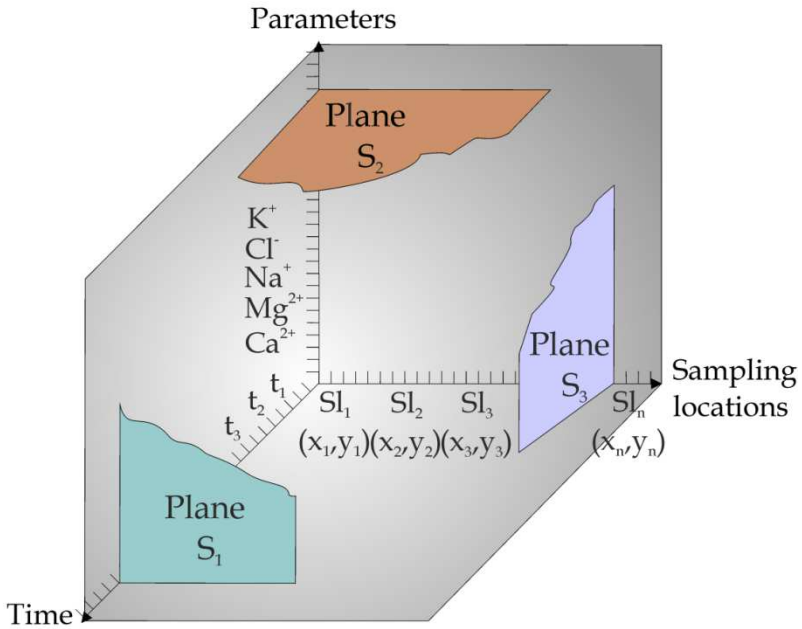


Fig. 1. Data in four dimensions (Sampling locations (x_i; y_i), parameters, time) (based on Kovács et al., 2008)

As an example let us imagine that in a certain area many parameters are sampled from more than one groundwater monitoring well (GMW) at the same time (plane S1). The data obtained is then recorded on worksheets where each row corresponds to a GMW and each column to a parameter. This data matrix is considered to be static. Besides the common univariate statistical methods multi-variate ones can be used, such as cluster-, discriminant-, principal component- and factor analysis along with multi-dimensional scaling.

Cluster analysis (CA) and multi-dimensional scaling can be used on sampling locations when there is a need to reveal similarities. Another aim can be to determine the background

processes that explain most of the original dataset's variance. This can be achieved using principal component analysis (PCA) or factor analysis (FA). During CA the rows of the data matrix and during PCA of FA its columns are the object and the analyses.

In most cases the datasets obtained contain the parameter determining the fourth dimension, time. In this case the data matrix is not static. Staying with the previous example, if more than one GMW is sampled equidistantly in time one is dealing with problems described in plane S2. In this plane the most frequent question is which background processes drive the sampled parameter's (in our case the water levels) temporal fluctuation. Because consecutive temporal samples are not independent of each other only dynamic factor analysis (DFA) can be applied to answer the question raised (Márkus et al., 1999; Ritter et al., 2007). Its application began only in the last years of the 20th century, and with its promising results its role in solving environmental problems is expected to increase (Kovács et al., 2004).

Plane S3 is where the time series of multiple parameters are examined at only one sampling location. Here "classical time series methods" (Shumway & Davis, 2000; Hans, 2005) can be employed, the use of which implies determining the parameters' trend and periodicity etc. In many cases determining these two characteristics is a key question of the study. If the periodicity and trend of a certain process is extracted it can be used for forecasting, but only if one is certain that the driving processes still exist and will exist in the distant future. However further expansion of this topic is not an aim of this chapter.

Returning to the spatial alignment of the sampling locations, by having two spatial coordinates arranged for every parameter sample one will be able to visualize its spatial distribution (on isoline maps) which could be of great use. However if a parameter's spatiotemporal changes are of interest -for example in non-stationary cases- only a few tools are at hand, and the research to solve four dimensional problems in ongoing. Here we refer the reader to Dryden et al. (2005).

2.1.2 Common problems in data handling

Accurate results can only be expected from multi-variate methods if the datasets used contain the desired information that describes the investigated processes precisely; in other words, the amount of data obtained is sufficient. Determining what amount is sufficient information is the duty and responsibility of the given discipline. As an important requirement the number of sampling locations should exceed the number of analyzed parameters (Füstös et al., 1986), in this way statistical stability would be ensured.

One of the most important criteria regarding the data matrix is that there should be no data missing. Unfortunately in many cases (mostly water quality data) data is missing from the datasets. The solution could be data replacement but this has to be done with caution. In practice it often happens that the missing data is replaced with *'0'*. This is a huge error and should be corrected at all times. Another frequent mistake is when a measured parameter's values are below detection limit and the analyst therefore sets its value in the matrix to, for example: the detection limit, or half of the detection limit etc. A dataset like this can give misleading results.

Extreme or outlying values can also lead to inaccurate results. To decide precisely which datum is really outlying or extreme and which one is mistyped is a key question. The certain parameters' variability can be of help in deciding this question.

In water quality data it is often observable that a parameter is the linear combination of one or more others. This of course cannot be used in the course of multi-variate data analysis. Every criterion can only be held if the data matrix is checked for these and other kinds of errors before analysis. This is the most annoying and time-consuming part of the research. However, skipping this step will inevitably lead to incorrect results and conclusions.

2.2 Applied methods

The most effective order in which the presented methods should be used is the following. After checking the data matrix for the errors described above it is necessary to examine the data using uni-variate methods like descriptive statistics, distribution analysis, hypothesis testing of some sort and finally determining the stochastic connections with correlation analysis (Helsel & Hirsch, 2002). Out of these methods, correlation analysis is the one that will be discussed.

The next step is the application of multi-variate methods. The first method suggested is cluster analysis. Its results are groups of similar sampling locations. As a verification tool discriminant analysis and as a tool for determining which parameters influenced the formation of the cluster groups the most, Wilks' lambda distribution is suggested, along with the overview of the groups' statistics (Box-and-whiskers plot).

As a last step it is proposed to determine the driving background processes using PCA and if possible visualizing the results on maps for better interpretation.

2.2.1 Correlation (stochastic connections)

A frequent question is what kind of relation can be revealed between two parameters. The connection's strength should be described numerically. The most common is the Pearson correlation coefficient which measures the strength and direction of a linear relationship between two variables. It is calculated as follows:

$$r_{(x,y)} = \frac{\sigma_{(x,y)}}{\sigma_x \sigma_y}$$

Where σ_x, σ_y are the standard deviation of variables *X* and *Y*, while the numerator is the covariance.

This means that the correlation coefficient includes the covariance's every good quality and with the division by the standard deviation it will be independent from measurement units, and the upper lower boundary problem will be solved as well. The properties of the correlation coefficients are the following:

- Its value is independent from the measurement unit of *X* and *Y*
- if $\sigma_x = 0$ or $\sigma_y = 0$, then $r_{(x,y)} = 0$
- $-1 \le r_{(x,y)} \le 1$;
- If the relation between *X* and *Y* is positive, then $0 < r_{(x,y)}$, if this relation is negative, then $r_{(x,y)} < 0$
- if $r_{(x,y)} = \pm 1$, then there is a linear functional relation between the two variables
- any variable's correlation with itself is 1

- if the correlation coefficient is zero the two variables are uncorrelated, however this does not mean that they are independent
- if two variables are independent then $r_{(x,y)}=0$

In the studies presented correlation coefficients (in absolute value) higher than 0.71 were considered to represent strong linear relationship (Füst, 1997).

2.2.2 Cluster

Clustering is a kind of coding, in which a certain sampling location -originally described with many parameters (runoff, chemical oxygen demand etc.) is now described with only one value, its group code (cluster number). It is important to note that during clustering not the number of parameters but the number of sampling locations is decreased by placing the similar ones into groups. It is an important criterion that every sampling location has to belong to a group, but only one group. It is obvious that there are many possible group conformations. The main aim is to settle the similar sampling locations into the same group, however this similarity has to be measured by assigning a distance (metrics) to each sampling location which is placed in an *N*† dimensional space. If the distance between two sampling locations is small, then they are highly similar to each other. If the distance is zero they are perfectly similar. From this it should be clear that choosing the right distance is a key question. It needs skill and practice. This means that the verification of cluster results is compulsory.

"Cluster analysis (CA) classifies a set of observations into two or more mutually exclusive "unknown" groups based on combinations of interval variables. The purpose of cluster analysis is to discover a system of organizing observations, usually people, into groups. where members of the groups share properties in common" (Stockburger, 2001).

There are basically two types of clustering, the K-Means CA and the Hierarchical CA. In the former, one has to predetermine how many groups are required and is frequently suggested to be used with large datasets, in latter one only has to determine the groups after obtaining the dendrogram, the graphical output of the CA. In this study divisive Hierarchical CA was applied, where one group is divided to many more and so on. Its opposite is the agglomerative Hierarchical CA when the number of groups is reduced during the analysis.

2.2.3 Discriminant analysis and Wilks' lambda distribution

To verify the accuracy of the results, discriminant analysis can be used. It shows to what extent the planes separating the groups can be distinguished by building a predictive model for group membership. The model is composed of a discriminant function (for more than two groups a set of discriminant functions) based on linear combinations of the predictor variables that provide the best discrimination between the groups. The functions are generated from a sample of cases for which the group membership is known; the functions can then be applied to new cases that have measurements for the predictor variables but their group membership is as yet unknown (Afifi et al., 2004). The result of the discriminant

†In this case the number of dimensions (*N*) is equal to the number of measured parameters.

analysis is often visualized on the surface stretched between the first two discriminating planes (function 1 & function 2, e.g. Fig. 13) (Ketskeméty & Izsó, 2005).

After the verification of the cluster groups the role of each parameter should be analyzed in determining the formation of the cluster groups. Using Wilks' λ distribution a Wilks' λ quotient is assigned to every parameter, where the quotient is:

$$\lambda = \frac{\sum_i \sum_j (x_{ij} - \bar{x}_i)^2}{\sum_i \sum_j (x_{ij} - \bar{x})^2}$$

Where x_{ij} is the j^{th} element of the i^{th} group, $\bar{x}_i$ the i^{th} group's mean, and $\bar{x}$ the total mean.

