Assessing temporal and spatial patterns of surface-water quality with a multivariate approach: a case study in Uruguay

A Gorgoglione1⁎, J Alonso1, C Chreties1 and M Fossati1

1Department of Fluid Mechanics and Environmental Engineering (IMFIA), School of Engineering, Universidad de la República, Montevideo 11300, Uruguay

⁎Email: agorgoglione@fing.edu.uy

Abstract. The ecological state of inland waters of the Santa Lucía watershed, the primary drinking water source of Uruguay, has raised interest since it presents the seasonal phenomenon of eutrophication. For this reason, an in-depth understanding of the behaviour in time and space of the water-quality variables that characterize this stream is essential. Therefore, this study aims to evaluate the occurrence of spatial and temporal patterns of water-quality variables (Q, turbidity, T, TN, NO3−, NO2−, NH4+, TP, DO, BOD5) in the Santa Lucía Chico watershed with the aid of multivariate statistical tools. The principal component analysis, coupled with k-means cluster analysis, helped to identify a seasonal variation (fall-winter and spring-summer). The hierarchical cluster analysis allowed us to classify the water-quality monitoring stations in three groups in the fall-winter season. The loadings values of the cluster analysis highlighted the most significant pollutants at each monitoring station. The outcomes of this work are expected to contribute valuable knowledge for determining effective management strategies to reduce stream pollution and protect the aquatic ecosystem health of the study area.

1. Introduction

Water quality is considered worldwide as a measure that can assess the usage of water for different purposes (drinking, industrial, agricultural, habitat, and recreational) using several parameters, like physical, biological, and chemical. Water quality plays a crucial role in all aspects of living organisms on the earth, which attracts the attention of environmental researchers. Water quality varies based on time, location, weather, and the presence of different types of pollution sources [1]. Preserving adequate water quality is a challenging task mainly due to point and non-point sources (NPS) of pollution.

Land use is one of the components of the landscape that plays a critical role in the generation of pollution. In particular, agricultural and urban land use are the primary NPS of nutrients that degrade surface-water quality [2]. Degradation of water quality intensifies phytoplankton biomass and algal bloom, the odor of water, and drinking water purification cost, which finally negatively impacts aquatic and terrestrial ecosystems [3]. Therefore, an effective and comprehensive long-term management of water resources requires a thorough understanding of the entire system, including hydro-morphological, chemical, and biological characteristics.

Water quality is influenced by numerous factors, including all the aspects related to the interaction of the stream with its contributing catchment. Multivariate-statistical techniques, such as cluster analysis (CA) and principal component analysis (PCA), are strategic approaches for i) characterizing water quality; ii) discovering spatial and temporal variations produced by natural and anthropogenic factors;
asses correlations among different land uses and water-quality parameters to recognize the key factors threatening water quality and affecting the aquatic ecosystem [2][4][5][6][7].

Based on these aspects, the goal of this study is to explore the spatial and temporal variability of water quality related to land use in the Santa Lucí a Chico (SLC) watershed (Uruguay). To analyze these variations, the occurrence of spatial and temporal patterns was assessed with the aid of multivariate statistical methods (CA and PCA). The application of these techniques will offer a valuable tool for reliable management of water resources and efficient solutions to pollution problems.

2. Materials and methods

2.1. Study watershed description

The Santa Lucí a (SL) catchment (Figure 1 – on the left, highlighted in light brown) is one of the principal sources of potable water of Uruguay. It provides water for more than half (60%) of the country population and maintains numerous agricultural and industrial activities [8][9]. The Paso Severino (PS) reservoir is positioned in Santa Lucí a Chico (SLC) watershed, a sub-catchment of SL (Figure 1 – on the right). It is characterized by a lake surface equal to 15 km$^2$, a maximum level of 20 m, and an accumulation capacity equal to 65 hm$^3$. Its contributing area is equal to 2500 km$^2$. A water-treatment plant is situated 35 km-downstream the PS.

Considering the national importance of this catchment, to satisfy all water uses, it is crucial to have suitable water-resource management strategies to assure the appropriate water-quality standard. For this reason, SLC was chosen as a case study.

The SLC-watershed surface is equal to 2570 km$^2$. Here, the total annual precipitation ranges between 1.0 m and 1.5 m, and the average temperature can fluctuate between 3$^\circ$C and 30$^\circ$C, respectively in winter and summer. SLC elevation varies between 25 m a.s.l. (PS lake) and 200 m a.s.l. (north-eastern area of the catchment). The soil is mainly formed by Bruno Sols with loam texture and it has a high content of soil organic matter. The southern part of the watershed is characterized by higher clay and silt proportions. The actual land use/cover is mainly natural and cultivated pastures for dairy and beef cattle grazing (60%) followed by rainfed croplands (30%) and forest (5%) [10].

