Journal Pre-proofs

Chemometric modeling for spatiotemporal characterization and self-depuration monitoring of surface water assessing the pollution sources impact of northern Argentina rivers

Marcelo A. Jurado Zavaleta, Mirta R. Alcaraz, Lidia G. Peñaloza, Analía Boemo, Ana Cardozo, Gerardo Tarcaya, Silvana M. Azcarate, Héctor C. Goicoechea

PII:	S0026-265X(20)33783-8
DOI:	https://doi.org/10.1016/j.microc.2020.105841
Reference:	MICROC 105841
To appear in:	*Microchemical Journal*
Received Date:	23 September 2020
Revised Date:	2 December 2020
Accepted Date:	8 December 2020

Please cite this article as: M.A. Jurado Zavaleta, M.R. Alcaraz, L.G. Peñaloza, A. Boemo, A. Cardozo, G. Tarcaya, S.M. Azcarate, H.C. Goicoechea, Chemometric modeling for spatiotemporal characterization and self-depuration monitoring of surface water assessing the pollution sources impact of northern Argentina rivers, *Microchemical Journal* (2020), doi: https://doi.org/10.1016/j.microc.2020.105841

Chemometric modeling for spatiotemporal characterization and self-depuration monitoring of surface water assessing the pollution sources impact of northern Argentina rivers

Marcelo A. Jurado Zavaleta[a], Mirta R. Alcaraz[b], Lidia G. Peñaloza[a], Analía Boemo[a], Ana Cardozo[c], Gerardo Tarcaya[c], Silvana M. Azcarate[d,*], Héctor C. Goicoechea[b]

[a] Consejo de Investigación (CIUNSa), Facultad de Ciencias Exactas, Universidad Nacional de Salta, Avenida Bolivia 5150, (4400) Salta, Argentina

[b] Laboratorio de Desarrollo Analítico y Quimiometría (LADAQ), Facultad de Bioquímica y Ciencias Biológicas, Universidad Nacional del Litoral-CONICET, Ciudad Universitaria, Santa Fe (S3000ZAA), Argentina

[c] Laboratorio de Efluentes, Planta Depuradora Sur, Aguas del Norte COSAySA, (4400) Salta, Argentina

[d] Instituto de las Ciencias de la Tierra y Ambientales de La Pampa (INCITAP-CONICET), Facultad de Ciencias Exactas y Naturales, Universidad Nacional de La Pampa, La Pampa, Argentina.

ABSTRACT

In Argentina, both surface and ground water are used for a diverse priority purposes, such as drinking and basic hygiene, but they are also utilized as receivers of different types of industrial and urban and suburban effluents that affect their natural composition. This activity accompanied by the increase of the population and climate changes have activated the alarms of organism water management forced to implement strict quality controls previous to its use. In this work, a systematic evaluation of a set of physicochemical and biological parameters measured in 19 sampling sites during the period 2017-2019 is presented. Principal component analysis (PCA) and matrix augmentation-PCA (MA-PCA) were applied as exploratory analysis tools to visualize and interpret the information contained in the dataset. Both studies allowed to detect the relevant variables and to differentiate the samples based on pollution areas. These models led to similar conclusions; nonetheless, MA-PCA provided a more straightforward overview of the spatiotemporal variation of the samples in comparison to classical PCA. Finally, a significant and sensitive discriminant model (93% non-error rate) was developed to analyze and predict the self-depuration of the rivers. The excellent predictive ability achieved by this model makes its application suitable for the monitoring of the water quality.

Keywords: *Argentina rivers; surface water quality; chemometric modeling; self-depuration monitoring; source pollution*

1. Introduction

Currently, by cause of the urbanization and industrialization activities, specific surface water systems, particularly rivers, are under constant threat by the action of multiple source pollution having a detrimental effect on the aquatic biodiversity, and compromise the water safety and river usages [1-4]. The U.S. Environmental Protection Agency (EPA) defines a non-point source pollution as "*a diffuse source that is difficult to measure and is highly variable due to different rain patterns and other climatic conditions*" [5], whereas point source pollution is an identifiable source from which pollutants are directly discharged into the environment. Among others, runoff from urban and suburban areas, farming, manufacturing, agricultural activities, and mining are the most common non-point source pollution, which is considered as the major cause of water quality degradation [4,6-11]. On the other hand, some factories, as paper or sugar mills, are common types of point sources.

The introduction of point and non-point source pollutants into waterbodies are of high social concern since natural water is probably the most appreciated and valuable natural resources in the world. [12]. Therefore, it is of outstanding importance recognizing the pollutant sources distribution and their spatiotemporal occurrence in order to find pollution patterns that aid to stablish accurate procedures for water quality control and environmental monitoring. To diminish the health risk from water pollution, many countries perform regular controls of the water quality of their most important water systems [4].

Argentina is a country with a large number of waterbodies, including rivers, lakes and ponds, which are the main sources for water supply, whereas, some communities rely on ground water as water supply.

Both surface and ground water are used for a diverse priority purposes, including drinking, basic hygiene, in addition to industrial, agricultural, and recreational uses. Despite all these benefits, they are also utilized as receivers for different types of industrial and urban and suburban effluents that affect their natural composition. This activity, accompanied by the increase of the population and climate, has trigged the alarms in water management departments. As a result, strict quality controls have been implemented.

To perform an extensive and accurate environmental evaluation, the sampling is accomplished in massive scale and multiple physical, chemical and biological parameters are evaluated. This procedure generates large-size data of high complexity [13-14], which usually preclude the right implementation of data analysis and, in consequence, its interpretation. To overcome this problem, chemometric methods have arisen as power tools allowing extracting information from diverse data arrays and exploring the underlying patterns that, otherwise, could become an outstanding challenge.

Chemometric methods would help to find relationships between groups of samples and/or variables and, eventually, to identify the pollution source that impact on the area under study. In this regard, principal component analysis (PCA) is one of the most established techniques utilized in environmental studies since it enables to reduce the dataset dimensionality, and, then, to provide an easy visualization of the relationships between variables and samples. Furthermore, important factors explaining the data variability have been statistically encountered that helped to identify sources of spatial variability in water quality and to interpret complex environmental monitoring data [4,**Error! Bookmark not defined.**14-25]. In general, classical PCA model has been applied for the interpretation of datasets arranged in two-way arrays.

A variant of PCA, called matrix augmentation-PCA (MA-PCA), has emerged as an interesting approach providing a comprehensive interpretation of numerous parameters that can affect the study, in particular, environmental studies. MA-PCA allows handling complex data arrays in an easy way concatenating the multiple two-way data arrays one on top of the other to provide a new augmented two-way data matrix [26,27]. Due to the information in two modes becomes mixed and the results could be difficult to interpret, the confounded information is recovered by rearranging each augmented score vector into a matrix and, then, averaging them in both directions [26,27]. As an additional advantage, it can be mentioned that MA-PCA can provide some insight into environmental studies which could not be undertaken using classical N-way methods, since the incomplete environmental data bases prevent their arrangement in n-arrays [28,29]. MA-PCA has been successfully applied to model and understand the spatiotemporal variations of polluting substances in the environment such as water samples from lakes and rivers [4,14,20,30]. Particularly, it provided the identification of the main sources of the pollutants in river waters from Portugal and the interpretation according to their chemical characteristics and their geographical and temporal profiles [27]. Similarly, the temporal evolution of water quality could be related to seasonal increments of the physicochemical parameters, defining the decomposition of the organic matter in a local study carried out in rivers of Spain [31].

The aim of this study was to assess the spatiotemporal variations of the water quality parameters of 6 rivers of Salta province, Argentina, (Arenales, Bermejo, Juramento, Mojotoro, Rosario and Horcones), which belong to 2 hydrographic basins (Bermejo and Juramento), in order to evaluate their self-depuration capacity. For this purpose, a systematic evaluation of a set of physicochemical and biological parameters, measured in 19 sampling sites during the period 2017-2019, was accomplished. PCA and MA-PCA were applied as exploratory analysis tools to visualize and interpret the information contained in the dataset. After determining the anthropogenic impact on the ecosystems, the key challenge was to obtain a complementary classification model from a PCA-discriminant analysis (PCA-DA) that would allow the self-depuration monitoring of the rivers. This work is attempting to provide a tool that can be used for evaluation of water quality in order to assess the ecosystem health and to provide early environmental warnings that might indicate adverse effects.

2. Materials and methods

2.1. General considerations

2.1.1. Study area description

The surface water resource in Salta province has an irregular spatial distribution. In addition to being strongly affected by a deficient and unfavorable temporal distribution, the rivers present a long and pronounced dry season, in contrast to summer periods [32].