The value of λ is the ratio of the within-group sum of squares to the total sum of squares. It is a number between 0 and 1. If λ=1, then the mean of the discriminant scores is the same in all groups and there is no inter-group variability, so, in our case the parameter did not affect the formation of the cluster groups (Afifi et al., op. cit.). On the contrary, if λ=0, then that particular parameter affected the formation of the cluster groups the most. The smaller the quotient is, the more it determines the formation of the cluster groups.

2.2.4 Box-and-whiskers-plots

Box-and-whiskers plots are great tools for visualizing more than one statistic of a parameter on one graph, making the interpretation clearer. The boxes show the interquartile range and the black line in the box is the median. Two upright lines represent the data within the 1.5 interquartile range. The data between 1.5 and 3 times the interquartile range are indicated with a circle (outliers), and the ones with higher values than 3 times the interquartile range are considered to be extreme values indicated with an asterisk (Norusis, 1993). For example see Fig. 8.

For better interpretation it is necessary in every case possible to visualize the results on maps of some sort.

2.2.5 Principal component analysis compared with factor analysis

The principal component analysis (PCA) and factor analysis (FA) methods are used to analyze multidimensional data. The goal is to determine the background processes, while describing the observed parameters with fewer hypothetical variables without any significant information-loss in the original data.

During PCA the measured chemical and physical parameters are correlated, whereas the hypothetical variables (called principal components) are uncorrelated and are obtained as a linear combination of the original parameters. The PCA decomposes the total variance of the original variables to principal components which explain the original variance in a monotonically decreasing way. The correlation coefficients between the original parameters and the principal components are the factor loadings. They explain the weights of the original parameters in the principal components, however they do not give an exact answer whether a weight has to be considered as significant or not, and how many principal components are important.

During FA the hypothetical variables are called factors. They are classified into three categories: (1) common factors (influencing multiple parameters), (2) specific factors (influencing one parameter) and (3) error factors (arising for example from inaccuracy during the measurement). In this method only the common factors and their factor loadings are determined, because identifying specific-, and error factors along with their factor loadings usually causes mathematical difficulties. As a result, the common factors only explain a part of the total variance of the original parameters. While the PCA uses a correlation matrix, the FA uses an adjusted correlation matrix. In this matrix the elements on the diagonal (commonalities) can be estimated in different ways, as a result different solutions (of the FA) may be acceptable. The basic model of FA suits the conditions of datasets in earth sciences better (Geiger, 2007), this may be the reason for its successful application (Voudouris et al., 1997; 2000). In this chapter only PCA was applied.

In regard to the software used, we suggest using STATISTICA for its good visual output and user friendly interface, SPSS because its more commonly used and for its "syntax system" and last but not least R because it is an open source freeware and the most up-to-date software available.

3. Case studies giving an example on the data analysis methods

In sections 3.1-3.3 three cases studies are presented to give an example of the most efficient application of the basic data analysis methods (section 2) using the datasets of a mitigation wetland, a river and a groundwater area (Fig. 2).

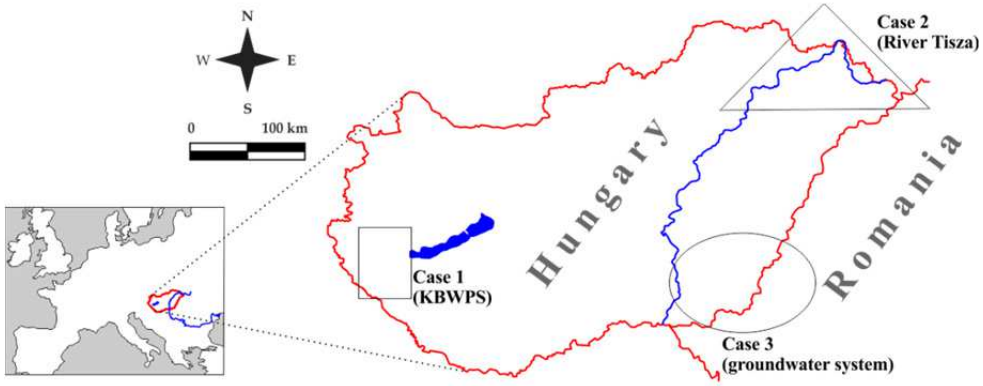


Fig. 2. The location of the three case studies numbered in order of appearance

3.1 The Kis-Balaton Water Protection System (KBWPS)

The KBWPS is a mitigation wetland located at the mouth of the River Zala at Lake Balaton, the largest shallow freshwater lake in Central Europe (Padisák & Reynolds, 2003). It was settled on the remains of the original Kis-Balaton Wetland (KBW), which due to artificial water level modifications decreased in area and functionality in the late 19th century. The original KBW used to naturally filter the waters of the River Zala which supplies 45% of Lake Balaton's water and 35-40% of its nutrient input (Lotz, 1988; Kovács et al., 2010).

During the time period between the function loss of the KBW and the construction of the KBWPS the Zala's waters were less filtered. This resulted in the water quality deterioration of Lake Balaton. As a solution to the problem of nutrient retention at the mouth of the River Zala and stop the degradation of Balaton's water quality the KBWPS was constructed (Tátrai et al., 2000; Kovács et al., 2010).

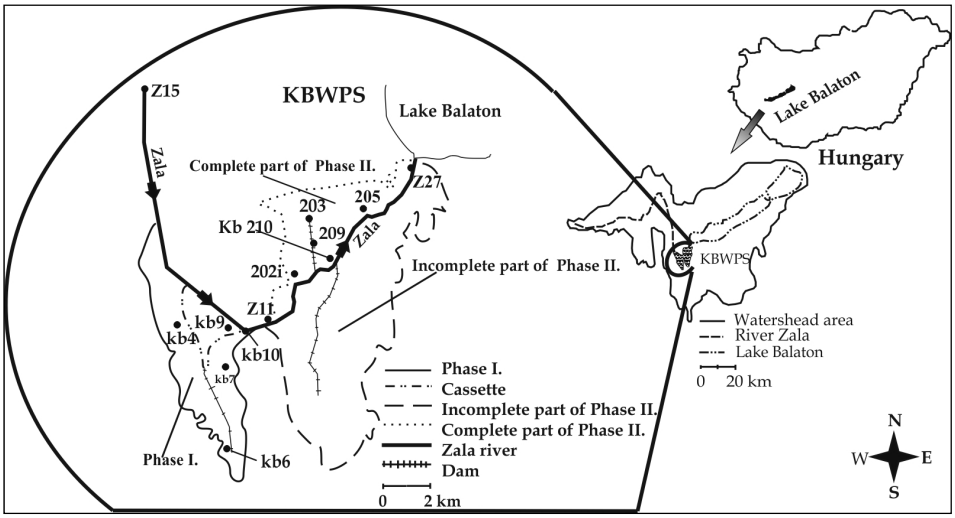


Fig. 3. The KBWPS and its sampling locations (based on Hatvani et al., 2011)

3.1.1 The structure of the KBWPS

The construction of the KBWPS was planned to take place in two phases (Fig. 3) as an extended wetland. Phase I was finished in 1985 after a five step flooding (Korponai et al., 2010). It resulted in a eutrophic pond, which gives ideal conditions for algae to reproduce. Phase II remains incomplete. Only a 16 km^2 area has been functioning since 1992. Its habitat could be described as a "classic" wetland, with 95% macrophyte coverage, primarily reeds (Nguyen et al., 2005; Tátrai et al., op. cit.). It is a highly protected nature conservation area and under the de-jure protection of the Ramsar Convention (1971).

3.1.2 Sampling locations and examined parameters on the KBWPS

Since the installation of Phase I weekly sampling has been carried out by the laboratory of the West Transdanubian Water Authority's Kis-Balaton Department on thirteen sampling locations. In the following the most important locations are described in detail heading downstream.

- "Z15", the inlet of Phase I, which typifies the water of the River Zala
- "Kb4; Kb6; Kb7"
- "Kb9", the sampling location that typifies the waters of the Cassette, which is the site of biological experiments, and is the most eutrophic place in the system
- "Kb10"

- "Z11", the interface between Phase I and Phase II, which typifies the water generated by Phase I, the eutrophic pond
- "202i; 203; 209; 210"
- "205" is the location where the Combined Belt Canal and drainage pipes from fishing lakes coming from Somogy County enter the KBWPS
- "Z27", the outlet of KBWPS, which typifies the water output of Phase II, the "classic" wetland area of the KBWPS

In this study the following parameters were examined for the time interval (1984-2008) with the methods described in section 2.2.

Runoff (m^3 s^{-1}); Chemical oxygen demand-potassium-dichromate original (COD^C); biological oxygen demand (BOD-5); dissolved O_2 (mg l^{-1}); chlorophyll-a (mg m^{-3}); Cl^-; SO_4^{2-}; HCO_3^-; CO_3^{2-}; Fe^{2+}; Mn^{2+}; Mg^{2+}; Ca^{2+}; Na^+; K^+; NH_4-N; NO_2-N, NO_3-N; total nitrogen (TN); total phosphorus (TP); dissolved phosphorus; soluble reactive phosphorous (SRP); suspended solids (mg l^{-1}); pH; conductivity (μS cm^{-1}). This resulted in around ~34,000 data per sampling location.

The sampling was conducted and the samples were prepared according to the current Hungarian standards (MSZ 12749 & MSZ 12750).

In case of the KBWPS the dataset was prepared by the same authority during the whole investigated time period. However in most cases the scientist is not this fortunate, he has to deal with the problems that originate from different sampling methods, and even standards (in case of an international research). To explore these problems and maybe even solve them a preanalysis and probably cross-verification of the data must be carried out before its analysis. Even in our case –the KBWPS- each-and-every data was checked whether it is an extreme-, an outlier- or a mistyped value or not. An x-y scatterplot was created for each parameter for the whole time interval and analyzed manually. It would not have been sufficient to set a high and low margin or even apply the three-sigma rule (Pukelsheim, 1994) because there were certain events that resulted in extremely high values (e.g. floods) that can easily be recognized from their surroundings in the graphs, and these would have been falsely discarded.

