Integrating Data from Suquía River Basin: Chemometrics and Other Concepts

María Verónica Baroni and Daniel A. Wunderlin

Abstract Assessing the water quality in a river basin seems to be an easy tool. However, some degree of expertise is required when planning and executing several tasks necessary to avoid either excessive or insufficient data. In this chapter we briefly describe some aspect of sampling and sample preparation, which have been discussed in deep in previous chapters. However, after sampling a river, it is necessary to analyse several parameters, arising from many monitoring stations, sampled at different time, etc. All of this generates an amazing database that needs to be carefully explored, looking to extract the most relevant information on changes in the water quality, probable pollution sources, temporal and spatial changes and so on. One simple approximation is the construction of water quality indices from both chemical and biological data, deep discussed in Bistoni et al. (Handb Environ Chem. https://doi.org/10.1007/698_2016_455, 2016) and Amé and Pesce (Handb Environ Chem. https://doi.org/10.1007/698_2015_434, 2015). The other way is using multivariate statistics (chemometrics), trying to evidence which changes are occurring, where and when. Here we discuss several chemometrics methods used to verify changes in the water quality of the Suquía River basin, starting with non-supervised methods, like cluster analysis (CA), factor analysis (FA, including principal components analysis – PCA) and going to supervised methods, namely, discriminant analysis (DA) and generalized procrustes analysis (GPA). Although PCA has become the most popular method of analysis for the evaluation of water quality, we think that PCA should be complemented with other methods that help to corroborate results from PCA. In this case, we used CA as a primary tool to evidence

M.V. Baroni and D.A. Wunderlin (✉)
ICYTAC: Instituto de Ciencia y Tecnología de Alimentos Córdoba, CONICET, Universidad Nacional de Córdoba, Ciudad Universitaria, Córdoba 5000, Argentina

Departamento de Química Orgánica, Facultad de Ciencias Químicas, Universidad Nacional de Córdoba, Ciudad Universitaria, Córdoba 5000, Argentina
e-mail: vbaroni@fcq.unc.edu.ar; dwunder@fcq.unc.edu.ar

D.A. Wunderlin (ed.), *The Suquía River Basin (Córdoba, Argentina): An Integrated Study on Its Hydrology, Pollution, Effects on Native Biota and Models to Evaluate Changes in Water Quality*, Hdb Env Chem, DOI 10.1007/698_2017_202,

spatial differences in the water quality along the basin, confirming these results by PCA, which also added evidence on temporal differences. DA allowed further confirmation of both temporal and spatial changes, with an important data reduction, which is important for the survey of a river basin when the budget is restrictive. Finally, GPA brings further confirmation of other chemometrics methods, enabling a clear differentiation between water quality at diverse river sections, during both dry and rainy season. So far, we truly expect that this chapter helps readers to better design future surveys to evaluate changes of the water quality in other rivers worldwide.

Keywords Chemometrics, Integrated evaluation, Multivariate statistics, River basin, Water quality

Contents

1 Introduction

River basins function with complex interactions between biotic and abiotic components. In order to understand these complex interactions, it is often necessary characterizing the basin considering its geology, hydrology, biology and human activities. Some measurements can be made directly in the basin by using sensors or collecting samples that can be analysed at the laboratory and in the field as well. For instance, pH, conductivity, water flow, temperature, etc. can be field measured; while organic and inorganic elements require laboratory work for measuring. Furthermore, the community structure of fish can be field obtained but needs further analysis at the laboratory to point out changes, etc. Other biological elements need to be sampled in the field but transferred to the laboratory for a detailed analysis (plankton structure, macroinvertebrates, microorganisms, etc.). Thus, an integrated analysis of a river basin requires the assemblage of several data, from different origin, units, magnitude, etc. So, it will be necessary to have appropriate statistical methods to examine the relationships among the parameters measured to describe the functioning of the chemical-biological processes that gave rise to the observed values, behaviour and changes [1].

It is noteworthy that the use of mathematical and statistical methods, including chemometrics and many other statistical methods/algorithms, in environmental and

other sciences has increased steeply during the last years. Conducting a quick search in the most recognized scientific databases (PubMed, ScienceDirect, Scopus), it is possible to observe that statistical methods have gained a huge space in different areas. Statistical analysis is at the core of most modern data mining models [2, 3].

During the last years, with the development of computational software, we can perform more complex statistical analysis, exploring our data deeper and in a more complex way. However, this means that the statistical models and analysis are complex too; therefore, we need to be familiar with the potential and limitations of a much greater range of statistical approaches [2, 4].

Furthermore, scientific journals demand proving that the experimental differences between data sets are statistically significant; thus, an increasing number of software tools and packages have been developed to cover this need. However, modern, user-friendly software has led to a generation of "click and go" users, who are eagerly destined to obtain the *P*-values and multivariate plots (projection of samples and variables on the factor plane) but with less or no idea on how the statistical parameters are calculated. Furthermore, some researchers do not know the theoretical and practical reasons for performing such tests [2, 5, 6].

Computational tools available can be used not only to run statistical analysis such as univariate and bivariate tests but also multivariate calibration and development of complex models, simulating different scenarios that consider a set of inputs or simply making predictions for specific data sets or conditions. Therefore, one should avoid and forget the word "test" and replace it with analysis. A test implies something simple and unified and gives a clear answer related to a *P*-value, something rarely for environmental data. In practice, one has to apply data exploration, check assumptions, validate the models, perhaps apply a series of methods and most importantly interpret the results [7].

For instance, ecologists evaluating the community composition in aquatic systems need to include the characterization of individual populations, the environment and, of course, relating the observed biological variation to the environmental characteristics. One has to take into account this multidimensionality, so a univariate analysis does not work in most cases. The most appropriate methods for statistical analysis of such data are the "multivariate statistical methods" [8].

One important point when analysing an extensive data set, like those produced during basin monitoring, is the data reduction. Data reduction means that not all obtained data are to be used to show the main basin characteristics, its variations and changes (e.g. temporal and spatial changes). In spite of using the whole data set, mathematically- statistically methods can be used to select those variables that are most representative of such changes, enabling a good analysis from a reduced data set, which points out critical variables that fit for the analysis purpose. Multivariate statistics helps scientists to discover the data structure, helping to reduce the amount of data but keeping the important information for easier comprehension. Multivariate analysis uses relationships between variables to order the objects of study according to the results of the measured variables and to classify monitoring sites, biological species or ecosystems in distinct classes, each containing entities with similar chemical and/or physiological characteristics. However, multivariate analysis is complicated in theoretical structure and in operational methodology [8].

Most of the statistical tests are based on a set of assumptions about the data that must be met prior to the application of the statistical analysis and testing of a hypothesis. For example, most of the parametric statistics analyses have an assumption that the data follow a certain distribution. Other assumptions include homoscedasticity, linearity and independence. The non-compliance of these assumptions may have little impact on the results or conclusions; yet others may arise the possibility of making errors type I or II (false positive and false negative), leading to incorrect inferences about the results and thus undermine meaningful research [3, 9]. In this sense, statistical procedure should be use to check that the statistical analyses meet the assumptions, and the results from these procedures need to be reported to verify that the conclusions raised from the statistical analysis are valid [9].

There are new statistical methods that can be applied to answer almost every concern in studies related to aquatic systems; however, the greatest challenge is to figure out how the various statistical methods relate to each other, determining which method is most appropriate for any particular problem.

Before questioning which statistical analysis should I apply? The scientist should answer: What are the underlying questions? What do I want to show? What am I looking for? Answering these questions will give you the basis for deciding the most appropriate statistical approach. Our objective as scientists is to be able to use these tools efficiently, without losing sight of the vision, that is, the motivation of the research done (e.g. base chemical/biological monitoring, evaluation of temporal and spatial changes along a basin, integrating chemical with biological data, etc.).