Upper Juramento Basin comprises a vast extension of the Argentine province of Salta (Capital city, Cerrillos, Chicoana, La Viña, Rosario de Lerma, Guachipas, Metán, General Guemes, Rosario de la Frontera, Anta, La Poma, San Carlos, Molinos, Cafayate y Cachi) along with other areas belonging to the neighboring provinces of Tucumán and Catamarca. The Basin physiology presents a clearly distinct asymmetry. To the west, it is framed by elevations over 4.000 m above sea level with peaks up to 6.700 m above sea level whereas the Eastern side generally has heights below 2.000 m above sea level, reaching heights of 400 m above sea level in Chaco region. The predominance of arid to semi-arid climates in the Basin determines the relevance of water resource exploitation. This is due to drought periods that resent the quality and availability of surface water in the hydrological cycle [33].

The main rivers which are part of different sub-basins have a particular rainfall hydrological system, depending on rainfall seasonality. This occurs during the summer months of maximum rainfall from January to March, with flooding peaks in the month of February. Some of the tributaries have a mixed rain-snow system, such as the headwaters of

Arenales, Rosario and Guachipas rivers which, in turn, are fed by meltwater. The dry season runs from April to November. The minimum flows are recorded between September and November. When a large part of these river flows located upstream "Cabra Corral" Dam flow down into Lerma Valley, they must be injected. This is due to a slope break and a course granulometry that greatly favors water infiltration [33].

Bermejo River Basin extends across an area of approximately 123,000 km^2, developing its natural resources in the Argentine provinces of Salta and Jujuy, as well as in Tarija city in the neighboring country of Bolivia. It comprises an hydrologically active part known as the Upper Basin, having watercourses with mountain features. From the hydrological point of view, Bermejo Basin presents a prolonged period of recession and a limited high flow period during summer heavy rainfalls. On the other hand, the high production of sediments in its Basin is the distinctive feature of Bermejo river, which contributes with 100 million annual tons of sediments to Paraguay- Paraná Delta and Rio de la Plata system [34].

2.1.2. Sampling procedure

A monitoring program was designed to evaluate the impact of anthropogenic pollutants and to detect possible pollution patterns in surface water of 6 rivers of Salta province (north-west Argentina, 24°47′S, 65°25′W). They will henceforth be referred to as Z1, Z2, Z3, Z4, Z5 and Z6. All samples were collected in each contamination site, to be evaluated on the river already impacted with the effluent, and also, upstream and downstream from that specific sites by the drinking water and sewage services staff of Aguas del Norte COSAySA (Compañía Salteña de Agua y Saneamiento S.A.), Salta, Argentina, following the protocols developed by its Quality Department. For this study, two sampling at each hydrological cycle (high and low flow season) were carried out for 3 years (2017, 2018 and 2019). A total of 19 sampling sites (S1-S19) distributed among Z1-Z6 were chosen for being representative of the rivers under study and the whole sampling was completed in 11 campaigns (C#, see table S1, supplementary information). In this way, the final set of samples comprised a total of 190 samples. Figure 1 and table 1 summarize the locations of the sampling zones (Z#) and the selected sampling site (S#).

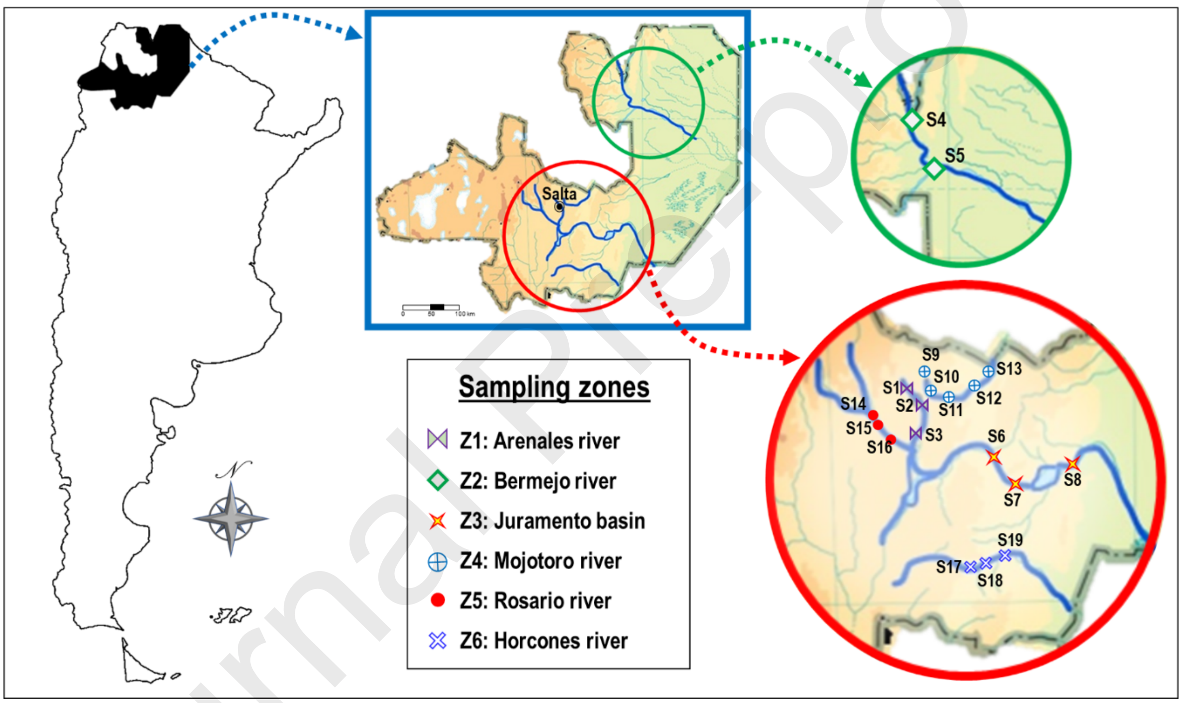

Figure 1. Map of the studied sampling zones (Z1: Arenales, Z2: Bermejo, Z3: Juramento, Z4: Mojotoro, Z5: Rosario and Z6: Horcones rivers) and the selected sampling sites (S1-S19)

*** Insert Table 1 ***

All the sampling zones were evaluated at specific locations considering the areas of pollutant discharge (DP) and the areas upstream and downstream (DU) of the discharge point.

Upstream areas are considered as reference states of water composition, and downstream, as index of self-depuration of the river.

Three locations were sampled at Arenales river (Z1), one of them is located at the South Treatment Plant outlet of Salta city, which constantly receives urban waste, urban and industrial sewage effluents from this discharge site. The 2 remaining sites are located upstream (after the junction of the Arias and Arenales rivers) and downstream from the discharge site. Bermejo River (Z2) does not include a specific DP site, but 2 locations were sampled near to urban areas (Aguas Blancas city, Embarcación city) in order to evaluate the impact of point and non-point source pollutions. Juramento basin (Z3) receives raw sewage effluents discharge from the El Galpón city and the effluents of agricultural activities from the surrounded rural areas. Mojotoro river (Z4) was studied at 5 different sites: 2 DP and the corresponding upstream and downstream areas. These DPs receive direct contributions of urban effluents and sewage from Capital and Campo Santo cities and industrial effluents from the Industrial Park of Güemes, Salta, Argentina. Rosario river (Z5) was sampled along 3 locations; one of them receives contributions of sewage effluents from stabilization ponds in the city of Rosario de Lerma. Horcones river (Z6) receives sewage discharges from the city Rosario de la Frontera and contributions from agro-industrial activities in the area.

2.2. Sampling procedure and sample preparation

For metal analysis, samples were collected in 1L polyethylene containers, previously washed with analytical quality nitric acid, and rinsed with distilled water. After arrival to the laboratory, 50 mL was taken for boron (B) analysis and the remaining volume was filtered through fiberglass paper (Whatman 934-AH), which was finally preserved with nitric acid 1:1.

For physicochemical and microbiological analysis, samples were collected in 2L polyethylene containers, previously washed with sodium hypochlorite solution, rinsed with water, then, 1: 1 HCl solution and finally rinsed with distilled water. All the samples were storaged in the dark at 4ºC until the analyses were performed.