3.1.3 Results of the analyses conducted on the KBWPS

3.1.3.1 Cluster-, discriminant and Wilks' lambda analyses

Using cluster analysis on the 13 sampling locations and the annual averages of the parameters specified in table for the years 1984-2008 we were able to point out the alignment of the similar sampling locations for each year. The sampling locations basically grouped up according to the two constructional phases, but there were interesting exceptions (Fig. 4).

- Between 1997 and 1998 sampling location 202i disconnects from Phase II and connects to the cluster group, covering Phase I. This is the only location that changes its orientation and keeps it for the rest of the time analyzed. This occurred because of the constantly high water level of the area. Because of that the reeds died out and the surroundings of sampling location 202i became similar to Phase I, an open and eutrophic water space became dominant and the area of the "classic" wetland decreased along with the system's efficiency (Hatvani et al., 2011).

- Sampling locations Kb9 and 205 form separate groups, the former in 1997 and 1999, and the latter in 1996 and 1997. Previous because it is located in the Cassette, this constituted a separate waterspace with almost no water flow and high water retention time, so it is highly eutrophic, while with the latter (Kb205), because of the Combined Belt Canal and drainage pipes of fishing lakes from Somogy County joining the system here.

If one uses cluster analysis the results always have to be verified. Discriminant analysis results pointed out that 100.0% of the original grouped cases proved to be correctly classified.

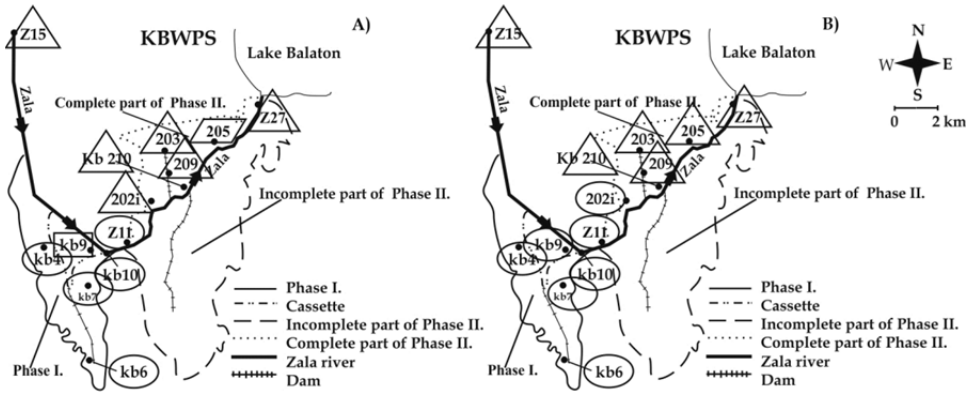


Fig. 4. Cluster groups of the KBWPS in 1997 (A) and control years 2007 & 2008 (B) (based on Hatvani et al., 2011)

Using cluster and discriminant analyses it was possible to point out key problems in the water level management of the system (Hatvani et al., 2011). For better interpretation the visualization of the spatial cluster results on a map of some sort in every possible case is suggested.

As discussed previously, Wilks' lambda distribution determines the parameters that effect the conformation of the previously discussed cluster groups the most. A Wilks' λ quotient was assigned to every parameter for every year and then clustered. Grouping the eleven quotients, the parameters were separated to three different groups according to how much they determine the original spatial cluster groups (Fig. 5).

Group 1 contains: chlorophyll-a, dissolved P, suspended solids and pH, CO_3^{2-}, HCO_3^-, dissolved O_2. These parameters had the smallest Wilks' λ quotients (average: 0.32), so they affected the conformation of the cluster groups the most and were responsible for the separation of the cluster group covering the eutrophic pond.

Group 2 contains the parameters which also stand in close relation to the eutrophication processes: Ca^{2+}, COD^C, total N, and total P, their average quotient was 0.55.

In fine the parameters in Group 3 (NH_4-N, NO_2-N , NO_3-N , Mg^{2+}, SO_4^{2-}, Cl^-, Fe^{2+}, Mn^{2+}, Na^+, K^+) generated an average quotient of 0.69 during the Wilks' λ distribution, so they affect the orientation of the spatial cluster groups the least and play a great role in the separation of the cluster group covering the Wetland (Phase II).

With these three methods used together a global picture was obtained regarding the similarities of the KBWPS' sampling locations and the parameters that drive these

similarities. After gaining knowledge concerning the connection between the sampling locations and the parameters behind them, the next step would be to familiarize oneself with the processes evolving in the different areas of the water protection system.

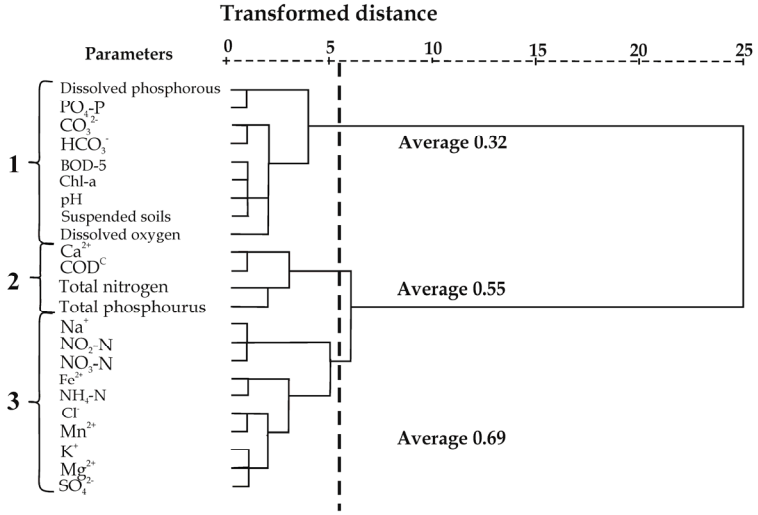


Fig. 5. Dendrogram of Wilks' lambda quotients

3.1.3.2 Stochastic connections

From the correlation matrix it became clear that at Z15 only the phosphorus forms (TP, SRP, dissolved P) in case of Z11 and Z27 besides the phosphorus forms the Na^+ and Cl^- ions have correlation coefficients (in absolute value) higher than 0.71. Meanwhile, at all three sampling locations the number of weak linear connections (values of correlation coefficients is between ±0.2) is 137-120-136 respectively.

In summary it can be said that in the eutrophic area of the KBWPS (Phase I) there are fewer weak linear connections than in the riverine or wetland areas. However, regarding the whole system, the linear connection between the 22 parameters cannot be considered generally significant. This is reflected in the following PCA results.

3.1.3.3 Principal component analysis

The PCA results presented concern the three cardinal, and one peculiar sampling locations of the KBWPS (Z15, Z11, Z27 and Kb9) and their surroundings. Their summer datasets were analyzed for the time interval 1984-2008.

The aim was to find the parameters that determine the processes evolving –mainly- in Phase I of the Kis-Balaton Water Protection System (KBWPS).

The biggest problem which had to be faced in the course of the PCA was that the datasets were time series, where data follow each other and are therefore not independent‡ because

‡The spatial or temporal distance when consecutive data turn independent can be measured for e.g. with variogram analysis (Kovács et al., 2011),

they are too close to each other in time, and the PCA cannot handle this kind of input[§]. To solve this problem only three summer months (May-July-September) were analyzed with two-month gaps in between. This is the growing season when primal production is most intensive (Wetzel, 2001).

Regarding the results, only the first factor can be considered as significant. It explains 25 to 30% of the data's variance, while the second PC only explained 15-19%, and so was discarded. The low explanatory value may originate from the parameters' weak linear connection, reflected in the correlation matrixes discussed in section 3.1.3.2.

If the constitution of the first component is analyzed in more detail it can be said that at:

- **Z15** (inlet, the River Zala) where the parameters TP, Dissolved P and SRP are present in the first PC. It corresponds what is already known about the River Zala, that there is technically no planktonic life in its water, and bentic eutrophication is not dominant in its sediment. These parameters originate from diffuse loads, thus their concentration only depends on the runoff of the River Zala. The highest concentrations were observable at peak flooding times.
- **Kb9** (Cassette) the parameters Chl-a and TP are present in the first PC. These are the main parameters that the OECD (Vollenweider & Kerekes, 1982) uses to classify the trophic conditions. Their presence is no surprise, because at Kb9 the water is still, with almost no water flow, and long hydraulic residence time. Bottom-up[**] processes dominate at this sampling location and its processes can be described with the Vollenweider model (Vollenweider, 1976).
- **Z11** (interface, representing Phase I) the parameters TP, Chl-a, TN Ca^{2+}, Suspended solids and Cl^- are present in the first PC, indicating Phase I's eutrophic and algae dominant waterbody. Again, the Vollenweider model can be used to describe this environment. The calcium ion indicates biogenic carbonate precipitation which is a dominant process in certain locations of Phase I.
- **Z27** (outlet, describing Phase II) the parameters K^+, Na^+, Cl^- and NO_3^- are present in the first PC, where decomposition processes are dominant. The reason for NO_3^- being one of the dominant parameters is that nitrification is the main process in this section of the system indicating aerobic conditions.

In summary it can be stated that PCA is a universal tool for determining the dominant processes of a certain sampling location's area or a whole system. The facts that were perceptible to the naked eye, such as the eutrophic conditions of the Cassette are now written down in numbers and can be subject of further scientific studies.

3.1.4 Conclusions regarding the results obtained from the KBWPS

The development and the functioning of the KBWPS is a good testing ground for new habitat remediation techniques and a great example of the presentation of the application and the use of the methods described above. The result of each method gave extra information to scientists and confirmed their previous suspicions. For example the constant water level deteriorates the wetland vegetation (Pomogyi et al., 1996), therefore decreases

[§] Instead of PCA dynamic factor analysis was developed to handle temporal dependence

[**]Bottom-up control: ecological scenario in which the abundance or biomass of organisms is mainly determined by a lack of resources and mortality owing to starvation (Pernthaler, 2005)

the efficiency of the system and that this process reached a peak in 1997 and 1998. It was always known that the water need of different vegetation types is different; the key result was that the irreversible change at 202i happened in those particular years. This was just one example highlighted regarding the mitigation wetland.