![Figure 1. Santa Lucí a Chico (SLC) catchment with hydrometric and water-quality monitoring stations (respectively green rhombus and red dots).](image-url)
the SLC mainstream (Figure 1). It is important to remark that three of these stations (SLC01, SLC02, and PS01) are positioned upstream the PS reservoir, while PS03, PS04, and PS02 can be found in the lake.

The streamflow dataset was developed by the Uruguayan National Water Board (DINAGUA). The water level was measured three times a day during 2004-2018 from the hydrometric station placed in the city of Florida (Figure 1).

For this study, the following variables (attributes) were taken into account: flow rate (Q), turbidity, water temperature (T), total nitrogen (TN), nitrate (NO$_3^-$), nitrite (NO$_2^-$), ammonium (NH$_4^+$), total phosphorus (TP), dissolved oxygen (DO), and biochemical oxygen demand (BOD$_5$).

It is important to remark that the mentioned-above water-quality attributes are aligned with the predominant agricultural land use of the study area.

2.3. Data analysis

Traditional multivariate-data analysis, PCA and CA (like k-means and hierarchical cluster analysis (HCA)), were used in this study to detect spatial similarities of the monitoring sites and the occurrence of temporal trends in water-quality parameters. Both PCA and CA were programmed and run in R by adopting the libraries “cluster,” “HSAUR,” “vegan,” “devtools,” and “ggbiplot.” These are unsupervised methods, e.g., any information about other response variables, or cluster belonging, are not involved in getting the outcomes. For this reason, these techniques are appropriate for exploratory data analysis, whose goal is assumption creation more than assumption confirmation [12].

The cluster analysis aims to divide a multifaceted database into different groups so that the within-cluster similarities are higher than the between-cluster ones [13]. These groups are able to identify patterns related to the studied process. The similarity between observations is measured by a distance function. First of all, the similarity is calculated among all data points. Afterward, once the latter starts grouping, the similarity is calculated among clusters too. Numerous metrics exist to calculate this similarity, and its choice could affect the results. A priori and arbitrary choice of the group number (k) is also required for this technique.

PCA is a method that reduces and transforms measured data into new uncorrelated variables known as principal components (PCs). It has the aim of revealing possible patterns among data points and variables [14]. The raw measured data are considered as independent variables. Every single PC is a linear combination of the original variables. The PCs make the basis of the respective vector space and they are organized by decreasing variance. Consequently, PC1 (the first PC) carries the most information about the original data, and so on [15].

3. Results and discussion

3.1. Temporal patterns

A PCA associated with k-means clustering was conducted at each water-quality monitoring site to evaluate temporal trends in water quality. With this aim, ten attributes were taken into account: Q, turbidity, T, TN, NO$_3^-$, NO$_2^-$, NH$_4^+$, TP, DO, and BOD$_5$. Taking into account the different monitoring frequency of the variables involved in this analysis, the monitoring dates of water-quality variables were taken as a reference along with an antecedent period (AP) equal to one month, one week, and one day. Since the results obtained for these three cases were quite similar, in this work, we are just presenting the results obtained for AP=1 week. The flow rate considered was the cumulative flow rate during the AP. Since the six monitoring locations do not have the same number (n) of samples, six matrices with a dimension of (n×10) were used to run PCA/k-means, being n=63 at SLC02; n=39 at SLC01, PS03, PS04, PS02; and n=51 at SLC03. Before the classification, the variables (raw-data) were normalized (mean = 0, standard deviation = 1) to give equal weight to each of them and deal with their various measurement units. The k-means was run for the first and second dimension (called Dim in Figure 2) or PC since they covered 46.4% of the variance. In Figure 2, the spider plot obtained for the sampling site
SLC01 is represented. This plot represents the PCA scores plot (data points), to which the k-means algorithm was applied.

In Figure 2, we can detect two distinct groups: Winter Cluster (k1), which incorporates fall and winter water-quality observations, and Summer Cluster (k2), which includes spring and summer water-quality observations. In particular, the month grouping chosen for the two detected seasons was arbitrary and it consists of Apr.-May-Jun.-Jul.-Aug.-Sep. for k1, and Oct.-Nov.-Dec.-Jan.-Feb.-Mar. for k2. Hence, October and April are “transition months,” that can belong indistinctly to the Winter or Summer Cluster, based on the month grouping. The latter clarifies the presence of two data points from April in k1. It is worth remarking the occurrence of two main seasons (spring/summer and fall/winter) in the water quality of SLC. Similar outcomes were achieved at the remaining five water-quality monitoring sites.