2.3. Analytical procedures for water quality parameter determination

For each sample, 27 water quality parameters were measured: 1)- water temperature (WT); 2)- pH; 3)- conductivity (C); 4)-settleable solids 10 min (SS10); 5)- settleable solids 2 h (SS2); 6)-oxygen dissolved (OD); 7)- sulfide (S); 8)- total nitrogen (TKN); 9)- ammonia nitrogen (NH_4); 10)- organic nitrogen (Norg); 11)- biological oxygen demand (BOD); 12)- chemical oxygen demand (COD); 13)- phenols (Phen); 14)- total phosphorus (TP); 15)- fecal coliforms (FC); 16)- total coliforms (TC); 17)- boron (B); 18)- iron (Fe); 19)- manganese (Mn); 20)- chromium (Cr); 21)- zinc (Zn); 22)- cadmium (Cd); 23)- copper (Cu); 24)- lead (Pb); 25)- mercury (Hg); 26)- arsenic (As) and 27)- selenium (Se). All the analytical determinations were performed according to the Standard Methods for the Examination of Water and Wastewater [35]. Table 2 summarizes water quality parameters, analytical techniques, methods and instruments implemented for the assay. In all cases, calibration, recovery tests, blank measurement and correction procedures were accomplished. All the experiments were performed in duplicates.

*** Insert Table 2 ***

2.4. Data analysis

The resulting dataset for spatiotemporal assessment of water quality of Salta rivers consisted in 19 sampling sites with 27 measured parameters, monitored at every hydrological cycle for 3 years. For chemometric analysis, several strategies were implemented by using different data structures. Prior to chemometric modeling, various preprocessing methods were tested and autoscaling was selected. Although it led to models explaining low raw variance, it achieved simpler chemical interpretations with reasonable groups of samples and their loadings become easier to interpret.

2.4.1. Exploratory Data Analysis

2.4.1.1. Principal component analysis

To identify the correlations between the multiple parameters and to consistently evaluate the water quality, PCA was conducted. PCA is a useful tool that help elucidating the complex nature of multivariate relationships and comprehending the structure of multivariate complex datasets by revealing intrinsic hidden patterns [36]. In the present case, similarities and differences among samples were analyzed by visual inspection of the achieved principal components (PC) scores, and the relevance of the variables were evaluated through the loading plots.

On one hand, to ascertain the most appropriate data pre-treatment procedure and to find outliers and main patterns, PCA models were developed for a set of samples belonging to the same Z# set. Thus, 6 PCA models were built and a comparative evaluation was done (PCA_Z). On the other hand, in order to assess similarities among Z# pattern behaviors and to examine the temporal distribution of the pollution patterns, PCA models were developed for each C# (PCA_C). This analysis enabled to understand the correlations among the multiple studied parameters and to consistently obtain water quality patterns.

2.4.1.2. Principal component analysis, matrix augmentation MA-PCA

Even though the main application of classical PCA is for two-dimensional dataset, it can be easily extended to the simultaneous analysis of multiple datasets through matrix augmentation. This matrix augmentation consists of arranging a three-dimensional ***X*** object (S# × variables × C#) into a two-dimensional **X**aug array ((S# × C#) × variables). In the present case, an augmented matrix comprising 19 × 11 = 209 rows (S# × C#) and 23 columns (variables) was built and subjected to MA-PCA.

As a result, an augmented score matrix containing information about geographical and temporal distribution of river pollution patterns is acquired. Notwithstanding loadings provide useful insights about the relationships among variables, the information comprised into the two other dimensions or modes (spatial and temporal) is intertwined in the scores, which is hardly interpretable and may hinder the usefulness of MA-PCA. Therefore, to overcome this difficulty, a strategy based on refolding the scores is applied allowing direct access to information [26-31]. For that, each column of augmented score matrix is refolded to give a new score matrix, where the columns would correspond to C# and the rows to sites#. If they are row-wise averaged, the resulting vector will represent the time-averaged geographical distribution of the corresponding PC of the augmented score matrix. On the other hand, if column-wise averages are calculated, the obtained vector will indicate the temporal evolution of such PC.

2.4.2. Classification by discriminant analysis

Discriminant analysis (DA) operates a strict classification by associating each of the samples with one and only one of the possible classes. DA approaches operate by partitioning the variable hyperspace into as many regions as the number of categories, calculating decision surfaces minimizing some sort of error criterion for the training samples being the most common the overall classification error. [37]. Its implementation requires data compression; hence, it is possible to use PC scores previous obtained from PCA.

In the present work, PCA model was first applied as a data reduction tool to extract the score values of the individual components and, then, they were used for DA [36]. Prior to classification model, the original dataset was divided into two datasets: training and test set, with 75% and 25% of samples, respectively. The split between training and test sets was done by keeping the samples ratio of each class equal to the original dataset [38]. Finally, PCA-DA was applied on the training set to develop a model that permit to classify the classes previously observed.

2.4.2.1. Evaluation of the built model

The built classification model was internally validated by using venetian blind cross-validation (VBCV) and the final model performance was confirmed through test set validation (TV). The quality of the model was assessed by its prediction capability. The optimal conditions were chosen by using primary measures related to single classes, as sensitivity (S), specificity (SP) and precision (PR) of the calibration and prediction stages, which were calculated on each class encoding different classification aspects. Additionally, to provide an overall evaluation of the classification quality, the global indices derived from the primary class measures, such as average sensitivity (non-error rate -NER) and average precision (AC) were also calculated [39,40].

2.5. Software

Data preprocessing and PCA were performed by using in-house codes written in MATLAB 9.2 (R2017a) (The Mathworks, Natick, MA, USA). PCA-DA classification models were calculated with the Classification toolbox for MATLAB [41].

3. Results and discussion

3.1. Physicochemical and microbiological parameters in surface water - General considerations.

Table 3 summarizes the dataset generated throughout this study, showing the minimum and maximum values detected for water quality variables. Se, Phen and S were not considered in the study inasmuch as they were not detected in the analyzed samples. Additionally, despite WT was measured for all samples, it was not considered as a variable since it strongly depends on the season and it would not represent a pollution parameter (min annual WT: 9.5ºC - max annual WT: 33.5ºC). In this way, 23 parameters were finally considered for the quality water assessment. As can be noticed, wide variation ranges for some variables were observed, which may be associated to the sampling site or to the seasonal variations in climate (temperature, precipitation, etc.).

Most of the parameter values obtained for DP samples exceed the limits established for the quality standards for the discharges of liquid and/or industrial residual effluents of receiving bodies [42], who reports that the surface water samples that outstrip the critical values of SS10, SS2, COD, BOD, NH_4, Norg, TKN, C, TC and FC poses a health risk for human (Table 3). The higher values acquired for these variables can be a consequence of the organic matter decomposition that is discarded from DP. It is noteworthy that total COD and NH_4 are typical indicators of organic matter decomposition (leaves, grass, algae or some sort of wastes). Moreover, these variables can also be related to agricultural, household and industrial activities as well as urban and domestic waste. In this context, the increase of N and P pollution density relates primarily to the excessive use of fertilizers and agrochemicals in rural areas, together with livestock and poultry farming wastes.

Metals and metalloids are released into the environment through natural processes and human activities. The weathering of parent rocks and soil particles are natural sources of metals, while urban runoff, municipal sewage discharges, agricultural and industrial activities represent anthropogenic sources [43]. Certain metals, such as Zn, Pb and Cu, are typical anthropogenic pollutants. Cu is mainly used in wiring, electronics circuits and plumbing and other uses like healthcare (bactericides) and pesticide manufacturing (fungicides and algaecides). Pb is utilized for manufacturing batteries, ammunition and ceramics and as a paint pigment. Zn is widely used in the steel industry for Zn-Fe protective coatings. These elements are introduced into water bodies by urban runoff, sewage disposal and industrial dumping; all pathways are possible in the sampled areas of Salta province since rainwater is usually discharged either direct to surface water or introduced into sewage treatment plants to achieve dilution, even though plant capacity is sometimes exceeded in the rainy season. Presence of As and B in surface water of the Andean region of Salta arises from natural sources, such as mining, thermal springs or volcanic ashes. In the study area, As may be found in groundwater because of the sedimentary profile of soils, but the presence of both elements in surface water is mainly due to B processing in industrial plants located near the riverbanks. Moreover, leaching and runoff from tailing dumps in mining areas are also sources of metal pollution; some tributaries of Bermejo river collect mining disposals originated in Bolivia. Thus, due to the many possible pollution sources, several metals and metalloids such Mn, Cr, Cu, Pb, Hg and As were found exceeding the respective maximum tolerable limit.

It is worth noticing that the relationship between samples and variables for different S# and C# is complex and difficult to interpret. For this reason, multivariate approaches were conducted and they are described in the following sections.