On the other hand, sometimes it is necessary to present our results in a simple way, which can be easily understood by the people, policy makers, magistrates, etc. Under these circumstances, we can choose non-statistical models, like the use of an index that reflects the overall quality status of a basin, temporal and spatial changes, etc. The use of water quality indice, for instance, can provide this kind of practical approach to show changes along a basin, a temporal trend during monitoring surveys, etc. [10]. However, it is worth to mention that the validity of these indices needs to be confirmed by a more strength mathematical-statistical model to demonstrate their usefulness and accuracy when informing results to the population.

This chapter will provide both a conceptual and practical understanding of the application of quality indices and multivariate statistics in the evaluation of changes along a river basin.

We hope that, after reading this chapter, you should be able to understand and know the validity and limitations of water quality indices as well as the assumptions of different statistical methods, identify the appropriate model for the sampling design, interpret the output of the method used for data mining and design the monitoring and analysis programme that best fits for the purpose, lowering cost and maximizing information.

2 Monitoring a River Basin

The design of a monitoring programme should fit for purpose. If the aim of the survey is bringing baseline information on the current status of the basin, or checking changes with respect to previous monitoring campaigns or evaluating the usefulness of the water for a specific purpose (e.g. drinking water, irrigation, industrial uses, tourism/recreation, aquaculture, etc.), it should be considered that each of these different purposes involves different water quality requisites, and, thus, the sampling strategy and analytical methods may differ. Some classical literature can be advised for those intending to start with the monitoring of a river basin [11, 12]. However, few practical tips are presented as follows:

1. Consider the river hydrology and its seasonal changes [13]. It is important to evaluate temporal changes considering variations in the river flow, which leads to dilution/concentration of contaminants, different degree of toxicity (related to the dilution of toxics), presence of particulate matter (PM) (affecting the absorption of diverse pollutants to its surface, in addition to transport phenomena of pollutants attached to PM, etc.), different land uses during winter and summer, different environmental conditions for the river biota, etc. (from bacteria to higher organisms, etc.). River hydrology also involves different areas within a basin (high, medium, lower basin), usually associated with different flow, turbulence, sediment structure (sand, clay, etc.) and many other conditions that can change from upper to lower basin. Consider that both native and non-native biotas are also affected by hydrological changes along the basin [14].
2. If natural or artificial lakes are present in the basin, their limnology should also be studied [15]. The study of limnologic issues can help to understand and prevent algae blooms, including those producing toxins that can affect both the native biota and human health. In addition, the study of the limnology helps to predict the quality of drinking water if the lake/reservoir is used for such purpose, etc.
3. Make an inventory of human activities along the basin. It helps to identify point pollution sources (industries, sewage exits, urban or agricultural run-off, etc.). This inventory usually helps to avoid unnecessary analyses in areas where some pollutants could not be expected. Of course, for a scientific work, it is necessary to consider pristine (reference) points to compare changes with respect to more impacted areas.
4. If possible, make an inventory of the biota inhabiting different basin areas, including both native and introduced biota. This is very helpful to evaluate how water quality changes are affecting the assemblages between different species, changes in dominance, endangered species, etc. [14].
5. Be prepared to evaluate changes in the water quality at different levels (surface water, sediment, interstitial water, etc.), at different areas (upper, middle and lower basin), with different pollution sources (human eutrophication, urban run-off, industrial effluents, wastewater, agricultural run-off, etc.). Each of these pollution sources could require different analytical methods and, accordingly, different budget to perform the monitoring survey [16–18]. In addition,

temporal changes should be considered as previously stated. To this point, when the budget is restrictive or even with less restrictive resources, a very careful analysis of requested information should be performed before sampling, never after sampling. Sometimes, professionals with less experience in water quality assessments just decide to get hundred samples, analysing a restricted set of parameters, losing valuable information that could be obtained from the same sampling campaign. Additionally, the possibility for safe sample storage should be considered, since after a first preliminary analysis, further elements could help to get a better diagnostic of the basin. If well-stored samples are available, preliminary results can be later completed with additional information without repeating the sampling, which would require the entire analysis of the new sample set. Moreover, stored samples can be used to compare changes along years when introducing new analytical methods, enhancing the scope of analysed parameters, etc.

6. Get enough number of samples from each environmental compartment (water, sediment, biota, etc.) at each area. Sampling campaigns are expensive, time consuming, requiring appropriate weather conditions, etc.; thus, getting only one sample from each area will be not enough if you need a serious, scientific-based, statistically sound result to be presented. On the other hand, excessive number of samples can be expensive to analyse, can be negative for endangered biota, etc. Just consider the appropriate number of samples necessary to get statistically representative results from your survey.
7. Repeat monitoring campaign for at least 2 years, better 3–4 different years. This is just for consider inter-year variations caused by different climatic conditions, etc.
8. Use appropriate tools for monitoring. Some practical tips can be found at WHO [12], Chapman [11] and recent scientific literature (see, for instance, [19–23]). It is quite common to find the use of metallic tools used to get samples designed for metal analysis. Particularly with trace elements, the use of metallic tools results in contamination of samples. In this case, pre-cleaned plastic tools should be used. On the contrary, to analyse organic elements, the use of plastic tools should be avoided as many organic compounds can be adsorbed onto some plastic surfaces, etc. [18]. Some particular cases need to be addressed. For instance, using glass bottles for sampling water for boron analysis should be avoided as the glass usually contains boron-silicate, which interferes with the measurement of boron. In addition, the common practice of adding mineral acids to stabilize samples should be considered using ultra-pure acids (ICP-MS grade) to avoid interferences when analysing trace elements by ICP-MS and related methods [17].

3 Data Mining from Monitored Basins

3.1 Use of Indices

Water monitoring for different purposes is well defined (e.g. aquatic life preservation, contact recreation, drinking water use) [11, 12]. However, the overall water quality is sometimes difficult to assess from a large number of samples, each containing concentrations for many parameters, including different magnitudes within analysed parameters [11]. Although any measured parameter is worth to be analysed by itself (univariate change using a single parameter, e.g. changes in water conductivity along a river basin), it is also quite common to analyse groups of parameters sharing a common feature (e.g. nitrogen load through the analysis of ammonia, nitrites, nitrates and organic nitrogen). Even so, the analysis of a single parameter provides only partial information on the overall water quality. Thus, the integration of several parameters to afford a better idea of changes along the basin seems a reasonable approach [10].

Mathematical-computational modelling of river water quality is possible but requires a previous knowledge of hydraulics and hydrodynamics [13]. Besides, mathematical models require extensive validation (see, for instance, [11, 24–32]).

The use of water quality indices (WQI) is a simple practice that overcomes many of the previous mentioned problems and allows the public and decision makers to receive water quality information, based on scientific criteria, from a complex data set, but presented in a simple form that can be used for regulatory purposes, public information, etc. [10]. WQI also permit assessing changes in the water quality along the basin, identifying trends and sudden changes produced by point pollution sources [11].

The use of WQI to assess the water quality in the Suquía River basin has been explained in [16] and will not be repeated here.

It is worth mentioning that other indices, like a Biotic Index, can be constructed using different parameters (e.g. fish diversity, etc.), not only chemical or physical parameters. This biotic index has been constructed for the Suquía River basin [33], considering valuable information on the biota present in the basin [14].