In the next section (3.2) the same methods were applied to the largest tributary of Europe's second longest river the Danube.

3.2 River Tisza

The River Tisza collects the waters of the Carpathian Basin's eastern region. According to Lászlóffy Woldemár (1982), its watershed area is 157,186 km^2. Less than one third of this is located in Hungary (Fig. 6).

Fig. 6. Watershed of the River Tisza (Based on Istvánovics et al., 2010)

From its spring in the Maramureşului Mountains to its confluence with the Danube, it stretches for 966 km across the Ukraine, Romania, Slovakia, Hungary and Serbia (Sakan et al., 2007). The Hungarian section of the Tisza from border to border is 594.5 km long. Its average runoff is 25.4 billion m^3 per year (Pécsi, 1969).

Despite the fact that in the last one and a half centuries numerous anthropogenic activities have influenced this area, in comparison to Europe's other large rivers it is still considered to have one of the most natural river valleys in Europe (Zsuga & Szabó, 2005). It is for this reason that it is in our common interest to protect it. If we only take Hungary into account, then approximately 400 settlements and 1,500,000 inhabitants' lives depend on its runoff and water quality.

3.2.1 Sampling locations and examined parameters on the River Tisza

Many surface waters are monitored as part of the National Sampling Network. In case of the Tisza, data from the first five Hungarian sampling locations were analyzed (258.7 river km) (Fig. 7). The exact specifications of these monitoring locations can be found in Hungarian Standard No. MSZ12749:1993.

The River Tisza reaches the Hungarian border at Tiszabecs. The next sampling location (SL) is at Záhony. There are two tributaries of considerable size between these two locations, the Szamos and the Kraszna. The next SL can be found at Balsa, just before the mouth of the Tokaj and Bodrog rivers. The Tiszalök SL is located just after the Eastern Trunk Sewer and the Tiszalök Water Barrage System. The last SL analyzed is at Polgár downstream from the River Sajó's mouth.

Fig. 7. River Tisza's first five sampling locations in Hungary

During the research, a 31-year-long dataset (1974-2004) was analyzed, consisting of 300,000 data. From 1970 only one sample a week was taken at Tiszabecs and Polgár, according to COMECON††'s specifications (T. Nagy et al., 2004). At Balsa only one sample was taken per month. At Záhony 26 samples were taken every year. In 1994 Hungarian Standard No. MSZ 12749:1993 came into force. As a result, since 1993 26 samples have been taken every year uniformly.

The parameters used were as follows: Runoff ($m^3\ s^{-1}$), pH, Conductivity ($\mu S\ cm^{-1}$), M-alkalinity ($mval\ l^{-1}$), Oxygen saturation (%), Dissolved O_2, BOD-5, COD^C, Ca^{2+}, Mg^{2+}, Na^+, K^+, Total hardness, Carbonate hardness, Cl^-, SO_4^{2-}, HCO_3^-, NH_4-nitrogen, NO_2-nitrogen, NO_3-nitrogen, Mineral nitrogen‡‡ ($mg\ l^{-1}$), PO_4-phosphorus, Chlorophyll-a (Chl-a) ($\mu g\ l^{-1}$).

3.2.2 Results of the analyses conducted on the River Tisza

During our work many temporal approaches can be employed, meaning the whole year can be examined or the dataset can be separated to seasons. In the former case the whole year's processes, while in the latter obviously the seasonal changes can be followed. To shed light on the seasonal changes the winter and summer data were analyzed separately. Summer was considered to last from June to October, while winter from November to March.

3.2.2.1 Cluster- and discriminant analyses and Wilks' lambda statistics

Regarding the River Tisza's data series, cluster analysis was conducted on averages formed from the parameters. This approach gives a longer perspective on the data. One average was formed at each sampling location from each parameter's total dataset.

The cluster analysis resulted in three groups. As a verification tool, discriminant analysis was applied.

††Council for Mutual Economic Assistance founded by the Soviet Union 1949.
‡‡Mineral-N is the summary of the NH_4-nitrogen, NO_2-nitrogen, NO_3-nitrogen.

The main question was whether these groups (formed from the averages) present at all and whether they are discoverable if the data is analysed in more detail (not in averages). The answer is yes.

By using discriminant analysis on the discrete data, the cluster groups formed from the averages were realized for example 94.8% in summer and 89% in winter. This means that the clustering from the averages is correct and representative.

To stay with our examples, Table 1 shows the Wilks' lambda coefficient in summer and winter for each parameter.

	Wilks' lambda coefficients	
Parameter	**Summer**	**Winter**
SO_4^{2-}	0.41	0.53
Na^+	0.74	0.87
Mg^{2+}	0.75	0.78
pH-lab	0.96	0.99

Table 1. Parameters' Wilks' lambda coefficients in summer and winter.

It is clear from Table 1 that the sulphate ion is the most determining in both seasons. Parallel to the Wilks' lambda distribution, it is suggested that the parameters' variability be analyzed, because it gives a much wider picture of certain parameters. In Fig. 8, three parameters are presented on box-and-whiskers plots. One with a small (sulphate, Fig. 8/A) one with a medium (calcium, Fig. 8/B) and one with a high (pH, Fig. 8/C) Wilks' lambda coefficient.

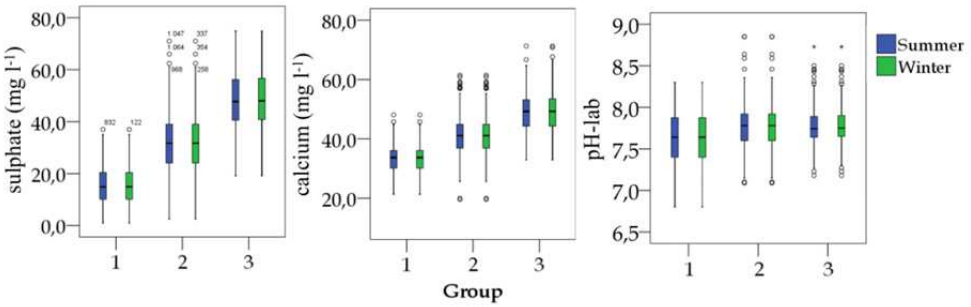


Fig. 8. Sulphate ion (A), calcium ion (B), pH (C) parameters' boxplots in increasing order of their Wilks' lambda quotients.

It is clear that the sulphate is the most variable, calcium is less so, and pH (which was the most influential regarding the Wilks' lambda distribution) is the least variable parameter.

3.2.2.2 Stochastic connections

The stochastic connections were analyzed using correlation analysis. The connection between the parameters was analyzed through different approaches. First the whole year, then the different seasons (winter, summer) were taken into account.

According to each approach (whole year, winter/summer), the number of strong correlations increases downstream. At Tiszabecs, the number of strong correlations is only seven (Table 2), at Tiszalök this number reaches 36. This can be explained by the flow conditions. In the area of the water barrage system, the water-flow slows down, and (according to the spiral model[§§]) so does physical transport; suspended solids are deposited, the water becomes more transparent and light limitation decreases (Padisák, 2005). This gives an opportunity for organisms to compose the nutrients into their systems faster and more efficiently. Usually the River Tisza is autotrophic during the summer months, but tributary input may considerably exceed net autochtonous production (Istvánovics et al., 2010).

	Number of linear connections [pcs.]		
	Annual	**Winter**	**Summer**
Tiszabecs	7	5	7
Záhony	14	17	9
Balsa	17	18	10
Tiszalök(wbs)	37	37	14
Polgár	36	41	18

Table 2. The number of strong linear connections ($|r| \geq 0.7$) at each sampling location by temporal distribution

During the summer months there were fewer strong linear connections than in the winter months. If the results from the whole year are compared to the summer and winter ones, it can be stated that the correlation matrix obtained from the winter data more closely resembles the annual correlation matrix than does the summer one.

As can be seen in Table 2, in the Tiszalök area the number of correlating parameters suddenly rises in both winter and summer.

Summarizing the correlation analysis, it can be stated that the number of correlations increases downstream. The annual, winter and summer results are different in the case of the number of correlations and in the case of the parameters which correlate as well. During the summer there are fewer linear connections, but these few are between the parameters in relation to the organic processes.

As previously stated, if the results for the whole year are compared to the summer and winter ones, it can be seen that the correlation matrix obtained from the winter data more closely resembles the annual correlation matrix than does the summer one. So, if just the whole year had been analyzed, vital differences between the summer and winter would have been lost.

After analyzing the connections between the SL based on the sampled parameters the next step is to take a closer look at the processes evolving in the river.

[§§]The riverine spiral model describes recycling of nutrients together with physical transport downstream (Padisák, 2005).

3.2.2.3 Principal component analysis

To answer the question which background processes determine the water quality and processes of the River Tisza, PCA was applied to the summer, winter and whole year data.

Before the PCA was conducted, the number of parameters had to be decreased, either because the parameter was not sampled during certain time periods or because it was unsystematically sampled over the whole investigated time period. Or, in other cases, the parameter itself contained information concerning other parameters (e.g. specific conductance). There are other cases when the dataset has to be decreased, more examples can be seen in section 3.3.2.3.

In terms of their importance, only the parameters with a factor score (in absolute value) higher than 0.7 were taken into account in the first and second principal components (PC). The summarized results of the PCA can be found in Table 3.

Season / Sampling location	Summer		Winter		Annual	
	First PC	Second PC	First PC	Second PC	First PC	Second PC
Tiszabecs	N-forms	Major ions	N-forms	*None*	Major ions	N-forms
Záhony	N-forms	Major ions	Major ions	N-forms	Major ions	N-forms
Balsa	N-forms	Major ions	Major ions	N-forms	Major ions	N-forms
Tiszalök	Major ions	N-forms	Major ions	N-forms	Major ions	N-forms
Polgár	Major ions	N-forms	Major ions	N-forms	Major ions	N-forms

Table 3. Summarized results of the PCA, None: There were no factor scores ≥ 0.7

Regarding the results, it can be said that the first two components explain approximately 50% of the data's total variance, independent of their spatial and temporal distribution.