*** Insert Table 3***

3.2. PCA analysis

3.2.1. PCA_Z: analysis for sampling zone

Due to high data variability, an independent PCA was conducted to evaluate individual Z# aiming to find the main points of pollution and to evaluate their own trophic state. The individual matrices corresponding to each Z# were subjected to PCA analysis. For this, an individual matrix (S# x variable) of dimension (32 × 23), (21 × 23), (29 × 23), (46 × 23), (29 × 23) and (33 × 23) for each Z# was build. As it can be noticed, Z# comprises a different number of S#, as local climatic conditions and river drought prevent the continuous sampling procedure on these respective site (for more details, see table S1, supplementary information).

As a result, the first 2 PCs were selected to represent the data variability. Figure S1 (Supplementary information) shows the scores and loading plots of each dataset defined by PC1 against PC2. The percentage of the explained variance for the individual components is shown on each axis. It is clear to observe that all the Z# behave similarly. The score plots evidence 2 groups that correspond to the DP samples and the DU samples. Thus, for all cases, PC1 clearly describes the separation of DU samples, on the negative side, from DP samples on the positive side. Furthermore, it could be observed that throughout the PC2, samples are scattered within each group.

Despite the different geographical localization of the S#; DU samples were located on the negative quadrant of PC1 as a unique group indicating that there were not significant differences in water quality between upstream and downstream samples.

One of the outcomes arisen from an in-depth evaluation of the loadings on PC1 is the relevance of the parameter DO for all the evaluated Z#, for which the higher levels were encountered in the Z# that does not have DP samples. SS10 and SS2 are common variables that exhibit, at low values, a significant correlation with the admissible water conditions [4]. For Z1 and Z3, pH was also an additional important parameter to define water quality. The variables responsible for this discrimination are mainly microbiological (such as TC and FC), organic matter indicators (COD, BOD, Norg, NH_4, TKN) and mineral indicators (B, Fe, Mn, Zn, TP and C). It is worth to highlight, that all these Z#, in particular, Z1, Z3 and Z5, receive large amounts of organic pollutants from identified sources, such as urban waste, domestic sewage and hatcheries and poultry farms effluents. The obtained results are in accordance with the previous observation and they reflect the impacts of these sources on the water quality, e.g., changes on the pH value, low DO concentration, high Norg, all these, as consequence of the fermentation processes of the organic matter.

All zones include DP samples and constitute well-discriminated groups with scores that are extensively shifted towards high values according to PC1. On the contrary, as expected for Z2, similar scores scattering were obtained from both downstream and upstream site samples. However, analyzing each zone from these individual PCAs, it was not possible to find some trends over time on the score plots when the samples were identified by their corresponding campaign (data not shown).

3.2.2. PCA$_C$: analysis for sampling campaign

In order to evaluate the behavior of the Z# at the same C#, PCA models were built considering individual sampling campaigns. To accomplish the analysis, 11 datasets with their corresponding S# taken from each C# (autumn, winter, summer and spring for the period 2017-2019) were built together with the 23 analyzed variables. Then, the matrices were subjected to PCA decomposition. According to the obtained results, the first three PCs were necessary to observe a clear differentiation among samples, with a >70% of the total data variability for all models. Figure 2 depicts the obtained results for the S# at C#, defined by the score on PC1, PC2 and PC3.

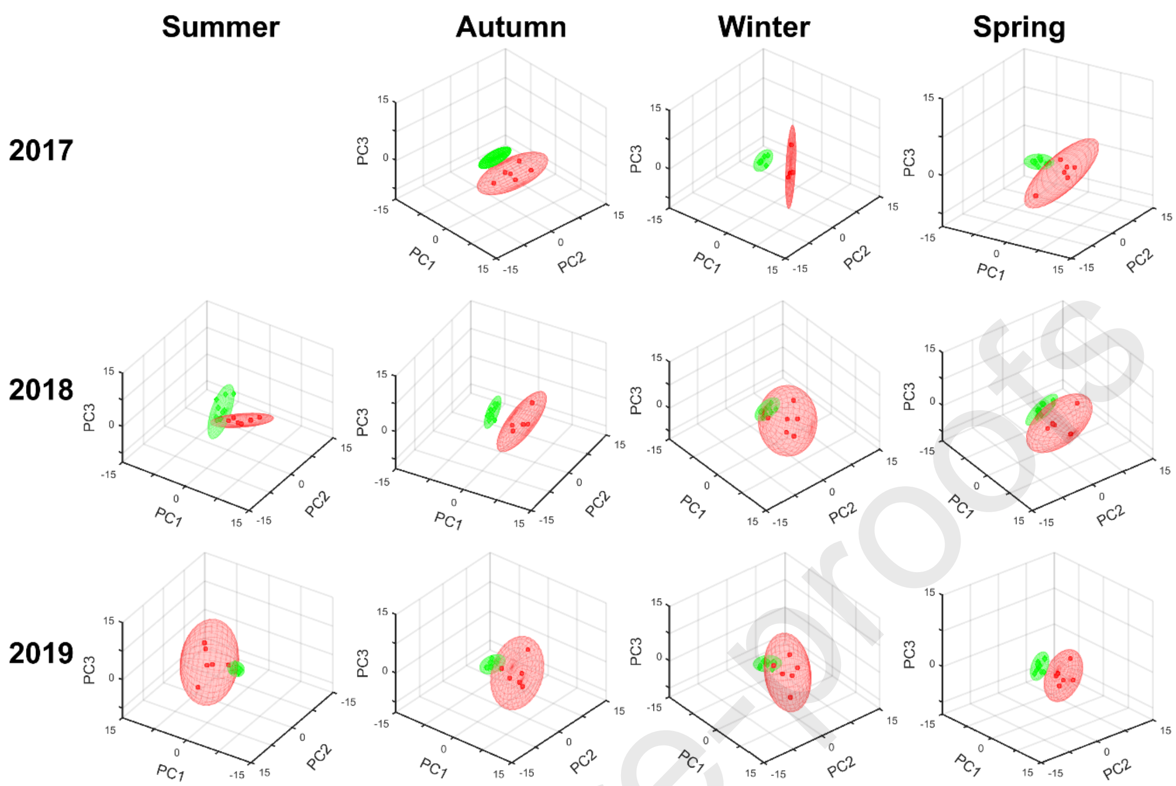

Figure 2. Score plots of the first 3 PCs obtained from PCA_C applied to the 11 datasets corresponding to each C#. The samples are shown according to the S# nature: DU samples (green) and DP samples (red). The three-dimensional projection of the confidence ellipsoids by applying the Student's t-distribution at 95% confidence level is included to facilitate visualization. Explained variance values of PC1, PC2 and PC3 in % are in table S2, supplementary information.

Strong similarities are clearly observed between upstream and downstream samples. Then the samples can be grouped into the same cluster. A distinguishable characteristic arose from the score plots, is the wide dispersion among the DP sample group for all the C#. On the other hand, in most cases, DU sample group showed a low dispersion indicating that the quality parameters remain stable regardless of their sampling point.

In addition, the obtained loadings behave similar to those acquired from PCA_Z. Figure S2 shows the loading plots of the first 3 PCs obtained from PCA_C applied to the 11 datasets corresponding to each C#. In general terms, several common variables in the different C# are responsible of samples discrimination. For all the C#, OD and pH seem to be the most relevant variable describing DU samples on PC1 and, particularly for C1, C3, C4, C5 and C9, some metals, such as Cr, Pb, Cr, Cu, Mn and Zn, display contributions on 3 PCs for this group of samples. According to the aforementioned for PCA_z, the main variables associated to DP cluster were TC, FC, COD, BOD, Norg, NH_4, TKN, TP. All of them showed slight variations between campaigns.

3.3. MA-PCA to evaluate spatiotemporal variability of the water quality

Spatiotemporal variability patterns of the full dataset were simultaneously studied by MA-PCA. For this purpose, an augmented matrix of dimension 190×23 was built.

Here, it is noteworthy that a number of factors as low as possible is required to facilitate the analysis understanding. Through a comprehensive comparison of only a few PCs, it was possible to explain the total variability of the original dataset. The first 3 PCs (47% variability) were considered to visualize the relationships between samples and variables.