3.2 Multivariate Statistics (Chemometrics)

Even though WQI provide a useful way to predict changes and trends in the water quality considering multiple parameters, WQI do not provide evidences on the pollution sources, mainly because they are calculated after normalization of analytical values and weighting of such normalized values, according to the importance of the measured parameter for the aquatic life preservation [10, 14]. The use of multivariate techniques (also called chemometrics) is another method that can be adopted to evaluate water quality changes [11]. Other words used to describe some

multivariate methods is "pattern recognition techniques" as these methods allow discovering different patterns, representing different behaviour of the data set (although not all multivariate methods led to patterns). In the case of the Suquía River basin, the use of chemometrics allowed differentiating different parameters associated with diverse point and non-point pollution sources [34].

3.2.1 Exploratory Methods: Cluster Analysis

Cluster analysis (CA) can be divided in hierarchical and non-hierarchical. Hierarchical CA forms clusters sequentially, starting with the most similar pair of objects and forming higher clusters step by step. The similarity between two samples is usually given by the Euclidean distance, and a "distance" can be represented by the "difference" between analytical values from both samples [35]. One efficient way to calculate such distance is the Wards method, which uses analysis of variance to verify the distances between clusters, minimizing the sum of squares of any two (hypothetical) clusters that can be formed at each step.

Non-hierarchical CA methods, including fuzzy clustering, evaluate overall distributions of objects by pairs, classifying them into groups.

CA repeats the process of forming and joining clusters until obtaining a single cluster containing all samples. CA results in a dendrogram (tree diagram), providing a picture of the groups and its proximity, in addition to data reduction.

Many times, it is convenient to standardize the data set to obtain better results with CA [35–40]. Thus, grouped data presented in a dendrogram can be recognized as belonging to dry or wet season in seasonal analysis or belonging to the upper, middle or lower river basin during spatial analysis.

Some examples of other river basins studied using CA can be found in Astel et al. [41]; [42–51]. CA can be also used for the study of variations in groundwater, lakes, etc. (e.g. [52, 53])

Figure 1 presents results of both temporal and spatial analyses of samples corresponding to the Suquía River basin. Figure 1 was constructed using the same chemical and microbiological parameters used to construct WQI [16]; however, using biological and biochemical parameters is also possible (see, for instance, [54]). But here comes one of the first limitations of chemometrics methods: increasing the number of parameters considered (translated into columns in the matrix used for statistical calculation) also requires increasing the number of samples (each sample represents a row in the matrix, although data arising from different analytical replicates can be used as separate rows, accounting for both sample and analytical variations). In Fig. 1 samples belonging to a common area/season were grouped, namely, group "A" accounts mainly for samples from the lower basin (with 15% samples corresponding to the middle basin); group "B" accounts for samples from the high basin (with 25% samples from the middle basin). Thus, this approach helps to better visualize differences/similitudes between studied areas/seasons.

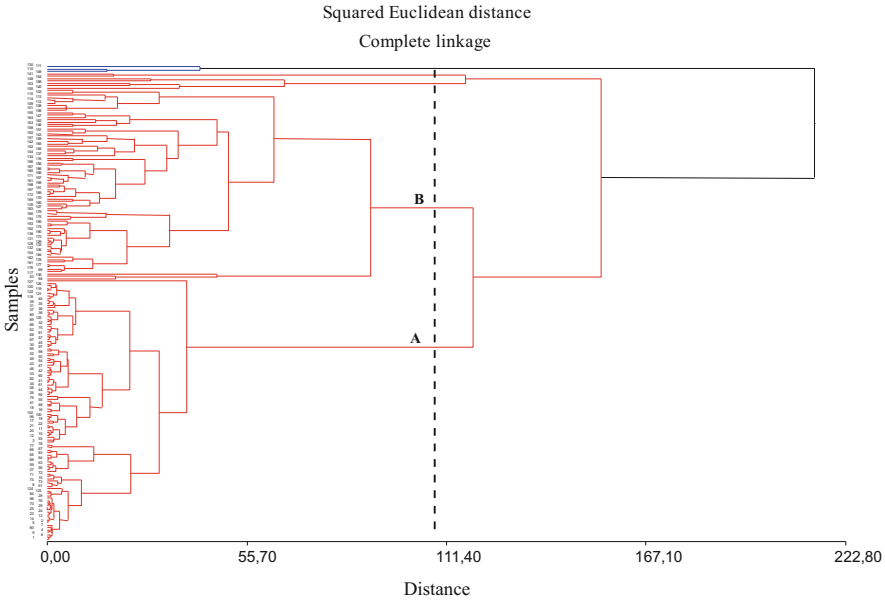

Fig. 1 Dendrogram constructed by cluster analysis (CA) of chemical and microbiological data evaluated in the Suquía River basin

As we can see from Fig. 1, data arising from 201 samples, taken along several years at the Suquía River basin (grouped according to basin area), are separated into two main groups, at approximately 50% of the maximal Euclidean distance (dotted line). What we can say from this plot is that the Suquía River Basin shows a clear difference in the water quality between the high and lower basin, meanwhile medium basin overlaps the other two zones.

It is also interesting to remark that CA does not show those parameters leading to the constructed dendrogram (Fig. 1). Thus, although CA can be used as a primary tool to see how to divide (grouping) our data set, it fails to show what is important and what is less important, to measure to reach such group separation.

3.2.2 Exploratory Methods: Factor Analysis/Principal Component Analysis

Factor analysis (FA), which includes principal components analysis (PCA), is used for data reduction without loss of information. Large data sets are common when evaluating temporal-spatial changes in river water quality, where several stations are included along the river basin to account for diverse environmental issues. In addition, several samples are taken at each site during one monitoring campaign, and many monitoring campaigns are organized to account for temporal changes, etc. In addition to the number of samples, several physical, chemical and biological parameters are usually evaluated, with two to five analytical replicates, etc. This

construct a big database that must be considered, evidencing changes along the basin or between seasons, pointing out key parameters to distinguish between different areas/seasons, etc.

PCA mathematically operates from the covariance matrix, which describes the dispersion of the multiple measured parameters, to obtain eigenvalues and eigenvectors. Linear combinations of the original variables and eigenvectors result in new variables, called principal components (PCs). A new set of axes, called factor axes, are obtained in a lower dimensional space onto which the original space of variables can be projected. Further rotation of the axis defined by PCA produces new groups of variables called varifactors (VFs). This last procedure is frequently known as factor analysis (FA), which is not the same with PC. The basic features associated with FA/PCA are data reduction and data grouping. Data reduction is obtained because we usually need only a few VFs/PCs to get a good description of the entire data set variability with a minimum of loss of information. Mathematically speaking, it is a linear combination of the variables that are most correlated with it. This further implies that the factor coordinates (or factor loadings) of a variable are the correlations between the variable and the factor or principal axes. Accordingly, interpretation of the PCs must be done in terms of the correlation. With this fact and the objective of factor interpretation in mind, given a set of variables, we should naturally be looking for those variables that have the highest (absolute) values of the factor coordinates for the given factors. Some other statistics that are useful for the purpose of interpretation are relative contribution of the factor axis to the eccentricity of the variables and the relative contribution of a variable to the variance of the factor axis. Besides, in FA, VFs usually group the studied variables in accordance with common features (i.e. soluble salts, organic pollutants, etc.). So FA and PCA are particularly valuable when a chemical, physical or biological interpretation of the data grouped in VFs/PCs is possible [11, 34, 40, 46–48].

Principal components analysis (PCA) has become one of the most popular methods used in multivariate statistics. Many reports on the use of PCA for the evaluation of changes in river water quality have been published (e.g. [34, 41, 43, 45, 46, 48, 49, 55–57]), including the combined use of physical, chemical, biological and biochemical parameters [58].