Regarding the summer results, at Tiszabecs, Záhony and Balsa mostly the N-forms can be found in the first PC. In the second PC, the ions responsible for halobity (Mg^{2+}, Na^{+}, K^{+}, Cl^{-}) take place. Between Balsa and Tiszalök the background processes show a peculiar change: the scale tilts from the organic components towards the inorganic ones. At Tiszalök in the summer (according to the first PC), the major ions (e.g. Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, Cl^{-}) play a determining role; the Polgár SL shows the same pattern. The fact that at Tiszabecs, Záhony and Balsa the N-forms are the most determining in the first PC leads us to the assumption that biological processes such as saprobity and trophic conditions are responsible for the background processes. Since there was no direct relationship observed between nutrient levels and phytoplankton biomass (Istvánovics et al., 2010) other factors are responsible for changes in the N-forms. In contrast, the results from Tiszalök and Polgár show a change in the determining processes. After Tiszalök, the inorganic processes (e.g. aggregation, dissolution) take the place of the N-forms in the first PC.

From the perspective of the winter results, the first PC's explanatory power varies between 20% and 40%. At each SL except for Tiszabecs the ions determine the background processes. In the second factor, the N-forms are dominant.

Regarding the whole year's PCA results it can be stated that the annual conditions resemble the winter ones to a high degree (just as in the case of the correlation results). In the first PC

(explanatory power: 21-39%) the ions take on the determining role, while in the second one the N-forms are dominant.

Regarding the temporal distribution, it is clear that during the winter inorganic processes are dominant in determining the Tisza's water quality and these results represent the annual conditions much more than the summer results do.

As in the case of the correlation analyses, and of the PC as well, results which are not temporally divided are not satisfactory, but this can only be confirmed if the summer and winter data are analyzed separately, as it was done in this case.

3.2.3 Conclusions regarding the result obtained from the River Tisza

It is clear that the methodology applied in the case of the KBWPS can successfully be applied to a river as well. The advantage of the methodology used is that it delimits sub-systems not based on geography but mathematics. It was known from the literature, that three large hydro-geomorphic sections can be distinguished along the river (Istvánovics et al., 2010). The meandering upstream and downstream sections are separated by the impact by the two reservoirs. The PCA for example clearly separated the Hungarian section of the River Tisza to two sub areas (Tiszalök SL (water barrage system)in the middle), where Tiszabecs and Balsa SLs belong to the upstream and Tiszalök and Polgár SLs to the reservoir section. This section however causes discontinuity in the environmental gradient along the river. The separation is the strongest in summer when autotrophic processes become driving forces in the reservoir. After presenting the two surface waters the final case study, the analysis of SE Hungary's groundwater will be discussed.

3.3 Groundwater

3.3.1 Sampling locations, and examined parameters

Drinking water has high arsenic content in south-eastern Hungary. Within the framework of the pilot project (Sustainable management and treatment of arsenic bearing groundwater in Southern Hungary (SUMANAS) —LIFE05 ENV/H/000418), the formation of the arsenic (geological in origin) and the development of its decontamination were examined. The geological part of the study was prepared by Smaragd GSH Ltd. Appropriate water chemistry analysis results made the utilization of multi-variate data analysis methods possible.

Subsurface water resources of the Pleistocene aquifer in the south-eastern part of the Great Hungarian Plain contain more or less arsenic. SE Hungary is a subsiding back arc basin that was loaded with river sediments (aleurit, sand and clay) in the Pleistocene. One of the two rivers, the Tisza, has a metamorphic, volcanic and re-accumulated sedimentary catchment area, while the Maros River mainly derives its sediments from volcanic territory (Nádor et al., 2007). Arsenic has been transported by fluvial fine-grained sediments (fraction: <2 μm), adsorbed on the surfactant amorphous iron-oxyhydroxides (Varsányi & Ó. Kovács, 2006), clay minerals and organic materials.

The primary transportation, accumulation and desorption of the arsenic from the sediments into the water is determined by absorbents (amorphous iron-hydroxides and the surface of the organic matter, clay minerals) (Lin & Puls, 2000; Varsányi & Ó. Kovács, 2006),

physicochemical conditions (Redox conditions, pH), and changes in the groundwater's parameters (recharge, flow regime). These agents display great variability over time and space, causing divergent arsenic concentrations in the groundwater.

Groundwater sampling was carried out at 202 groundwater monitoring wells (like those mentioned in section 2.1.1) All the wells were located in SE Hungary and the bordering area of Romania). Most of these were supply wells plus a few monitoring wells situated at an average elevation of 93 m above Baltic sea level. Groundwater temperature varies according different depth intervals. The average water temperature is 19.6 ^{o}C, the lowest is 12.1 ^{o}C, while the highest is 81 ^{o}C.

In the course of water chemistry analysis, the form of the separate As formulas is the most important step, because As (III) is 60 times more toxic to the human body than As (V). For this reason determining the quantity of the different arsenic formulas present in the groundwater is a crucial point in this study. The method for measuring As (V) and As (III) separately was developed during the project by Bálint Analitika Ltd. In: *Körös Valley District Environment and Water Directorate, 2008*.

The parameters analysed can be found in Table 4, excluding Na^+ (mg l^{-1}) and conductivity ($\mu S\ cm^{-1}$). One of the biggest problems concerning the database was when the concentration of the analysed parameter was beneath the detection limit. This occurred in many cases regarding a few parameters (cadmium, mercury, lead etc.), so these were simply left out of the calculations.

In the case of a few other parameters (ammonium, nitrite, nitrate, sulphate), concentrations both above and below detection limit were observable. On occasions when a parameter is beneath the detection limit, a common practice is that the values are supplemented with for example half of the detection limit (other solutions are mentioned in section 2.1.2). If it is done in this way so the examination of stochastic relationship (PCA) may generate huge errors. In the case of cluster analysis, however, supplementing data (beneath detection limit) with values close to zero may only lead to minor errors, which can be accepted. In summary, it must be stated that keeping in mind which method can handle the values under the detection limit and which cannot is a key question. In some cases using these values will lead to huge errors (e.g. PCA) in other cases the results can still be considered to be correct (CA).

3.3.2 Results of the analyses conducted on the study area

3.3.2.1 Stochastic connections

The correlation matrix shows whether there is a linear relationship between the measured parameters or not. The example is as follows.

The Correlation matrix shed light on a strong linear connection between conductivity (measured on site) and Na^+ and HCO_3^-. If conductivity and Na^+ content are plotted on a diagram (Fig. 9), besides the linear relationship one may easily recognize the different character of the connection in cases of different concentration ranges. Thus, the application of diverse regression functions should be practical if the aim is to determine the relationship to this resolution. The above-mentioned case is depicted in Fig. 9/A. This graph is split into three parts (Fig. 9/ B, C & D). The groups of sampling points presented in these figures are the results of the subsequent cluster analyses. They represent different geographical regions.

The figure series draws attention to the fact that great differences exist between the presented groups that can be interpreted as the result of different hydrogeological conditions (Fig. 9/ B, C, D). Regarding the correlation relationships, space does not allow us to enter into further details in the present paper.

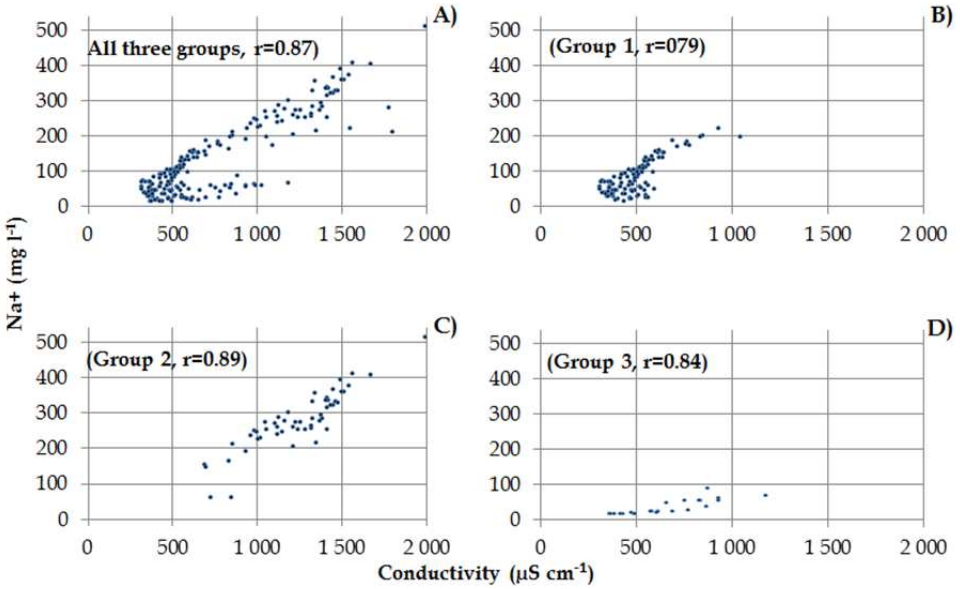


Fig. 9. x-y scatterplot showing the relationship between Na^+ and conductivity for all the groups (A), group 1 (B), group 2 (C), and group 3 (D)

3.3.2.2 Cluster-, discriminant analyses and Wilks' lambda statistics

In the next section the cluster results are presented, one containing the "problematic***" values and one conducted only on the subsurface waters' anions and cations used for facies determination.

During cluster analysis, the first step was to examine the datasets containing the parameters with values lower than the level of detection, and then to replace non detectable values with half of the detection limit (as mentioned above); as a result, the amount of data available for the cluster analysis increased. Two out of three strongly correlating parameters were removed (conductivity and Na^+) in order to reduce their collective effect. This is the reason why these cannot be found in the following analyses (e.g. Table 4).