Figure 3A and 3B display the score plot (PC2 vs. PC1) for S# samples and the corresponding loading plot for the first 2 PCs. As can be observed in Figure 3A, a clear division along the PC1 axis is obtained, with the DP samples on positive PC1 side and DU samples on the negative side; however, PC2 does not seem to contribute to a group differentiation. By a comparison of the groups, it is possible to conclude that, at 95% confidence level for PCA, some of the evaluated samples behave as outliers (see figure 3A). Nonetheless, they were included in the subsequent analysis. For instance, S7 (DP) displayed high PC1 values, which can be associated to the large discharge of wastes that are directly unloaded into the river, without any prior treatment.

Otherwise, S4, S14 and S16 (DU) exhibited higher values on the PC2 axes. These 3 samples were collected in the campaign C5, in autumn 2018, where two rivers (Z2 and Z5) presented a substantial decrease in their water volume, probably producing a concentration effect on the studied parameters.

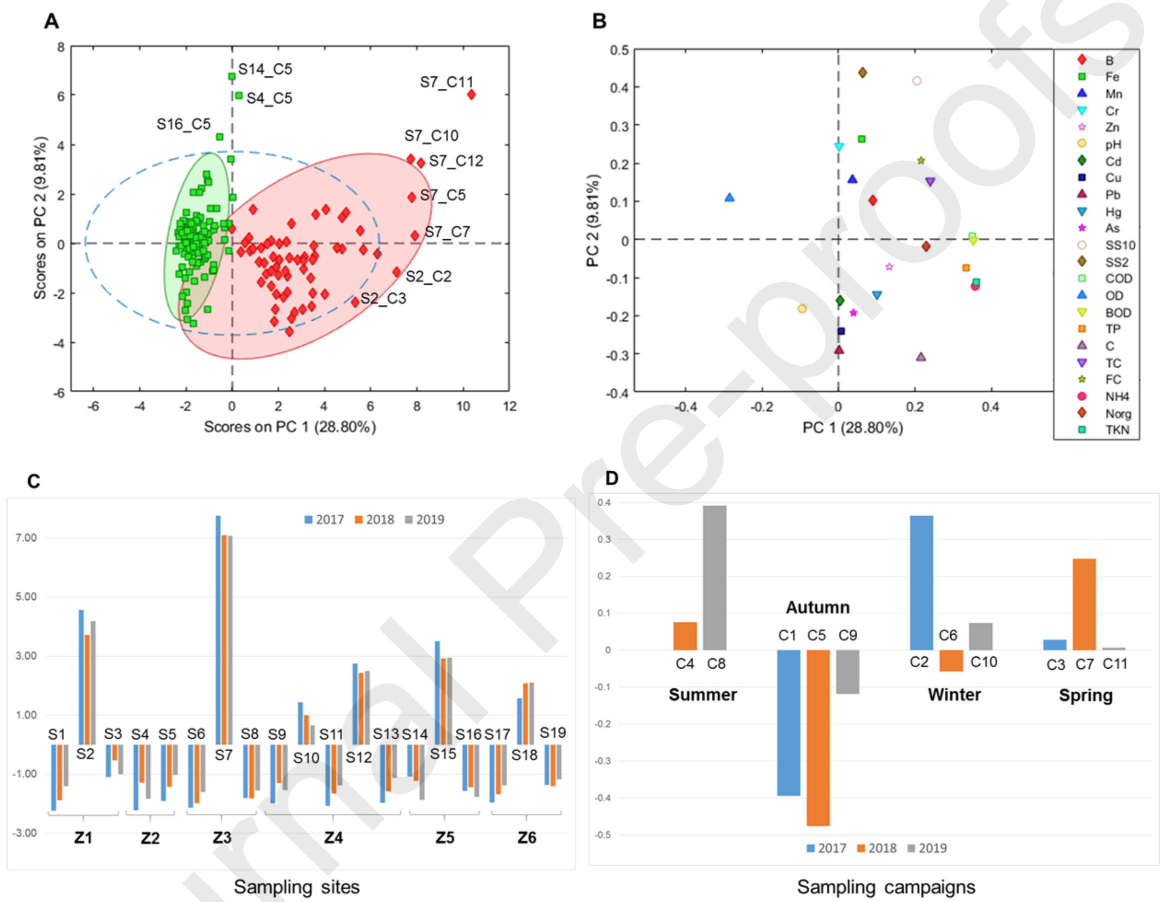

Figure 3. (**A**) Score plot acquired from MA-PCA. DP and DU samples are represented as red diamonds and green squares, respectively. The bi-dimensional projection of the confidence ellipse at 95% level for each class and for the global dataset (dashed light blue line) are included. (**B**) Loading plot obtained from MA-PCA. (**C**) Temporally averaged geographic scores after refolding of PC1 and (**D**) spatial averaged temporal scores after refolding of PC1 obtained from the simultaneous analysis of the 11 C# applying MA-PCA.

By inspection of the loading plot (Figure 3B), it is possible to observe that PC1 has highly positive values, being larger for NH_4, TKN, TP, BOD and COD, while SS10, FC, TC, Norg and C presented moderate positive contribution. On the contrary, pH and DO have relevant negative contributions to PC1. These results are in accordance with previous publications, reporting that PC1 can be associated to the simultaneous contribution of the DO and pH parameter [4,12,31]. This phenomenon reveals that both parameters tend to decrease with increasing anthropogenic water pollution. The rest of the parameters have negligible contributions of this component.

In order to assess the spatiotemporal variability of water quality, the scores estimated by MA-PCA were refolded according to the methodology proposed by Felipe-Sotelo *et al.* [31]. Figure 3C displays a bar plot built with the temporally averaged geographical MA-PC1 scores. This graph exposes a clear differentiation between the two groups of samples, DP and DU, with positive and negative values, respectively. Notwithstanding this conclusion arose from the MA-PCA plots (figure 3C and 3D), with this approach it is possible to make two inferences regarding the spatial variation of the water quality parameters. It can be seen that the source pollution impact is higher in 2017 period, while the parameters remain constant over period 2018-2019. Then, regarding DU samples, non-significant differences were observed between them, which led to the conclusion that rivers are able to return to the initial stage throughout natural processes. Figure 3D shows the representation of the spatially averaged temporal scores considering each season of the period 2017-2019. It can be noticed that the different C# showed different patterns, with strong dependence on the S#. The samples corresponding to autumn and winter 2017, autumn and spring 2018 and summer 2019, showed the stronger dependence on the S#.

Considering all the aforementioned observations, it can be concluded that the use of MA-PCA with refolded scores yields simple and straightforward representation that facilitates a quick and comprehensive understanding of spatial and temporal information.

An issue to highlight from the obtained results is the fact that downstream samples present similar characteristics to upstream samples. This allowed to assume that the rivers are capable to reach their initial natural quality. Along the river, waste and sewage discharges have direct detrimental impact on water quality. However, these results unravel the river ability to recover its water quality after passing through a pollution zone, i.e., water quality seems to be restored due to self-purification. This outcome is in accordance to a recent report that demonstrates that the water quality can be recovered in downstream sites of cities due to self-purification of surface waters [43]. However, it is not in agreement with the results reported by Daou et al. [19], who observed a significant downstream impact due to runoff arriving from some specific sources of pollution.

3.4. Self-depuration monitoring through a classification model

The results achieved by the different PCA models demonstrated the feasibility to build a model that permits classifying DP samples from the rest, and inferring about the self-depuration of the rivers under study. To corroborate this fact, an embedded method applying PCA-DA as classifier was performed. The motivation for the use of this kind of method arises from demand to ensure the most relevant variables in model.

First, the entire dataset (n = 190) was split into training (n = 128) and test (n = 62) subsets. Then, PCA-DA was applied as binary classification model on the training set by grouping the samples into 2 classes: DP and DU. Then, the optimum number of factors was determined by using VBCV. The optimal number of factors was chosen based on the lower error rate, being 3 the selected number for this analysis.

The scores on the first canonical variable of each S# are plotted in Figure 4, in which a clear discrimination between both groups is observed. In addition, this differentiation has an evidently linear behavior that was previously assessed with PCA models. Hence, a linear fitting function was implemented to build the model.