When PCA was applied to the same data set used to calculate WQI [16] and CA (Fig. 1), we obtained two complementary information on the water quality of the Suquía River. Figure 2 shows only spatial differences along the basin, pointing out high, medium and low basin areas as well as physical and chemical parameters associated with such basin areas. From Fig. 2 we can see that using only two principal components (PC1 and PC2), it is possible to account for 59% of the variance.

Different symbols and colours indicate different basins and seasons. Even though only 59% of the variance is explained, analysing PC1, samples from the high basin (wet and dry season) and samples from medium basin (wet season) are located to the left of the graph, in agreement with results obtained in CA (Fig. 1). On the right side of PC1, samples from the low basin (dry season) are located, while

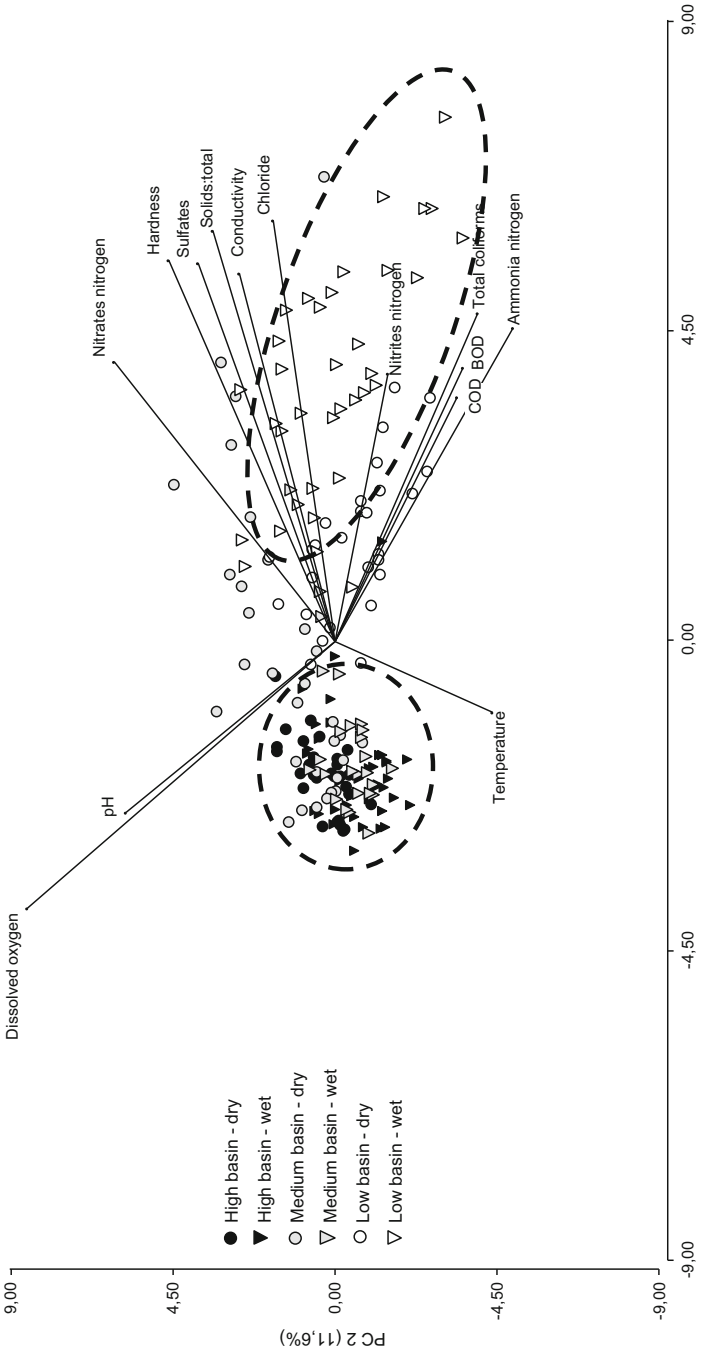

Fig. 2 Biplot produced from the entire data set used to calculate the water quality indice [16], obtained by PCA of this data matrix. Each point represents a sample (replicates are averaged), belonging to the high, medium or lower river basin

close to the zero point are samples belonging to middle basin (dry season) and to the low basin (wet season). It is worth to remark that samples located close to the zero point are not characterized by any of the analysed variables; instead, samples located to the left of PC1 are characterized by dissolved oxygen, pH and temperature. Conversely, samples located to the right of PC1 are characterized by ammonia nitrogen, conductivity, dissolved oxygen, chloride, COD, hardness, nitrates nitrogen, nitrites nitrogen, total solids, sulphates, total coliforms and BOD.

When data belonging to a particular area are averaged, the PCA biplot looks like Fig. 3. In Fig. 3 we can see even better differences between the low and the high basin, with the middle basin in the middle. When comparing Fig. 2 with Fig. 3, it is evident that real data points (Fig. 2) show some degree of overlaps, which is the natural situation as small changes are observed along the year, between years, etc. However, Fig. 3 better summarizes the trend observed along the basin. Similar interpretation as in Fig. 2, related to the variables and samples distribution along the axes, can be made.

Figure 3 shows that principal components 1 and 2 account for ca. 90.9% of the total variance. So far, PCA is able to differentiate well between higher, middle and lower basin, with most of the chemical parameters showing a strong change from higher to lower basin (from left to right along CP1 in this case). In addition to spatial differences, Fig. 3 also shows that the Suquía River basin also presents temporal variations. As it is evident from Fig. 3, variations along CP1 are much more pronounced than the corresponding to CP2, which means that temporal variations are less significant than spatial variations. However, further analysis evidences that the lower basin has bigger differences between wet (rainy) and dry season along CP1. So far, Fig. 3 resembles results obtained by CA (Fig. 1), showing that similar results are possible by two independent methods, using different mathematical approaches, diverse modes of graphical representation but same trends. This provides with additional certainty on the analysis of results, which is important to reinforce our conclusions but not so frequently performed by many researchers, policy makers, etc.

3.3 Supervised Methods: Discriminant Analysis (DA)/ Generalized Procrustes (GPA)

3.3.1 Discriminant Analysis (DA)

In contrast with CA, PCA and FA, which can be carried out without previous indication of the group to which a particular sample/data belongs, supervised methods require a grouping variable. In river monitoring this grouping variable is usually the monitoring station or the sampling period (summer/winter, etc.), although grouping variable can also refer to a group of monitoring stations (high basin, low basin, etc.) or to a group of months with some common characteristic (dry season/wet season, etc.).