Four groups were determined and placed on a map, three of which separated explicitly (Fig. 10). Different groups are marked with different colours. The fourth group contains outliers (e.g. shallow monitoring wells and a deep thermal well), thus there is as yet no explanation for its constitution. For the purposes of better interpretation, the three groups were given names after the geographical region they are located in and the fourth one Outlier. From

***In section 3.3.2.2 the term „problematic" refers to parameters with few datum under the detection limit. If their numbers is high the dataset's variance is low.

now on these will be referred to using the following names. The **first group** is the **Maros group** in the area of Maros alluvial fan and Makó graben. The **second group** is the **Körös group** in the area of the Körös basin. The **third group's** wells are geologically and geographically situated in the Maros alluvial fan, however it was given the name **Arad group** after the city around which they are located. As mentioned before, the clustering was repeated with a dataset without the "problematic" values. This time only anions and cations used for subsurface waters' facies determination were considered to be valid. Results were again visualized on a map (Fig. 11).

In the next paragraph the similarities and differences will be discussed between the results of the two clusterings (Fig. 10 & 11). The main difference is that the wells in the **second group (Körös)** in Fig. 10 after the second clustering (Fig. 11) with the decreased database (containing parameters: Ca^{2+}, Mg^{2+}, Na^{+}, K^{+}, HCO_3^{-}, Cl^{-} and SO_4^{2-}.) connect to the **first (Maros)** and the **fourth (Outlier) groups'** wells. Nevertheless a few wells kept their original relationship and stayed in the **second group (Körös)**.

Using the database without the "problematic" parameters, the data was plotted on a Piper diagram (Fig. 12) (the colour codes of the groups were retained). It is easy to see that the spatially separated groups are no longer present. This is no surprise because the application of the two methods has different intentions.

To verify the grouping, discriminant analysis was used, which pointed out that 94.6% of the original grouped cases were correctly classified using the grouping based on the data including the "problematic" parameters. Then the contents of the actual groups were changed according to the software's (SPSS) suggestions. Finally as a result 100% of the original grouped cases were proved to be correctly classified with the cross-validation resulting in a value of 96%.

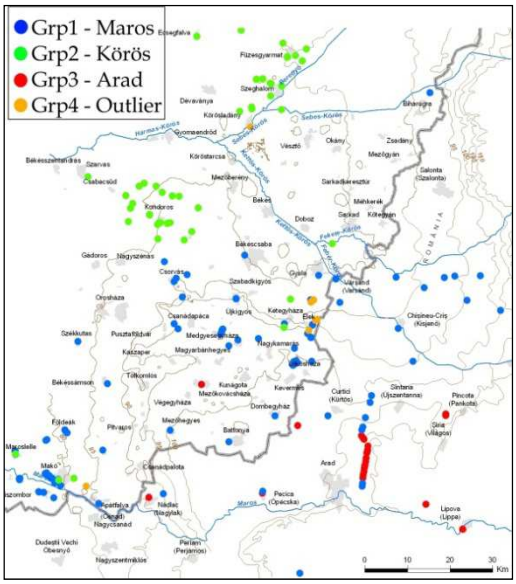


Fig. 10. Cluster results obtained from the database including the "problematic" parameters, where the same colour indicates the same group.

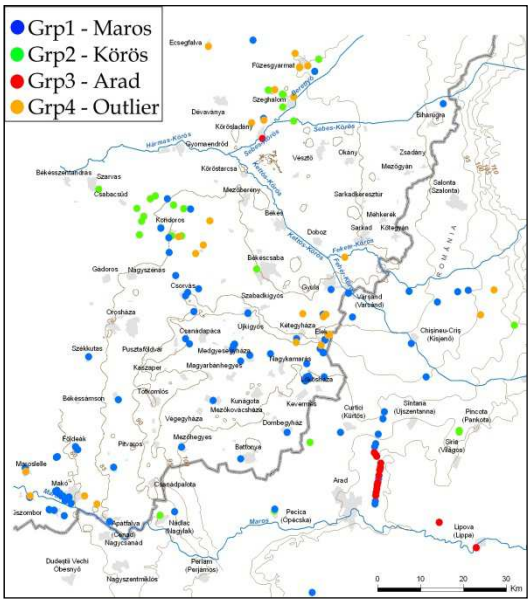


Fig. 11. Cluster results obtained from the database excluding the "problematic" parameters. The same color indicates the same group.

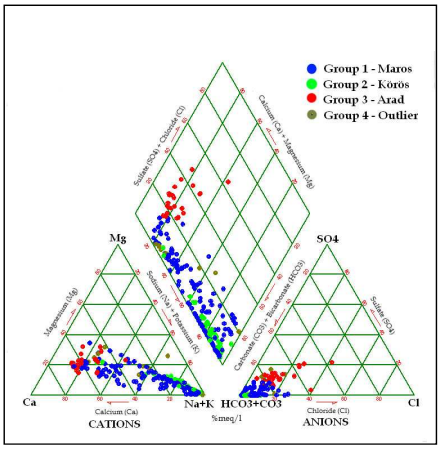


Fig. 12. Piper diagram obtained from the database excluding the "problematic" parameters

The result of the discrimination analysis can be visualized on Fig. 13 on the surface stretched between the first two discriminating planes (function 1 & function 2). Separation of the different groups can be easily observed. The wells in the **Outlier group** separated significantly from all the other groups.

In order to demonstrate the effect of the individual parameters on the formation of the cluster groups, a Wilks' lambda quotient was calculated for each parameter (Table 4).

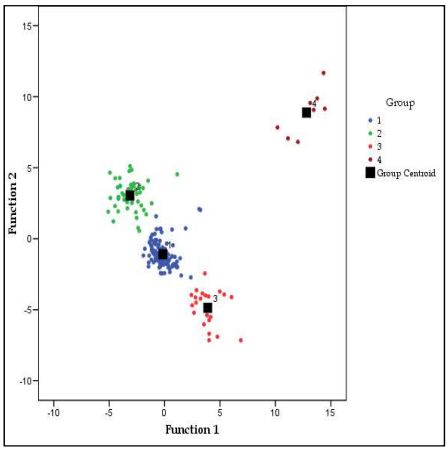


Fig. 13. Visualized results of the cross-validated discriminant results with the "problematic" parameters included.

Parameter	Measurement unit	Wilks' λ coefficients
NO_3^-	mg l^{-1}	0.249
HCO_3^-	mg l^{-1}	0.362
Fe_{total}	mg l^{-1}	0.404
SO_4^{2-}	mg l^{-1}	0.544
NH_4^+	mg l^{-1}	0.608
pH		0.610
Ca^{2+}	mg l^{-1}	0.690
As_{tot}	µg l^{-1}	0.723
Mg^{2+}	mg l^{-1}	0.748
F^-	mg l^{-1}	0.759
K^+	mg l^{-1}	0.797
Cl^-	mg l^{-1}	0.808
As (III)	µg l^{-1}	0.817
As (V)	µg l^{-1}	0.818
Dissolved O_2	mg l^{-1}	0.841
Br^-	µg l^{-1}	0.852
NO_2^-	mg l^{-1}	0.880
Water temperature	°C	0.889
Total organic carbon	mg l^{-1}	0.908
Mn^{3+}	mg l^{-1}	0.935

Table 4. Wilks' lambda quotients of each parameter in increasing order

As expected the most influencing parameter was the organic NO_3^- . Among the inorganic chemical components, the cluster grouping was notably influenced by anions such as: HCO_3^-, $SO_4^{2-}$and cations Mg^{2+} , Ca^{2+} and pH and Fe_{total} as well.

At this point enough results had been obtained to be able to see clearly why the two cluster results (Figs 10 & 11) show a resemblance to each other. The reason for this is that the main parameters that are necessary for determining hydrogeological facieses play an important part in forming the groups (based on the Wilks' lambda quotients), while the origin of the differences is the other influencing parameters. Spatial separation is much more obvious in the latter case. It is suggested to present the statistics of the different groups' parameters on box-and-whiskers diagrams. As an example HCO_3^- and As (V) are presented (Figs 14 & 15). Based on these results, the individual groups can be described.

The most important aim of the introduced area's hydrogeological investigation was to determine the distribution of arsenic in the groundwater. The groundwater's different As (V) concentrations can be seen in Fig. 15. It is obvious that the arsenic accumulated mostly in the waters of the deep and a few shallow wells (**Outlier group**), and the least in the **Arad group's** wells. In the **Arad group** arsenic was traceable in 5 out of 33 wells. Two of the five sampled wells fall to the area of the Maros alluvial fan. The different forms of arsenic show great variance in the groundwater of the other areas but the arsenic content of groundwater is clearly higher in the area of Körös basin.

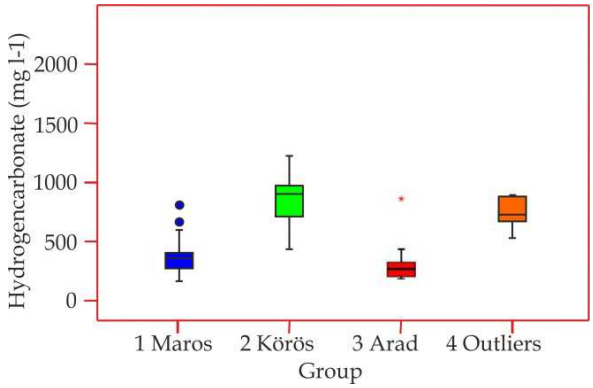


Fig. 14. Staticstics of hydrogencarbonate for each group

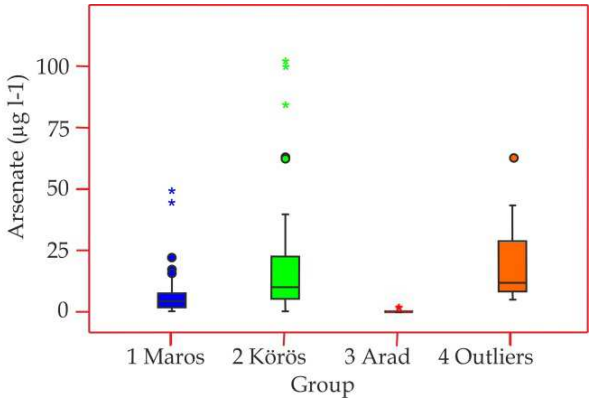


Fig. 15. Statistics of Arsenic (V) for each group

3.3.2.3 Principal component analysis

According to the literature, two out of the three **groups (Arad and Maros)** are located in different parts of the same flow system, while the third represents another flow direction. It is worth examining the dominant processes taking place in the different groups. However, a new problem had to be faced: a few parameters' variance was critically small in the different groups. In the cases of many hydrogeologically important parameters, their standard deviation is too small, so they had to be discarded. Examples included: sulphate in the **Körös group**, or As (III) and As (V) in **Arad group**, but the same thing can be said regarding the total iron content in the Arad area.