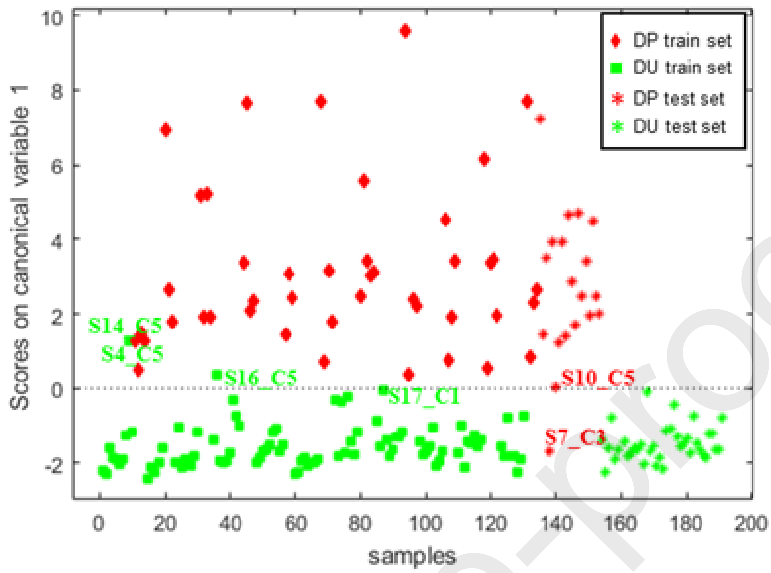

Figure 4. Projections of S# scores on first canonical variable for the PCA-DA model from linear decision boundary, showing the classification of the 2 evaluated sample class, DU and DP, for train and prediction subsets.

As result, this classifier only assigns each sample to a unique class. Under this condition, a well-known confusion matrix could be built from classification results, including information about actual and predicted classes disposed in rows and columns, respectively. The diagonal elements of the matrix contain the number of correctly classified samples, while off-diagonal elements include the number of misclassified samples. Table 4 summarize the confusion matrix built for VBCV and the statistical performance parameters of the classification model related to single classes. The classification results are expressed as the percentage of correct classification and the number of misclassified samples for each class. According to this binary classification task, several indices can be defined in terms of true/false positive/negative values to evaluate the model performance. The global parameters obtained for training stage were 0.93 and 0.95 for NER and AC, respectively.

*** Insert table 4 ***

Here, it is important to mention that the most useful indexes utilized to analyze samples and classes are sensitivity (S), which describes the ability of the model to correctly recognize samples belonging to a class, and specificity (SP), which characterizes the ability of a class to reject the samples of all the other. These indices have values ranging between 0 and 1 for non-class classification and perfect class classification, respectively. In the present study, the values obtained for S and SP to each class indicate meaningful model performance in this stage. However, it can be appreciated that 4 samples belonging to DP class and 2 samples of DU class were misclassified.

The predictive ability of the model was evaluated by analyzing classification indices from an independent test set. This prediction stage achieved NER and AC values of 0.96 and 0.96, respectively, which were better than those obtained from calibration stage.

In addition, only two samples of the DP group and four in the DU group were misclassified. Referring to DP samples (S10-C5 and S7-C3), despite these results can be undesired from the health and security standpoints, it is important to highlight that the parameters of the misclassified were below the maximum permissible. On the other hand, the fact that four DU samples (S17-C1, S4-C5, S14-C5 and S16-C5) were classified as sources of pollution indicates the lack of sewage treatment in many of the rural areas, cattle and human disposals spill directly into the water resource, being more perceptible in dry seasons (C1, C3 and C5).

Moreover, some implication on hydrological changes related to these results can be mention. All evaluated rivers have some common features: they run through low to medium slope terrain, carry great load of organic matter and silt and greatly increase their flow during summer. Downstream sample sites are located from 4 to 70 km away from the point source; only Juramento river (Z3) should not be considered since downstream sample is taken from a dam discharge, and so, after a dilution effect. High concentration of organic matter and silt help to remove heavy metals by suspension and precipitation; moreover, turbulent flows oxygenate water. On the other hand, eutrophication occurs in some extent, diminishing concentration of nutrients. However, self-recovery in a 4 km distance is quite remarkable.

Under this scenario, it can be concluded that the selected model was able to classify samples according to the proposed classes. Furthermore, classification results are in agreement with those reached from PCA models, i.e., misclassified samples were the same that behaved differently to the rest. In addition, it can be observed that OD, BOD, COD, TP, NH_4 and TKN were responsible of this discrimination. These variables were the same than those that presented higher variation in MA-PCA. The excellent predictive ability of the developed classification model makes it suitable for the water quality evaluation and for verification of self-recovery ability of the rivers by considering only a scarce number of parameters.

4. Conclusions

In this work, the application of chemometric techniques to model spatiotemporal water quality variations of Salta rivers, in the northwest area of Argentina, is presented. Although it can be considered as a case study, chemometric methodology can be used in similar studies where detection and characterization of point source pollution and self-recovery monitoring of the water resource are required. Twenty-seven physicochemical, chemical and biological parameters were quantified in 190 surface water samples collected during 11 sampling campaigns in the period 2017–2019. After a preliminary evaluation, 23 parameters were considered relevant to the study, from which, it can be concluded that samples from discharge areas can be considered as point source pollution according to the relevance of their load in organic matter, since that most of these quality parameters values were higher than those established as maximum tolerable limits.

PCA and MA-PCA were implemented as exploratory techniques for data recognition, and PCA-DA classification model was successfully built to predict the self-depuration capability of rivers.

Multivariate statistical techniques represent powerful and useful tool to understand the spatiotemporal variations of river water quality, as well as to identify main patterns arisen from the analyzed variables. It has proved that rivers are able to self-purify pollutants and return to an initial state of equilibrium in a distance that range from 4 to 70 km from the DP. This phenomenon sheds light on the fact that the physicochemical and biological environmental synergy aids the river to recover its water quality, ensuring the sustainability of future supplies. When using PCA-DA as classification model, not only was possible to point out pollution sources and establish self-recovery of resources, but also highlight events, such as satisfactory sewage treatment and diffuse organic pollution, represented by the misclassified samples founded.

This report is the first systematic study on Salta rivers and contains valuable information that can be established as a basis for future studies.

Acknowledgements

The authors are grateful to CONICET (Consejo Nacional de Investigaciones Científicas y Técnicas, Project PIP-2015 Nº 0111) and ANPCyT (Agencia Nacional de Promoción Científica y Tecnológica, Projects PICT 2017-0340) for financial support.

References

CRediT author statement

Marcelo A. Jurado Zavaleta: Software, Formal analysis, Visualization.

Mirta R. Alcaraz: Writing - Review & Editing.

Lidia G. Peñaloza, Ana Cardozo, Gerardo Tarcaya: Formal analysis

Analía Boemo: Supervision.

Silvana M. Azcarate: Conceptualization, Methodology, Investigation, Writing - Review & Editing.

Héctor C. Goicoechea: Conceptualization, Investigation, Resources, **Supervision, Funding acquisition.**

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests:

Highlights

1) Evaluation of surface water quality of northern Argentina rivers.

2) Systematic evaluation of physicochemical and biological parameters.

3) MA-PCA to model spatiotemporal variations of water quality parameters.

4) Evaluation of pollutant discharge, upstream and downstream areas

5) Classification model to predict river self-depuration.

[1] A. Cundy, J.S. Neil, B.B. Paul, International Encyclopedia of the Social & Behavioral Sciences, Pergamon, Oxford, 2001.

[2] C.H. Walker, S.P. Hopkin, R.M. Sibly, D.B. Peakall, Principles of Ecotoxicology, Taylor & Francis, Glasgow, 2001.

[3] D.A. Wright, P. Welbourn, Environmental Toxicology, Cambridge University Press, New York, 2002.

[4] F.D. Cid, R.I. Antón, R. Pardo, M. Vega, E. Caviedes-Vidal, Modelling spatial and temporal variations in the water quality of an artificial water reservoir in the semiarid Midwest of Argentina, Anal. Chim. Acta 705 (2011) 243–252. https://doi.org/10.1016/j.aca.2011.06.013

https://cfpub.epa.gov/si/si_public_record_Report.cfm?Lab=NHEERL&dirEntryID=60116 (last access: September 18th, 2020).