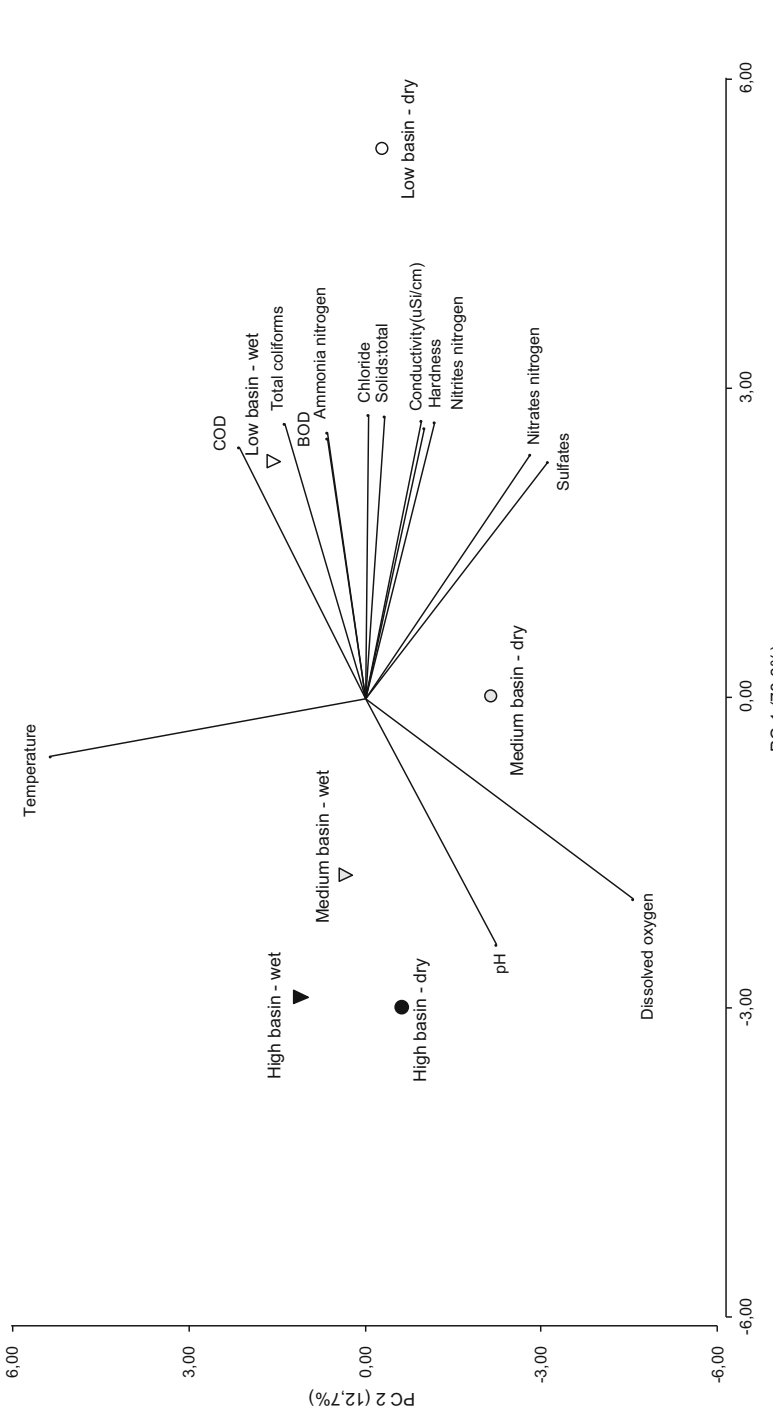

Fig. 3 Biplot produced from the entire data set used to calculate the water quality indice [14], obtained by PCA of this data matrix. Each point represents the average of samples belonging to the high, medium or lower river basin

Among others, DA technique is probably the most popular for evaluating changes along a river basin, within different seasons, etc. [34, 46, 48, 49]. DA builds up a discriminant function for each group, this function has the form presented in Eq. (1), and it is similar to a mathematical function having multiple terms (to the corresponding variables parameters analysed), with different load (weight) each for the final result [34, 39].

$$f(\mathrm{G}i) = k1 + \sum_{J=1}^{n} w_{ij}.p_{ij} \tag{1}$$

where

i is the number of groups (G).
ki is the constant inherent to each group.
n is the number of parameters used to classify a data set into a given group.
w_{ij} is the weight coefficient, assigned by DA, to a given parameter/variable (j), measured within a particular group (i).
(p_{ij}) is the analytical value of the parameter/variable (j), corresponding to a particular group (i).

During river monitoring it is necessary to measure parameters having different magnitudes in their value (e.g. pH, 1–14 and conductivity, 100–>1,000, etc.). Although these differences in magnitudes can be compensated assigning different values to w_{ij}, it is very frequent to standardize the entire data set (matrix) to mean 0 and variance 1 [34, 39].

The efficiency of these discriminant functions (DF) needs to be checked. Here two alternative procedures can be used:

(a) Construct the DF using one part of the data set; check its goodness with the remaining data (which implies a large number of samples to be able to divide the data set in two).
(b) Use the entire data set for constructing DF, but use the cross-validation method to verify it, which refers to the process of assessing the predictive accuracy of the model in a test sample (cross-validation), relative to its predictive accuracy in the learning sample from which the model was developed. If the model performs as well in the test sample as in the learning sample, it is said to cross-validate. One common method for cross-validation is the so named "one in, one out," which first constructs the discriminant function with the entire data set, removes one data line (set of variables corresponding to a particular site, monitored at a particular time), recalculates new DFs and checks if the removed data fits to its original group using the new constructed function. This procedure is repeated for each data line from the original matrix.

One additional feature that DA offers is the possibility to include all the variables measured to calculate DFs (standard mode), or constructing DFs by a stepwise procedure. Stepwise mode can be also divided into two alternatives, forward or backward stepwise. Forward stepwise mode starts including only one parameter

within DF (the most significant to differentiate among groups); in the next step, it adds the second most significant parameter and verifies significant changes (improvements) in the ability to discriminate groups. If so, a third step is performed and so on until adding a new variable does not significantly improve the previous discriminating power, which causes that the stepwise procedure is stopped and DFs are constructed using only those parameters significant to the discrimination. The second alternative is the backward stepwise mode, which starts with all the variables measured (as in the standard mode) and then removes the less significant variable in a second step and so on until removing a variable causes a significant drop in the predictive ability. So far, stepwise modes allow reducing the starting number of variables, keeping only those that are significant to discriminate among river sections, seasons, etc. This is an important issue of DA, because it allows reducing the number of parameters to be measured in monitoring campaigns but keeping the same valuable information on changes in the river water quality, identification of both point and non-point pollution sources, etc. [34].

DA was applied to the same data set used in [16] for calculating water quality indice and in this chapter to show how CA (Fig. 1) and PCA (Figs. 2 and 3) perform. Results from DA, corresponding to the Suquía River basin, are shown in Table 1 and Fig. 4.

From Table 1 it is evident that the classification power of DA did not change so much from the standard to the backward stepwise mode, which rendered the best compromise between right assignations (classification) and number of parameters to be used. In this case, using only six parameters would be enough to distinguish water from the high, medium or lower basin of the river and according to the season of sampling. Moreover, these six parameters are relatively simple to measure and could be used in a surveillance programme of the basin, at low cost.

Table 1 Parameters used by DA in different modes to evaluate spatial and temporal changes in the water quality of the Suquía River basin

Parameters used	Standard mode	Forward stepwise	Backward stepwise
Ammonia nitrogen	✓*		
BOD-5	✓*		
Chloride	✓	✓	✓
Chem.Oxig.demand	✓	✓	
Conductivity	✓*		
Dissolved oxygen	✓	✓	✓
Hardness	✓	✓	✓
Nitrates nitrogen	✓	✓	
Nitrites nitrogen	✓	✓	
pH	✓	✓	✓
Solids: total	✓*	✓*	
Sulphates	✓	✓	
Temperature	✓	✓	✓
Total coliforms	✓*		✓*
Classification goodness	87.2%	87.5%	85.2%