The most important results of the PCA – regarding the parameters and groups – are presented in Table 5. Absolute values of factor loadings higher than 0.71 are indicated in bold and **red**. Regarding the PCA result it can be said that no consistent conclusion can be drawn without discrepancy, either globally, or regarding the different groups and parameters.

	Group 1 (Maros)			Group 2 (Körös)			Group 3 (Arad)		
	PC 1	PC 2	PC 3	PC 1	PC 2	PC 3	PC 1	PC 2	PC 3
Water temperature	**-0.721**	0.267	-0.374	**0.714**	-0.277	0.061	0.117	**0.860**	-0.289
Conductivity	0.287	**0.940**	0.112	0.644	**0.737**	-0.136	**0.981**	-0.112	0.011
pH	**-0.748**	-0.061	0.173	0.365	-0.673	0.339	-0.272	0.489	0.481
Dissolved O_2	-0.184	-0.277	0.420	-0.395	0.325	-0.378	-0.656	-0.423	0.410
Na^+	-0.498	**0.850**	0.076	**0.891**	0.393	-0.155	**0.888**	0.260	-0.345
K^+	**0.796**	0.132	-0.165	0.289	0.585	0.575	0.090	-0.572	-0.593
Ca^{2+}	**0.928**	-0.263	-0.021	**-0.791**	0.504	0.160	**0.795**	-0.488	0.129
Mg^{2+}	**0.902**	0.026	-0.114	**-0.729**	0.602	-0.141	**0.898**	-0.083	0.322
Mn^{3+}	**0.872**	0.077	0.150	**-0.713**	0.319	0.391	**0.812**	0.259	0.310
Cl^-	0.159	0.329	**0.842**	0.388	0.520	0.469	**0.799**	0.038	-0.226
HCO_3^-	0.278	**0.885**	-0.240	0.594	**0.741**	-0.230	**0.829**	-0.112	0.385

Table 5. The factor scores of the parameters which suited the conditions of the PCA in each cluster group

3.3.3 Conclusions regarding the results obtained from the groundwater system's analysis of SE Hungary and the bordering Romanian area

The explicit separation of the groundwater characteristics in the different parts of the Great Hungarian Plain has been a well-known fact for a long time (Rónai, 1985). Based on the dissolved cations, the individual water types are related to the three extended hydrogeological units (Duna-Tisza interfluve South-Tiszántúl, Maros alluvial fan, Körös basin) (Rónai 1985; Varsányi & Ó. Kovács, 2006). Our investigations confirm this. Furthermore, the investigations expand what is already known with the result from the Romanian area. Based on the results of the applied multi-variate data analysis methods, the groundwater sampled in the Makó graben originating from the Duna-Tisza interfluve does not separate from the characteristic water type of the Maros alluvial fan.

Based on hydraulic modelling, results obtained from the upper 600 m (screen depths of drinking water supply wells) and water age data, the Maros alluvial fan consists of one uniform gravitational flow system. Towards Romania hydraulic heads gradually increase.

Based on the literature and results obtained, it seems that the **Arad group** is located at the beginning of the regional flow system, while its middle part and discharge area is situated in the Hungarian part of the Maros alluvial fan. Along the flow path in the Maros alluvial fan, depending on the quantity of bounded cations in the clay minerals, the ions with one and two valences may interchange. This results in the systematic change of the dissolved cations' concentration in the direction of the flow. In the cases of the $Ca(HCO_3)_2$ and $Mg(HCO_3)_2$ water types, the concentration of cations with two valences decreases, while the concentration of Na^+ increases in the direction of the flow (Varsányi, 2001).

The Körös basin, (bordering the Maros alluvial fan), is an individual hydrogeological system that is – based on the water ages and high Na^+content of the groundwater – situated at the end of a NE-SW oriented gravitational flow system. Besides the gravitational flow, the area can be characterized with slow up flow originating from sediment compaction.

These statements do not contradict the results explained above; nevertheless, our results show that differences in the three areas' groundwater chemistry are affected not only by gravitational flow systems. It is important to mention that data analysis methods may provide significant extra information during the exploration of a certain area's hydrogeological conditions, but separating different flow systems and flow regimes based only on data analysis is not possible.

The results of the PCA may reveal background processes taking place in a gravitational flow system, like cation change processes, or the role of Na^+, which has an important place in both **Körös** and **Arad group**. In the case of the **Arad group** this fact contradicts with the group's location in the flow system. Regarding other parameters (for example high chloride and sulphate concentrations), anthropogenic contamination can be in the background or regarding river's ablation area geological origin is feasible as well.

In order to determine the origin of the contaminants further investigations are needed. The place of the **Körös group** in the flow system does not contradict the high factor score of the Na^+ , however this high value in comparison to its factor score in the Maros alluvial fan implies a background process, the albite-montmorillonite reaction in the sediments of the basin at depths of 60-500 meters (Varsányi, 2001).

4. Summary

In our chapter we introduced a few methods known for decades in earth sciences and geology (Davis, 2003). We developed an order of application, which seemed beneficial during our work. The results obtained were of great use in studies when a hypothesis needed verification or discarding. When choosing the cases studies our direct aim was to present data sets with problems commonly faced by scientist through the three water environments: Kis-Balaton Water Protection System, River Tisza and a groundwater system of SW Hungary and SW Romania. We hope this chapter will be of use for every scientist who has to work with water quality data.

5. Acknowledgements

We the authors would like to thank Paul Thatcher for his work on our English versions, Antal Füst D.Sc. for his helpful advices, and say thanks for the stimulating discussions with our colleagues and students at the Department of Physical and Applied Geology of Eötvös Loránd University.

6. References

Afifi, A., Clark, V. A., May, S. & Raton B. (2004). Computer-Aided Multivariate Analysis (4th ed.). Chapman & Hall/CRC, p. 489. ISBN 1-58488-308-1, USA

Davis, J. (2003). *Statistics and data analysis in geology*, Wiley, p. 656, ISBN 978-0-471-17275-8, USA

Dryden, I., Márkus, L., Taylor, C. & Kovács, J. (2005). Non-Stationary spatio-temporal analysis of karst water levels, *Journal of the Royal Statistical Society*, Series C-Applied Statistics, Vol.54., No.3., pp. 673-690, ISSN 1467-9876

Füst, A. (1997). *Geostatistics* (in Hungarian: *Geostatisztika*). Eötvös Kiadó, p. 427, Budapest Hungary

Füstös, L., Meszéna, Gy. & Simonné, M. N. (1986). *The methods of multi-variate data analysis* (in Hungarian: *A sokváltozós adatelemzés statisztikai módszerei)*, Akadémiai Kiadó, ISBN 963-05-3859-8, Budapest Hungary

Geiger, J. (2007). *Geomathematics (in Hungarian: Geomatematika)*. Jate Press, p.116, Szeged Hungary

Hans, G. (2005) *Uni-variate Time Series in Geosciences*, Springer-Verlag, p. 718, ISBN 3540238107, New York USA

Hatvani, I. G., Kovács, J., Kovácsné Székely, I., Jakusch, P. & Korponai, J. (2011). Analysis of long term water quality changes in the Kis-Balaton Water Protection System with time series-, cluster analysis and Wilks' lambda distribution. *Ecol. Eng.*,Vol. 37, No.4, pp. 629-635, ISSN 0925-8574

Helsel, D. R. & Hirsch, R. M. (2002). *Statistical Methods in Water Resources Techniques of Water Resources Investigations*, Book 4, chapter A3. U.S. Geological Survey. 522 p.

Isvánovics, V., Honti M. & Vörös L. (2010). Phytoplankton dynamics in relation to connectivity, flow dynamics and resource availability—the case of a large, lowland river, the Hungarian Tisza. *Hydrobiologia*. Vol. 637, No. 1. pp. 121-141, ISSN 0018-8158

Ketskeméty, L. & Izsó, L. (2005). *Introduction into the SPSS system* (in Hungarian: *Bevezetés az SPSS programrendszerbe)*, ELTE Eötvös Kiadó, ISBN 963-463-823-6, Budapest Hungary

Körös Valley District Environment and Water Directorate [KÖRKÖVIZIG], Makó Térségi Vízmű, Smaragd-GSH Kft., Bálint Analitika, Selor eeig, Geological Survey of Finland, Regia Autonomia Apa – Canal Arad (2008): LIFE 05 ENV/H/000418: Sustainable management and treatment of arsenic bearing groundwater in Southern Hungary (SUMANAS)

Korponai, J., Braun, M., Buczko, K., Gyulai, I., Forro, L., Nedli, J. & Papp, I. (2010). Transition from shallow lake to a wetland: a multi-proxy case study in Zalavari Pond, Lake Balaton, Hungary. *Hydrobiologia* Vol. 641, No. 1. pp. 225–244, ISSN 0018-8158

Kovács, B. & Szanyi, J. (2005). *Hydrodinamic and transport modeling in MODFLOW environment* (in Hungarian: *Hidrodinamikai és transzportmodellezés Processing MODFLOW környezetben*) II., Miskolci Egyetem – Szegedi Tudományegyetem – GÁMAGEO, p. 160, ISBN 963 661 637 X

Kovács, J. & Kovácsné Székely, I. (2006). The problems in interpreting the sample: theory and practice (in Hungarian: A minta értelmezési problémái: elmélet és gyakorlat). *Földtani Közlöny*, Vol. 136, pp. 139–146, ISSN 0015-542X

Kovács, J. (2007). The application of modern geomathematical methods in practical hydrogeology, Ph. D. dissertation p. 122 (manuscript)

Kovács, J., Czauner, B., Kovácsné Székely, I., Borsodi, A. & Reskóné Nagy, M. (2008). The description of the changes in the different water quality areas of Lake Balaton and their time series' patterns (in Hungarian: A Balaton eltérő vízminőséggel rendelkező térségeinek változásai és a mért vízminőségi idősorok mintázatai, 1985-2004 között). *Hidrológiai Közlöny*, Vol. 88,No. 6, pp. 172-174, ISSN 0018-1323

Kovács, J., Hatvani, I. G., Kovácsné Székely, I., Tanos, P. & Korponai, J. (2011). Key question of sampling frequency estimation during system calibration, on the example of the Kis-Balaton Water Protection System's data series. *Georgikon for Agriculture* Vol. 14, No. 1. pp. 53-68 ISSN 0239-1260

Kovács, J., Hatvani, I. G., Korponai, J. & Kovácsné Székely, I., (2010). Morlet wavelet and autocorrelaiton analysis of long term data series of the Kis-Balaton Water Protection System (KBWPS). *Ecol. Eng*., Vol. 36, No.10, pp. 1469-1477, ISSN 0925-8574

Kovács, J., Márkus, L. & Halupka, G. (2004). Dynamic Factor Analysis for Quantifying Aquifer Vulnerability. *Acta Geol. Hung*, Vol. 47, No. 1, pp. 1–17, ISSN 1588-2594

Lászlóffy, W. (1982). *The Tisza River. Water development and management in its river basin* (in Hungarian: *A Tisza, vízi munkálatok és vízgazdálkodás a tiszai vízrendszerben*) Akadémiai Kiadó, p.610, Budapest Hungary.