[6] K. Wang, P. Wang, R. Zhang, Z. Lin, Determination of spatiotemporal characteristics of agricultural non-point source pollution of river basins using the dynamic time warping distance, J. Hydrol. 583 (2020) 124303. 10.1016/j.jhydrol.2019.124303

[7] C.J. Vörösmarty, P.B. McIntyre, M.O. Gessner, D. Dudgeon, A. Prusevich, P. Green, S. Glidden, S.E. Bunn, C.A. Sullivan, C.R. Liermann, P.M. Davies, Global threats to human water security and river biodiversity. Nature 467 (2010) 555-561. https://doi.org/10.1038/nature09440

[5] M. A. Lewis, Nonpoint source pollution. Presented at Escambia County Task Force Meeting, Pensacola Junior College Media Center, Pensacola, FL, 9 November 1999. Website

[8] R. Aguilera, S. Sabater, R. Marcé, A Methodological Framework for Characterizing the Spatiotemporal Variability of River Water-Quality Patterns Using Dynamic Factor Analysis, J. Environ. Inform. 31 (2016) 97-110. https://doi.org/10.3808/jei.201600333

[9] R.P. Schwarzenbach, T. Egli, T.B. Hofstetter, U. Von Gunten, B. Wehrli, Global water pollution and human health. Annu. Rev. Environ. Resour. 35 (2010) 109–136. https://doi.org/10.1146/annurev-environ-100809-125342

[10] S. Liu, D. Ryu, J.A. Webb, A. Lintern, D. Waters, D. Guo, A.W. Western, Characterisation of spatial variability in water quality in the Great Barrier Reef catchments using multivariate statistical analysis, Mar. Pollut. Bull. 137 (2018) 137-151. https://doi.org/10.1016/j.marpolbul.2018.10.019

[11] H. Zia, N.R. Harris, G.V. Merrett, M. Rivers, N. Coles, The impact of agricultural activities on water quality: A case for collaborative catchment-scale management using integrated wireless sensor networks, Comput. Electron. Agric. 96 (2013) 126-138. https://doi.org/10.1016/j.compag.2013.05.001

[12] K. Luo, X. Hu, Q. He, Z. Wu, H. Cheng, Z. Hu, A. Mazumder, Using multivariate techniques to assess the effects of urbanization on surface water quality: a case study in the Liangjiang new area, China, Environ. Monit. Assess. 189 (2017) 174. https://doi.org/10.1007/s10661-017-5884-8.

[13] N.M. Gazzaz, M.K. Yusoff, M.F. Ramli, A.Z. Aris, H. Juahir, Characterization of spatial patterns in river water quality using chemometric pattern recognition techniques, Mar. Pollut. Bull. 64 (2012) 688–698. https://doi.org/10.1016/j.marpolbul.2012.01.032.

[14] A. Astel, M. Biziuk, A. Przyjazny, J. Namiesnik, Chemometrics in monitoring spatial and temporal variations in drinking water quality, Water Res. 8 (2006) 1706–1716. https://doi.org/10.1016/j.watres.2006.02.018

[15] X. Fan, B. Cui, H. Zhao, Z. Zhang, H. Zhang, Assessment of river water quality in Pearl River Delta using multivariate statistical techniques, Procedia Environ. Sci. 2 (2010) 1220–1234. https://doi.org/10.1016/j. proenv.2010.10.133

[16] R.L. Olsen, R.W. Chappell, J.C. Loftis, Water quality sample collection, data treatment and results presentation for principal components analysis—literature review and Illinois River watershed case study, Water Res. 46 (2012) 3110–3122. https://doi.org/10.1016/j.watres.2012.03.028

[17] Y. Wang, P. Wang, Y. Bai, Z. Tian, J. Li, X. Shao, L.F. Mustavich, B.L. Li, Assessment of surface water quality via multivariate statistical techniques: a case study of the Songhua

River Harbin region, China, J. Hydro-Environ. Res., 7 (2013) 30–40. https://doi.org/10.1016/j.jher.2012.10.003

[18] D. Walker, D. Jakovljević, D. Savić, M. Radovanović, Multi-criterion water quality analysis of the Danube River in Serbia: a visualisation approach, Water Res. (2015) 79 158–172. https://doi.org/10.1016/j.watres.2015.03.020

[19] C. Daou, M. Salloum, B. Legube, A. Kassouf, N. Ouaini, Characterization of spatial and temporal patterns in surface water quality: a case study of four major Lebanese rivers, Environ. Monit. Assess. 190 (2018) 485-500. https://doi.org/10.1007/s10661-018-6843-8

[20] K.P. Singh, A. Malik, D. Mohan, S. Sinha, Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)—a case study, Water Res. 38 (2004) 3980–3992. https://doi.org/10.1016/j.watres.2004.06.011

[21] Y.-H. Yang, F. Zhou, H.-C. Guo, H. Sheng, H. Liu, X. Dao, C.-J. He, Analysis of spatial and temporal water pollution patterns in Lake Dianchi using multivariate statistical methods, Environ. Monit. Assess. 170 (2010) 407–416. https://doi.org/10.1007/s10661-009-1242-9

[22] Q. Zhang, Z. Li, G. Zeng, J. Li, Y. Fang, Q. Yuan, Y. Wang, F. Ye, Assessment of surface water quality using multivariate statistical techniques in red soil hilly region: a case study of Xiangjiang watershed, China, Environ. Monit. Assess. 152 (2009), 123–131. https://doi.org/10.1007/s10661-008-0301-y

[23] F. Zhou, H. Guo, Y. Liu, Y. Jiang, Chemometrics data analysis of marine water quality and source identification in Southern Hong Kong, Mar. Pollut. Bull. 54 (2007a.) 745–756. https://doi.org/10.1016/j.marpolbul.2007.01.006

[24] S. Mitra, S. Ghosh, K.K. Satpathy, B.D. Bhattacharya, S.K. Sarkar, P. Mishra, P. Raja, Water quality assessment of the ecologically stressed Hooghly River Estuary, India: A

multivariate approach, Mar. Pollut. Bull. 126 (2018) 592-599. https://doi.org/10.1016/j.marpolbul.2017.09.053

[25] A. Bouguerne, A. Boudoukha, A. Benkhaled, A.-H. Mebarkia, Assessment of surface water quality of Ain Zada dam (Algeria) using multivariate statistical techniques, Int. J. River Basin Manag. 15 (2017) 133–143. https://doi.org/10.1080/15715124.2016.1215325

[26] R. Tauler, D. Barcelo, E.M. Thurman, Multivariate Correlation between Concentrations of Selected Herbicides and Derivatives in Outflows from Selected U.S. Midwestern, Reservoirs Environ. Sci. Technol. 34 (2000) 3307-3314. https://doi.org/10.1021/es000884m

[27] R. Tauler, S. Lacorte, M. Guillamon, R. Cespedes, P. Viana, D. Barceló, Chemometric Modeling of Main Contamination Sources in Surface Waters of Portugal, Environ. Tox. Chem. 23 (2004) 565-575. https://doi.org/10.1897/03-176b

[28] E. Pere-Trepat, L. Olivella, A. Ginebreda, J. Caixach, R. Tauler, Chemometrics modelling of organic contaminants in fish and sediment river samples, Sci. Total Environ. 371 (2006) 223–237. https://doi.org/ 10.1016/j.scitotenv.2006.04.005

[29] A. Navarro, R. Tauler, S. Lacorte, D. Barcelo, Chemometrical investigation of the presence and distribution of organochlorine and polyaromatic compounds in sediments of the Ebro River Basin. https://doi.org/10.1007/s00216-006-0451-0

[30] R. Pardo, M. Vega, L. Deban, C. Cazurro, C. Carretero, Modelling of chemical fractionation patterns of metals in soils by two-way and three-way principal component analysis, Anal. Chim. Acta 606 (2008) 26-36. https://doi.org/10.1016/j.aca.2007.11.004

[31] M. Felipe-Sotelo, J.M. Andrade, A. Carlosena, R. Tauler, Temporal characterisation of river waters in urban and semi-urban areas using physico-chemical parameters and

chemometric methods, Anal. Chim. Acta 583 (2007) 128–137. https://doi.org/10.1016/j.aca.2006.10.011

[32] H. Paoli, H. Elena, J. Mosciaro, F. Ledesma, Y. Noé, Caracterización de las cuencas hídricas de las provincias de Salta y Jujuy, Instituto Nacional de Tecnología Agropecuaria (INTA), Argentina, 1 December 2011.

Website:

https://inta.gob.ar/documentos/caracterizacion-de-las-cuencas-hidricas-de-las-provincias-de-salta-y-jujuy

(last access: November 15th, 2020).

[33] M.M. Salusso, Evaluación de la calidad de los recursos hídricos superficiales en la Alta Cuenca del Juramento (Salta). Tesis Doctoral Inédita, Facultad de Ciencias Exactas y Naturales. Universidad de Buenos Aires. 2005. https://bibliotecadigital.exactas.uba.ar/collection/tesis/document/tesis_n3872_Salusso?p.s=TextQuery.

[34] J.D. Brea, P. Spalletti, Generación y transporte de sedimentos en la Cuenca Binacional del Río Bermejo. Caracterización y análisis de los procesos intervinientes, first ed., COBINABE, Buenos Aires, 2010.

Website:

https://www.ina.gov.ar/pdf/Libro-Generacion-Transporte-Sedimentos-Cuenca-Bermejo.pdf

(last access: November 15th, 2020).