* $P > 0.05$

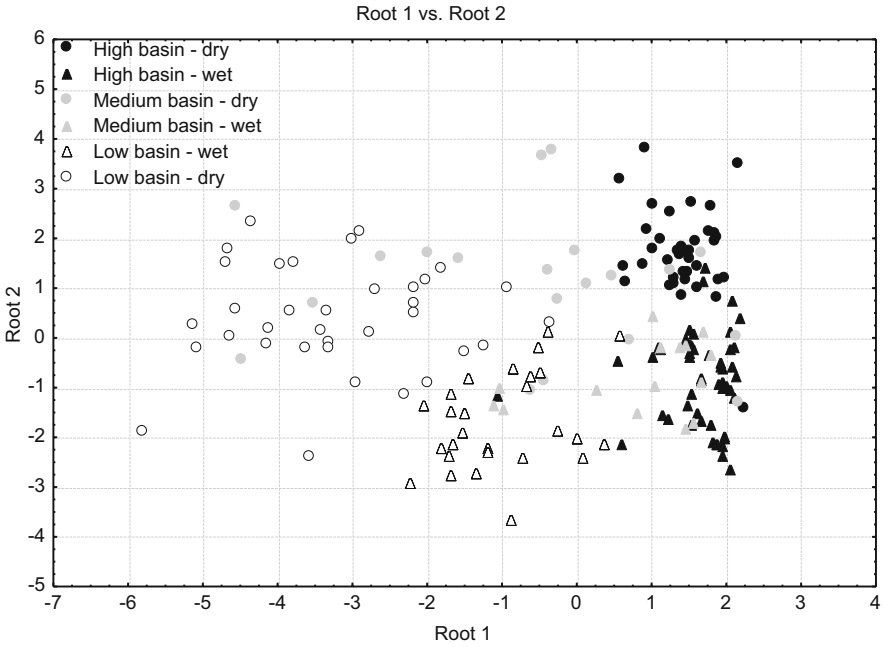

Fig. 4 Biplot obtained from DA (backward stepwise mode), corresponding to monitoring stations throughout the basin at both dry and wet season

From Fig. 4 it is clear that the best separation is obtained between the higher and the lower basin, which is expected considering the presence of the big city in the middle. Also from Fig. 4 we can see that the middle basin is overlapped with both high and lower basins. Additionally, Root 1 separates spatially well the river (from right = high basin to left = low basin). Conversely, Root 2 separates the basin according to the season (dry vs. wet). So far, DA affords similar results than those obtained with CA and PCA but now with a notorious reduction in the amount of data necessary to point out both spatial and temporal differences observed along the river basin throughout a year.

3.3.2 Generalized Procrustes Analysis (GPA)

Generalized procrustes analysis (GPA) is one of the family of methods that are concerned with the analysis of data arising from different group of variables/parameters, and it is frequently used to find a consensus from different set of variables analysed to a sample. In other words, the main goal is to acquire a consensus from the different groups of variables after they have undergone procrustes transformations that reduce individual differences by means of translation, rotation and reflection as well as isotropic scaling [59].

The statistical problem is to find a set of transformations (rotation, reflection, translation and an optional isotropic scaling factor), so that there is maximal agreement among the transformed configurations. The consensus is simply the average of all the transformed configurations. Here we used the Grower algorithm that minimizes within-samples variance by applying translation, scaling and rotation to generate a p-dimensional average configuration Yc. Following this, a q-dimensional group average space ($q \leq p$) is constructed from Yc by PCA. Therefore, GPA theory and algorithms can be applied to match chemical, physical and microbiological data arising from the different basin locations.

The aim of GPA is to evaluate the correspondence between different data from the same object, in this case the different river basin and stations. We evaluate the correspondence between microbiological parameters (Configuration 1; Fig. 5: includes BOD and total coliforms, accounting for changes in the microbiological quality of the river water), chloride (Configuration 2; Fig. 5: representing changes in the water quality probably due to urban run-off [34]) and chemical parameters (Configuration 3; Fig. 5: includes several variables: conductivity, temperature, pH, dissolved oxygen, ammonia nitrogen, COD, hardness, nitrates and nitrites nitrogen, total solids and sulphates). This last group generally accounts for chemical changes in the water quality, sometimes related to urban activities but in many cases related with sewage pollution. In this analysis, two components explained almost 97% of the total variance, contributed by three groups of variables: CP1 explains 94.2%, while the CP2 explained 2.7% (see Fig. 5). As shown in Fig. 4, similar results to PCA analysis are obtained. Samples from the high basin during both seasons are differentiated from the rest. On the other hand, the low basin (dry season) is on the opposite side of the biplot, while the rest is around the zero point. Table 2 shows the consensus values for each basin and season among the different groups of variables under study. The average consensus was 88.7%, which means that the three groups of variables (configurations 1–3) describe samples in similar ways. The low basin showed the highest level of consensus (97.1%) during the dry season, followed by the high basin in both seasons, while the lowest values were for the middle basin. This is in accordance with the results obtained along all the statistical analysis, which always showed that samples from the middle basin are overlapped, sometimes with the higher basin and sometimes with the lower one. This overlapping of the middle basin is expected, as previously stated, because of seasonal differences in the river flow, causing more or less dilution of different analytes, more or less turbidity; also changes in the water temperature from winter to summer cause changes in dissolved oxygen, etc. So far, the middle basin is by far the most difficult area to characterize in the Suquía River basin. We were not able to find other reports on the use of GPA in water quality assessment, being this probably the first.

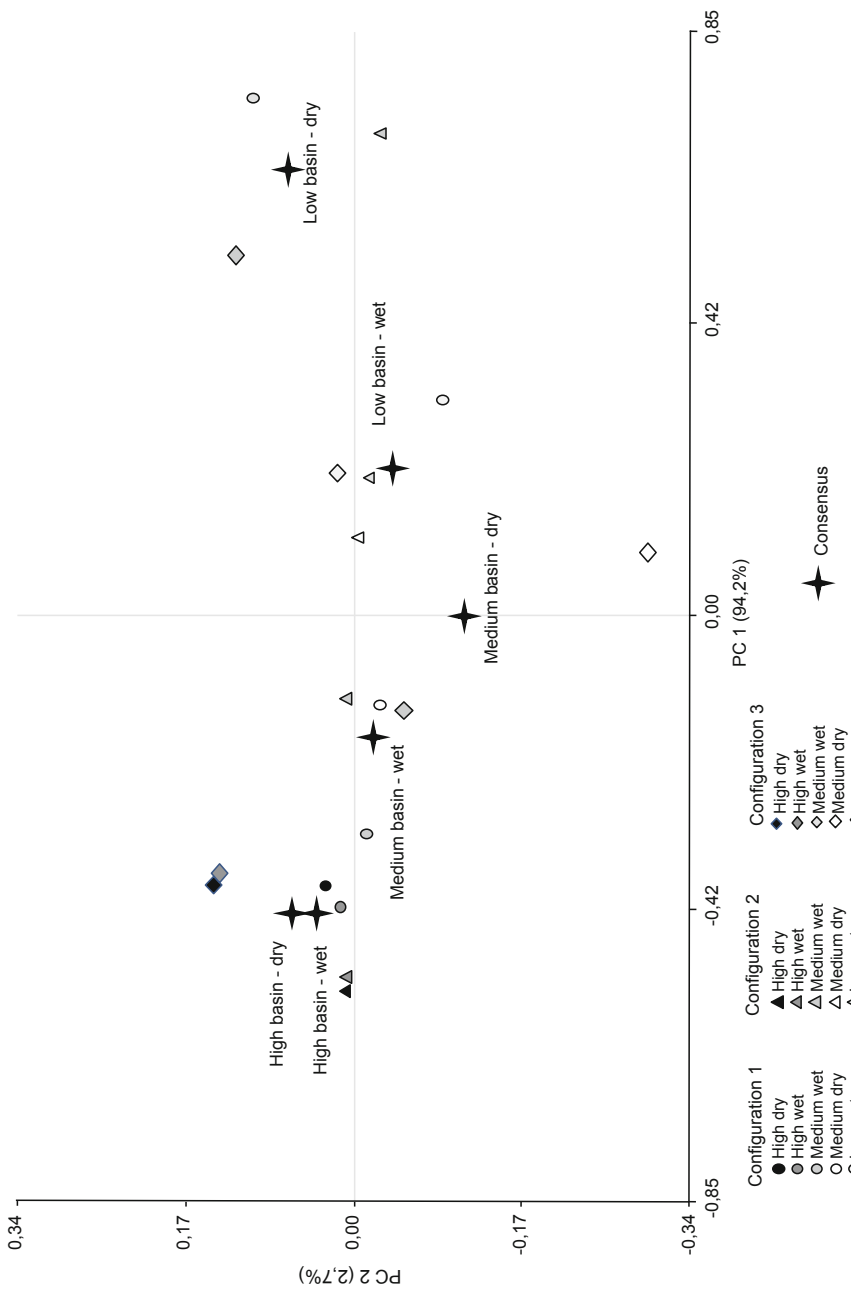

Fig. 5 Generalized procrustes analysis of the water quality in the Suquía River basin