Lin, Z. & Puls, R. W. (2000). Adsorption, desportion and oxidation of arsenic affected by clay minerals and aging process. *Environmental Geology,* Vol. 39 , No. 7 , pp 753-759, ISSN 1866-6299

Lotz, Gy. (1988). The Kis-Balaton Water Protection System (in Hungarian: A Kis-Balaton Vízvédelmi Rendszer). *HidrológiaiTájékoztató,* Vol. 28, No.2

Márkus, L., Berke, O., Kovács, J. & Urfer, W. (1999). Spatial Prediction of the Intensity of Latent Effects Governing Hydrogeological Phenomena. *Environmetrics*, Vol. 10, No. 5., pp. 633-654, ISSN 1180-4009

Nádor, A., Thamó-Bozsó, E., Magyari, Á. & Babinszki, E. (2007). Fluvial responses to tectonics and climate change during the Late Weichselian in the eastern part of the Pannonian Basin (Hungary). *Sedimentary Geology*, Vol. 202 , No. 1-2 , pp. 174–192, ISSN 0037-0738

Nguyen, H. L., Leermakers, M., Kurunczi, S., Bozo, L. & Baeyens, W.,(2005). Mercury distribution and speciation in Lake Balaton, Hungary. *Science of the Total Environment,* Vol. 340, No.1-3, pp. 231-246, ISSN 0048-9697

Norušis, M. J. (1993). SPSS for Windows Professional Statistics Release 6.0. SPSS Inc. p. 385. ISBN 0-13-178831-0, USA

Padisák J. (2005). *General limonology* (in Hungarian: *Általános limnológia)*. ELTE Eötvös Kiadó, p. 310 ISBN 963-463-721-3, Budapest Hugary

Padisák, J. & Reynolds, C. S. (2003). Shallow lakes: the absolute, the relative, the functional and the pragmatic, *Hydrobiologia, Vol.506-509 , No.1-3* , 506-509, 1-11, ISSN 0018-8158

Pécsi, M. (Eds.) (1969) *The River Tisza in the Great Hungarian Plane* (in Hungarian: *A tiszai Alföld)*. Akadémiai Kiadó, p. 381, Budapest Hungary.

Pernthaler, J. (2005). Predation on prokaryotes in the water column and its ecological implications. *Nature Reviews Microbiology*, Vol. 3, pp. 537–546 ISSN 1740-1526.

Pomogyi, P., Szeglet, P. & Csato, E. (1996). The change in the reed community of the KBWPS, Feneki pond, based on vegetation-mapping results (In Hungarian). In: Pomogyi, P. (Ed.), 2. Kis-Balaton Conference: Summary of the Research Result on the KBWPS from Between 1991–1995. University of Pannonia Georgikon, Faculty of Agriculture, Keszthely, pp. 206–207.

Pukelsheim, F. (1994). The Three Sigma Rule. *The American Statistician*, Vol. 48, No. 2 pp. 88-91, ISSN 00031305.

Ramsar Convention, "Convention on Wetlands of International Importance especially as Waterfowl Habitat. Ramsar (Iran), 2 February 1971. UN Treaty Series No. 14583.As amended by the Paris Protocol, 3 December 1982, and Regina Amendments, 28 May 1987."

Ritter, A., Munoz-Carpena, R. & Bosch, D.D. (2007) Agricultural land use and hydrology affect variability of shallow groundwater nitrate concentration in South Florida, *Hydrological Processes* Vol. 21, No. 18., pp. 2464-2473, ISSN: 1099-1085

Rónai, A. (1985): The Quaternary of the Great Hungarian Plain. Acta *Geol. Hung., Ser. Geol.*, Vol. 21, p. 446, ISSN 1588-2594, Budapest Hungary

Sakan, S., Gržetić, I. & Đorđevoć, D. (2007). Distribution and Fractionation of Heavy Metals in the Tisa (Tisza) River Sediments. *Environmental Science and Pollution Research*, Vol. 14, No. 4., pp. 229-236.

Shumway, R. H. & Davis, S. S. (2000) Time Series Analysis and Its Applications, Springer-Verlag, p. 549, ISBN 0-387-98950-1, New York USA

Stockburger, D. W. (2001). Multivariate Statistics: Concepts, Models, and Applications, *Missouri State University*, Available from <http://www.psychstat.missouristate.edu/multibook/mlt00.htm>

T. Nagy, M., Csépes, E., Aranyné Rózsavári, A., Bancsi, I., Kovács, P., Végvári, P. & Zsuga, K., (2004). A hosszú-távú adatsorok értékelésének korlátai. *Hidrológiai Közlöny*, Vol. 84. No. 5-6., pp. 162-165.

Tátrai, I., Mátyás, K., Korponai, J., Paulovits, G. & Pomogyi, P. (2000). The role of the Kis-Balaton Water Protection System in the control of water quality of Lake Balaton. *Ecol. Eng.*, Vol.16, No.16, pp. 73–78, ISSN 0952-8574

Varsányi, I. & Ó. Kovács, L. (2006). Arsenic, iron and organic matter in sediments and groundwater in the Pannonian Basin , Hungary, *Appled Geochemistry*, , Vol. 21, No. 6, pp. 949-963, ISSN 0883-2927

Varsányi, I. (2001). Groundwater of the southern Great Hungarian Plain: hydrogeochemical processes and hydrogeological conclusions. D.Sc. Thesis, Szeged, p. 126. (manuscript)

Vollenweider R. A. (1976). Advances in defining critical loading levels of phosphorus in lake eutrophication. *Mem. Ist. Ital. Idrobiol,* Vol. 33, No. 1. pp. 53-83.

Vollenweider, R. A. & Kerekes, J. (1982). Eutrophication of waters. Monitoring, assessment and control. OECD Cooperative programme on monitoring of inland waters (Eutrophication control), Environment Directorate, OECD, Paris, 154 p.

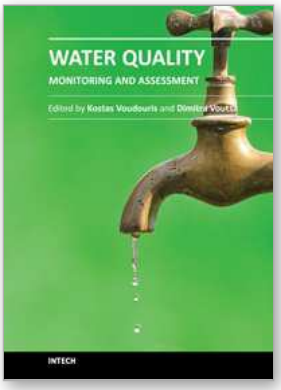
Voudouris, K., Panagopoulous & A., Koumantakis, J. (2000). Multivariate Statistical Analysis in the Assessment of Hydrochemistry of the Northern Korinthia Prefecture Alluvial Aquifer System (Peloponnese, Greece). *Natural Resources Research*, Vol. 9, No. 2. Pp.135-146. ISSN 1573-8981

Voudouris, K.S., Lambarkis, N.J., Papatheothorou & G., Daskalaki, P. (1997). An Application of Factor Analysis for the Study of the Hydrogeologycal Conditions in Pilo-Pleistocene Aquifers of NW Achaia (NW Peloponnesus, Greece). *Mathematical Geology*, Vol. 29. No. 1., pp. 43-59. ISSN 0882-8121

Wetzel, R. G. (2001). *Limnology, Lake and River Ecosystems*. 3rd edition. Academic Press, Elseveir Science, p.1006 ISBN 0-12-744760-1, California USA

Wilkinson, D.J. (2006). Stochastic Modelling for Systems Biology. Chapman & Hall/CRC, p. 254. ISBN 978-158488-540-5, USA

Zsuga, K. & Szabó, A. (2005). The ecological problems of the Tisza's Hungarian cathchment area, and its environmental problems (in Hungarian: A Tisza hazai vízgyűjtőterületének ökológiai állapota, környezetvédelmi problémái.) *Hidrológiai Közlöny*, Vol. 85. No. 6., pp. 168-170, ISSN 0018-1323



Water Quality Monitoring and Assessment
Edited by Dr. Voudouris

ISBN 978-953-51-0486-5
Hard cover, 602 pages
Publisher InTech
Published online 05, April, 2012
Published in print edition April, 2012

The book attempts to covers the main fields of water quality issues presenting case studies in various countries concerning the physicochemical characteristics of surface and groundwaters and possible pollution sources as well as methods and tools for the evaluation of water quality status. This book is divided into two sections: Statistical Analysis of Water Quality Data;Water Quality Monitoring Studies.

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

József Kovács, Péter Tanos, János Korponai, Ilona Kovácsné Székely, Károly Gondár, Katalin Gondár-Sőregi and István Gábor Hatvani (2012). Analysis of Water Quality Data for Scientists, Water Quality Monitoring and Assessment, Dr. Voudouris (Ed.), ISBN: 978-953-51-0486-5, InTech, Available from: http://www.intechopen.com/books/water-quality-monitoring-and-assessment/analysis-of-water-quality-data-for-researchers

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821