[35] APHA, Standard Methods for the Examination of Water and Wastewater, American Public Health Association (APHA), American Water Works Association (AWWA) and Water Environment Federation (WEF), Washington, DC, 2017.

[36] R. Bro and A. K. Smilde, Principal component analysis, Anal. Methods. 6 (2014) 2812–2831. http://dx.doi.org/10.1039/c3ay41907j

[37] M. Cocchi, A. Biancolillo, F. Marini, Chemometric Methods for Classification and Feature Selection, Compr. Anal. Chem. 82 (2018) 265-299. https://doi.org/10.1016/bs.coac.2018.08.006

[38] R. W. Kennard and L. A. Stone, Computer aided design of experiments, Technometrics 11 (1969) 137–148. https://doi.org/10.1080/00401706.1969.10490666

[39] D. Ballabio, F. Grisonia, R. Todeschini, Multivariate comparison of classification performance measures, Chemom. Intell. Lab. Syst. 174 (2018) 33–44. https://doi.org/10.1016/j.chemolab.2017.12.004

[40] S.M. Azcarate, A. de Araújo Gomes, A. Muñoz de la Peña, H.C. Goicoechea, Modeling second-order data for classification issues: Data characteristics, algorithms, processing procedures and applications, TrAC – Trend. Anal. Chem. 107 (2018) 151-168. . https://doi.org/10.1016/j.trac.2018.07.022

[41] D. Ballabio, V. Consonni, Classification tools in chemistry. Part 1: Linear models. PLS-DA, Anal. Methods 5 (2013) 3790-3798. https://doi.org/10.1039/C3AY40582F

[42] Ministerio de Asuntos Agrarios y Producción. VERTIDO DE EFLUENTES - MODIFICACIÓN RES. AGOSBA 389/98. Resolución (MAAyP) 336/03. 15/10/2003. B.O.: 25/11/2003. Modificaciones a la Resolución N° 389/98., Buenos Aires, Argentina.

Website:

http://www.ecofield.net/Legales/BsAs/res336-03_BA.htm

(last access: November 15th, 2020).

[43] L. Lupi, L. Bertrand, M.V. Monferrán, M.V. Amé, M.P. Diaz, Multilevel and structural equation modeling approach to identify spatiotemporal patterns and source characterization of metals and metalloids in surface water and sediment of the Ctalamochita river in Pampa region, Argentina, J. Hydrol. 572 (2019) 403–413. https://d 10.1016/j.jhydrol.2019.03.019

TABLES

Table 1. Details of the studied sampling zones and sites.

Sampling zone	Sampling site	Sample Location	Description
Z1 (Arenales river)	S1	Upstream S2	8 km upstream discharge point
	S2	South Treatment Plant of Salta city discharge	Discharge of urban effluents
	S3	Downstream S2	Entrance to the Cabra Corral Dam
Z2 (Bermejo river)	S4	Near to Aguas Blancas city	Urban region
	S5	In the catchment of the water treatment plant in Embarcación city	Drinking water source
Z3 (Juramento	S6	Upstream S7	52 km upstream discharge point

basin)	S7	El Galpón city discharge	Discharge of urban effluents
	S8	Downstream S7	Exit of El Tunal Dam
Z4 (Mojotoro river)	S9	Upstream S10	4 km upstream discharge point
	S10	Capital and Capo Santo cities discharge	Discharge of urban effluents
	S11	Downstream of S10	10 km downstream discharge point
	S12	Downstream of S11	Discharge of urban effluents
	S13	Industrial Park of Güemes discharge	8 km downstream discharge point
Z5 (Rosario river)	S14	Upstream S15	1 km upstream discharge point
	S15	Rosario de Lerma city	Discharge of urban effluents
	S16	Downstream S15	3 km downstream discharge point
Z6 (Horcones river)	S17	Upstream S18	1 km upstream unloading
	S18	Rosario de la Frontera city	Discharge of urban effluents
	S19	Downstream S18	Agricultural area. 12 km downstream discharge point

Table 2. Water quality parameters, analytical methods and instrumentation.

Parameter	**Coded analytical method ***	**Analytical technique**	**Materials and instruments**
WT	SM 2550B	Direct measurement	Stainless steel digital thermometer
pH	SM 4500 B	Potentiometry	pH-meter HACH sensION pH1
C	SM 2510B	Conductimetry	Conductivity meter HACH sensION EC5
SS10	SM 2540 F	Volumetry	Imhoff Cones
SS2h	SM 2540 F	Volumetry	
OD	SM 4500 G	Potentiometry ISE	Oximeter HACH sensION DO6
S	SM 4500 F	Iodometry	

TKN	SM 4500 Norg B, C (TKN)	Titration	
NH_4	SM 4500 NH_3, C	Titration	
Norg	SM 4500 Norg B, C	Titration	
BOD	SM 5210 B	Dilution	BOD Incubator
COD	SM 5220 D	Molecular absorption spectroscopy (Colorimetry at 600 nm)	Spectrophotometer UV-Vis HACH DR5000
Phen	SM 5530 B, D	Molecular absorption spectroscopy (Colorimetry at 500 nm)	
TP	SM 4500 C	Molecular absorption spectroscopy (Colorimetry at 800 nm)	
FC and TC	SM 9221 B, C, E	Multiple tube fermentation technique	Test and durham tubes, water bath
B	SON-A-1982-1323	Molecular absorption spectroscopy (Colorimetry at 414 nm)	Spectrophotometer UV-Vis Cintra GBC UV
Fe, Mn, Cr, Zn, Cd, Cu and Pb;	SM 3111B	Flame Spectrometry Atomic Absorption (FSAA)	Atomic absorption spectrophotometer Agilent AA 55B
Hg	SM 3112B	Cold Vapor-Hydride Generator-Spectrometry Atomic Absorption (CV-HG-SAA)	Atomic absorption spectrophotometer GBC AA 904 coupled to a hydride generator GBC HG3000
As and Se	SM 3114C	Hydride Generator- FSAA (HG-FSAA)	

*SM: Standard Methods for the Examination of Water and Wastewater, 23nd edition. [27]
SON: State official newsletter. Official methods of water analysis. Spain [43].

Table 3. Minimum and maximum limits obtained for the physiochemical and microbiological parameters of the Salta rivers and the standard water quality recommended by [42].

Variable	Unit	Minimum	Maximum	Maximum permissible
pH		6.12	9.33	6.5-10
C	$\mu S\ cm^{-1}$	105	2250	
SS10	$mL\ L^{-1}$	ND	2.5	Absence
SS2	$mL\ L^{-1}$	ND	12.5	≤ 1.0
DO	$mg\ L^{-1}$	ND	12.5	
TKN	$mg\ L^{-1}$	ND	65.1	≤ 10
NH_4	$mg\ L^{-1}$	ND	58.8	≤ 50
Norg	$mg\ L^{-1}$	ND	11.34	
BOD	$mg\ L^{-1}$	ND	209	≤ 25

COD	mg L^{-1}	1	623	≤ 250
TP	mg L^{-1}	ND	4.5	≤ 10
FC	MPN[b]/100 mL	ND	90000000	≤ 2000
TC	MPN[b]/100 mL	ND	90000000	
B	mg L^{-1}	0.03	1.22	≤ 2.0
Fe	mg L^{-1}	ND	1.5	≤ 2.0
Mn	mg L^{-1}	ND	1.28	≤ 0.5
Cr	mg L^{-1}	ND	0.12	≤ 0.1
Zn	mg L^{-1}	ND	0.31	≤ 2.0
Cd	mg L^{-1}	ND	0.03	≤ 0.1
Cu	mg L^{-1}	ND	0.17	≤ 1.0
Pb	mg L^{-1}	ND	0.18	≤ 0.1
Hg	mg L^{-1}	ND	8	≤ 0.005
As	µg L^{-1}	ND	5	≤ 0.5

[a] ND: Not detected (< Detection limit)
[b] MPN: Most probable number

Table 4. Confusion matrices corresponding to PCA-DA and sensitivity (S), specificity (SP) and precision (PR) resulting from training and prediction sets.

	Real/Predicted	**DU***	**DP***	**Sensitivity**	**Specificity**	**Precision**
Training set (CV)	**DU***	88	2	0.98	0.89	0.95
	DP*	4	39	0.89	0.98	0.95
Prediction set	**DU***	38	0	1	0.89	0.95
	DP*	2	17	0.89	1	1

*Upstream and downstream (DU) from discharge point (DP).