Table 2 GPA: analysis of variance: sum of squares within each case

	Consensus	Residual	Total	Prop cons
High basin – dry	0.499	0.039	0.539	0.927
High basin – wet	0.486	0.038	0.524	0.928
Medium basin – dry	0.06	0.13	0.19	0.314
Medium basin – wet	0.082	0.046	0.129	0.64
Low basin – wet	0.228	0.041	0.269	0.849
Low basin – dry	1.311	0.039	1.35	0.971
Total	2.667	0.333	3	0.889

4 Conclusions

Monitoring and analysing a river for different purposes are difficult tasks. A rational analysis of the river hydrology, climate, urbanization, industries, etc. is required before starting a surveillance programme. After deciding the monitoring stations and a set of parameters to be monitored and measured, a big data set is constructed. This big data set can be simplified and transformed to a water quality indice (WQI), reflecting the change in the water quality along the basin or between seasons. Although WQI are easy to explain, there is a deficit of detailed information in a WQI. Thus, multivariate statistical methods (chemometrics, pattern recognition methods, etc.) help to fully evaluate changes in the water quality at many levels, in our case spatial and temporal changes. Among multivariate methods, CA can be used as a primary, unsupervised method to verify bulk differences in the water quality, for instance, between the higher and the lower basin in the Suquía River. However, CA does not indicate which parameters are causing such differences. PCA/FA can also be used to evaluate differences between areas (spatial changes) and seasons in an unsupervised way (without indicating the real origin of each sample). PCA and FA enable differentiating spatial areas and/or seasons or both, pointing out a set of parameters (variables) associated with each area/season. This is usually visualized through a biplot (Figs. 2 and 3), showing the dispersion of data points and their association with a group of variables. Finally, supervised methods (DA and GPA) enable a better differentiation between areas and seasons (or both), using the entire data set but also leading to an important data reduction, with less deficiency of information, in many cases. So far, DA and GPA are, in our criteria, best-suited methods for the evaluation of changes in the water quality of a river basin. In the case of the Suquía River basin, DA enables differentiating both spatial and temporal changes, with 85% certainty, using only six parameters. In addition, GPA evidenced that less changes are observed in the high basin during both spatial and temporal analyses, which means that the Suquía River maintains a relative constant water quality along the year in the high basin. Conversely, GPA shows big temporal differences in the lower basin. So far, the water quality of the Suquía River is better during the rainy season (more close to the higher basin quality) than during the dry season, where the low river flow causes concentration of most of the pollutants, decrease in the dissolved oxygen, etc. Finally, also GPA shows that

the middle basin has an intermediate water quality, being more similar to the quality of the lower basin during the dry season and close to the quality observed in the high basin during the wet season.

References

1. Hopke PK (2015) Chemometrics applied to environmental systems. Chemom Intel Lab Syst 149:205–214
2. Nunes CA, Alvarenga V, de Souza Sant'Ana A, Sousa Santos J, Granato D (2015) The use of statistical software in food science and technology: advantages, limitations and misuses. Food Res Int 75:270–280
3. Zuur AA, Ieno EL, Elphick CS (2010) A protocol for data exploration to avoid common statistical problems. Methods Ecol Evol 2010(1):3–14
4. Quinn GP, Keough MJ (2002) Experimental design and data analysis for biologists. Cambridge University Press, New York, NY, 539 p
5. Daszykowski M, Walczak B (2006) Use and abuse of chemometrics in chromatography. Trends Anal Chem 25(11):1081–1096
6. Pretsch E, Wilkins CL (2006) Use and abuse of chemometrics. Trends Anal Chem 25(11):1045
7. Zuur AF, Ieno EN, Smith GM (2007) Analysing ecological data. Springer, New York, 672 p
8. Leps J, Smilauer P (2003) Multivariate analysis of ecological data using CANOCO. Cambridge University Press, New York, NY, 283 p
9. Bilotta GS, Milner AM, Boyd IL (2015) How to increase the potential policy impact of environmental science research. Environ Sci Eur 27:9
10. Pesce SF, Wunderlin DA (2000) Use of water quality indices to verify the impact of Córdoba City (Argentina) on Suquía River. Water Res 34:2915–2926
11. Chapman D (1992) In: Chapman D (ed) Water quality assessment. Chapman & Hall, London, p 585. (On behalf of UNESCO, WHO and UNEP)
12. WHO (World and Health Organization) (1987) GEMS/WATER operational guide. World Health Organization, Geneva
13. Díaz É et al (2016) Hydrology and hydraulics of the Suquía River Basin. Handb Environ Chem. https://doi.org/10.1007/698_2016_466
14. Bistoni MA et al (2016) Biota along the Suquía River Basin. Handb Environ Chem. https://doi.org/10.1007/698_2016_455
15. Rodriguez MI, Ruiz M (2016) Limnology of the San Roque Reservoir. Handb Environ Chem. https://doi.org/10.1007/698_2016_467
16. Amé MV, Pesce SF (2015) Spatial and temporal changes in water quality along the basin. Handb Environ Chem. https://doi.org/10.1007/698_2015_434
17. Monferrán MV (2015) Metals and metalloids in water and sediment of the Suquía River Basin: spatial and temporal changes. Handb Environ Chem. https://doi.org/10.1007/698_2015_444
18. Santiago AN et al (2015) Organic pollutants in the Suquía River Basin. Handb Environ Chem. https://doi.org/10.1007/698_2015_431
19. Aboutalebi M, Bozorg-Haddad O, Loáiciga HA (2017) Multiobjective design of water-quality monitoring networks in river-reservoir systems. J Environ Eng 143(1), art. nr. 04016070
20. McCarthy DT, Jovanovic D, Lintern A, Teakle I, Barnes M, Deletic A, Coleman R, Rooney G, Prosser T, Coutts S, Hipsey MR, Bruce LC, Henry R (2017) Source tracking using microbial community fingerprints: method comparison with hydrodynamic modelling. Water Res 109: 253–265
21. Stipaničev D, Dragun Z, Repec S, Rebok K, Jordanova M (2017) Broad spectrum screening of 463 organic contaminants in rivers in Macedonia. Ecotox Environ Safety 135:48–59

22. Tousova Z, Oswald P, Slobodnik J, Blaha L, Muz M, Hu M, Brack W, Krauss M, Di Paolo C, Tarcai Z, Seiler T-B, Hollert H, Koprivica S, Ahel M, Schollée JE, Hollender J, Suter MJ-F, Hidasi AO, Schirmer K, Sonavane M, Ait-Aissa S, Creusot N, Brion F, Froment J, Almeida AC, Thomas K, Tollefsen KE, Tufi S, Ouyang X, Leonards P, Lamoree M, Torrens VO, Kolkman A, Schriks M, Spirhanzlova P, Tindall A, Schulze T (2017) European demonstration program on the effect-based and chemical identification and monitoring of organic pollutants in European surface waters. Sci Total Environ 601–602:1849–1868
23. van der Oost R, Sileno G, Janse T, Nguyen MT, Besselink H, Simoni BA (2017) (Smart integrated monitoring) as a novel bioanalytical strategy for water quality assessment: part II-field feasibility survey. Environ Toxicol Chem 36:2400–2416
24. Arnold JG, Fohrer N (2005) SWAT2000: current capabilities and research opportunities in applied watershed modelling. Hydrol Process 19(3):563–572
25. Gassman PW, Reyes MR, Green CH, Arnold JG (2007) The soil and water assessment tool: historical development, applications, and future research directions. Transact ASABE 50(4): 1211–1250
26. Hawkins CP, Norris RH, Hogue JN, Feminella JW (2000) Development and evaluation of predictive models for measuring the biological integrity of streams. Ecol Appl 10(5): 1456–1477
27. Moriasi DN, Arnold JG, Van Liew MW, Bingner RL, Harmel RD, Veith TL (2007) Model evaluation guidelines for systematic quantification of accuracy in watershed simulations. Transact ASABE 50(3):885–900
28. Rauch W, Henze M, Koncsos L, Reichert P, Shanahan P, Somlyódy L, Vanrolleghem P (1998) River water quality modelling. I. State of the art. Water Sci Technol 38:237–244
29. Romagnoli M, Portapila M, Rigalli A, Maydana G, Burgués M, García CM (2017) Assessment of the SWAT model to simulate a watershed with limited available data in the Pampas region. Argentina Sci Total Environ 596–597:437–450
30. Shanahan P, Henze M, Koncsos L, Rauch W, Reichert P, Somlyódy L, Vanrolleghem P (1998) River water quality modelling II. Problems of the art. Water Sci Technol 38:245–252
31. Somlyódy L, Henze M, Koncsos L, Rauch W, Reichert P, Shanaham P, Vanrolleghem P (1998) River water quality modelling III. Future of the art. Water Sci Technol 38:253–260
32. Zhai X, Xia J, Zhang Y (2017) Integrated approach of hydrological and water quality dynamic simulation for anthropogenic disturbance assessment in the Huai River Basin. China Sci Total Environ 598:749–764
33. Hued AC, Bistoni MA (2005) Development and validation of a biotic index for evaluation of environmental quality in the central region of Argentina. Hydrobiologia 543:279–298
34. Wunderlin DA, Díaz MP, Amé MV, Pesce SF, Hued AC, Bistoni MA (2001) Pattern recognition techniques for the evaluation of spatial and temporal variations in water quality. A case study: Suquía River basin (Córdoba-Argentina). Water Res 35(12):2881–2894
35. Otto M (1998) Multivariate methods. In: Kellner R, Mermet JM, Otto M, Widmer HM (eds) Analytical chemistry, Wiley-VCH, Weinheim, Germany, 916 p
36. Adams MJ (1998) The principles of multivariate data analysis. In: Ashurst PR, Dennis MJ (eds) Analytical methods of food authentication. Blackie Academic & Professional, London, UK, 350 p
37. Graça MAS, Coimbra CN (1998) The elaboration of indices to assess biological water quality. A case study. Water Res 32:380–392
38. Helena B, Pardo R, Vega M, Barrado E, Fernandez JM, Fernandez L (2000) Temporal evolution of groundwater composition in an alluvial aquifer (Pisuerga River, Spain) by principal component analysis. Water Res 34:807–816
39. Johnson RA, Wichern DW (1992) Applied multivariate statistical analysis. 3rd edn. Prentice–Hall International, Englewood Cliffs, NJ, USA, 642 p
40. Vega M, Pardo R, Barrado E, Debn L (1998) Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. Water Res 32:3581–3592

41. Astel A, Tsakovski S, Barbieri P, Simeonov V (2007) Comparison of self-organizing maps classification approach with cluster and principal components analysis for large environmental data sets. Water Res 41(19):4566–4578
42. De Las Heras Jiménez A, Melgar Riol MJ (2005) Rivers water quality at Galicia-Cost River Basin during 1973-2002. Predictive analysis to 2010 [Calidad del agua de los ríos de la Cuenca Hidrográfica Galicia-Costa durante el período 1973-1995. Análisis predictivo a 2010]. Tecnologia del Agua 25(261):54–61
43. Kannel PR, Lee S, Kanel SR, Khan SP (2007) Chemometric application in classification and assessment of monitoring locations of an urban river system. Anal Chim Acta 582(2):390–399
44. Khalil B, Ouarda TBMJ, St-Hilaire A (2011) A statistical approach for the assessment and redesign of the Nile Delta drainage system water-quality-monitoring locations. J Environ Monit 13(8):2190–2205
45. Kowalkowski T, Zbytniewski R, Szpejna J, Buszewski B (2006) Application of chemometrics in river water classification. Water Res 40(4):744–752
46. Shrestha S, Kazama F (2007) Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan. Environ Model Softw 22(4):464–475
47. Simeonov V, Stratis JA, Samara C, Zachariadis G, Voutsa D, Anthemidis A, Sofoniou M, Kouimtzis T (2003) Assessment of the surface water quality in Northern Greece. Water Res 37(17):4119–4124
48. Singh KP, Malik A, Mohan D, Sinha S (2004) Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India) – a case study. Water Res 38(18):3980–3992
49. Singh KP, Malik A, Sinha S (2005) Water quality assessment and apportionment of pollution sources of Gomti river (India) using multivariate statistical techniques – a case study. Anal Chim Acta 538(1–2):355–374
50. Sundaray SK, Nayak BB, Lin S, Bhatta D (2011) Geochemical speciation and risk assessment of heavy metals in the river estuarine sediments-a case study: Mahanadi basin, India. J Hazard Mater 186(2–3):1837–1846
51. Varol M (2011) Assessment of heavy metal contamination in sediments of the Tigris River (Turkey) using pollution indices and multivariate statistical techniques. J Hazard Mater 195: 355–364
52. Devic G, Djordjevic D, Sakan S (2014) Natural and anthropogenic factors affecting the groundwater quality in Serbia. Sci Total Environ 468-469:933–942
53. Li G, Liu G, Zhou C, Chou C-L, Zheng L, Wang J (2012) Spatial distribution and multiple sources of heavy metals in the water of Chaohu Lake, Anhui, China. Environ Monit Assess 184(5):2763–2773
54. Cao Y, Bark AW, Williams WP (1997) A comparison of clustering methods for river benthic community analysis. Hydrobiologia 347(1–3):25–40
55. Chabukdhara M, Nema AK (2012) Assessment of heavy metal contamination in Hindon River sediments: a chemometric and geochemical approach. Chemosphere 87(8):945–953
56. Clarke RT, Wright JF, Furse MT (2003) RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. Ecol Model 160(3):219–233
57. Marchant R, Hirst A, Norris RH, Butcher R, Metzeling L, Tiller D (1997) Classification and prediction of macroinvertebrate assemblages from running waters in Victoria, Australia. J N Am Benthol Soc 16(3):664–681
58. Monferrán MV, Galanti LN, Bonansea RI, Amé MV, Wunderlin DA (2011) Integrated survey of water pollution in the Suquía River basin (Córdoba, Argentina). J Environ Monit 13: 398–409
59. Tomic O, Berget I, Næs T (2015) A comparison of generalised procrustes analysis and multiple factor analysis for projective mapping data. Food Qual Prefer 43(2015):34–46