3.2. Spatial similarity
HCA was adopted to spatially compare water-quality sampling stations and eventually identifying the occurrence of spatial similarities among them. Considering the seasonal trend previously discovered, two matrices with a dimension of \( (6 \times 9) \) were used as input of the HCA, where six are the sampling stations and nine are the hydrologic/water-quality attributes (Q, turbidity, T, TN, NO_3, NO_2, NH_4, TP, and DO). An extra column was used as a label for the sampling site identification. The two data matrices were constructed by dividing all data points into two groups depending on the season (fall-winter and spring-summer) and, for both of them, calculating the median of each attribute at each sampling station. The variable BOD_5 was not taken into account for this analysis since it was not measured at three stations (PS02, PS03, and PS04). Also for this analysis, the input matrices were normalized. In Figure 3, the dendrograms resulted from the two periods (fall-winter and spring-summer) using Euclidean distance and Ward linkage are presented.
Summer and Winter at PC1 is contributed by turbidity and $Q$ at all monitoring stations. This component can be interpreted as representative of the physical water characteristics. PC2 can be explained mainly with TN. This can justify the presence of turbidity, for the particle-bound nitrogen, and NO$_3^-$, for the dissolved nitrogen, as significant variables to represent PC2. The lower the variance of a PC, the more water-quality attributes characterize that PC. An example is represented by PC3 and PC4, where it is impossible to identify a single contributing variable.

It is also worth remarking that SLC01, SLC02, and PS01 (the three upstream stations) are always characterized by turbidity and $Q$ with the same sign (positive or negative) in PC1 and PC2. This confirms the water-quality spatial characterization at each monitoring site are described in Table 1.
the previous outcome of the net cluster founded in winter for these three stations, justified by the fact that water balance is considered as the driving force of sediment transport.

Table 1. PCA loadings outcomes at: (a) SLC01, (b) SLC02, (c) PS01, (d) PS03, (e) PS04, and (f) PS02. The variance of each PC is represented by numbers in parenthesis.

| Variables  | PC1 (31.5%) | PC2 (17.5%) | PC3 (15.5%) | PC4 (12.5%) |
|------------|-------------|-------------|-------------|-------------|
| TP         | -0.26       | 0.15        | -0.50       | 0.39        |
| TN         | 0.09        | 0.59        | 0.35        | -0.04       |
| NO₂        | -0.05       | 0.42        | 0.06        | 0.71        |
| NO₃        | 0.32        | -0.32       | 0.09        | 0.06        |
| NH₄⁺       | 0.13        | 0.43        | -0.24       | 0.49        |
| Turbid.    | 0.49        | -0.41       | -0.01       | 0.27        |
| T          | -0.51       | -0.18       | 0.16        | 0.06        |
| DO         | 0.34        | -0.09       | 0.50        | 0.05        |
| Q          | 0.26        | 0.15        | -0.50       | 0.39        |

| Variables  | PC1 (33.3%) | PC2 (16.2%) | PC3 (13.1%) | PC4 (10.5%) |
|------------|-------------|-------------|-------------|-------------|
| TP         | -0.23       | 0.47        | -0.08       | -0.48       |
| TN         | -0.35       | 0.41        | 0.21        | 0.03        |
| NO₂        | -0.38       | 0.25        | 0.03        | -0.20       |
| NO₃        | -0.09       | 0.11        | -0.67       | -0.16       |
| NH₄⁺       | 0.22        | 0.22        | 0.53        | -0.42       |
| Turbid.    | -0.43       | 0.30        | 0.07        | 0.13        |
| T          | 0.39        | 0.41        | 0.00        | 0.34        |
| DO         | -0.38       | -0.34       | 0.03        | -0.48       |
| BOD₅       | 0.00        | -0.28       | 0.46        | 0.13        |
| Q          | -0.37       | 0.17        | 0.05        | 0.38        |

| Variables  | PC1 (32.1%) | PC2 (15.3%) | PC3 (13.1%) | PC4 (9.3%)  |
|------------|-------------|-------------|-------------|-------------|
| TP         | -0.44       | 0.05        | -0.01       | 0.27        |
| TN         | -0.31       | 0.46        | 0.19        | 0.08        |
| NO₂        | -0.15       | 0.39        | -0.58       | -0.05       |
| NO₃        | -0.34       | 0.05        | 0.02        | 0.47        |
| NH₄⁺       | -0.34       | 0.24        | 0.29        | 0.14        |
| Turbid.    | 0.19        | 0.59        | 0.29        | -0.16       |
| T          | -0.41       | -0.25       | -0.01       | -0.39       |
| DO         | 0.35        | -0.06       | -0.09       | 0.70        |
| BOD₅       | 0.04        | 0.26        | -0.65       | -0.05       |
| Q          | 0.37        | 0.30        | 0.20        | -0.09       |

In Table 2, it is clear the lack of spatial patterns of the water-quality variables. This reflects the results obtained with the HCA in the spring-summer season. Overall, it exists a spatial distribution in the fall-winter season. However, this spatial pattern is not clear anymore when a comprehensive analysis is run at each monitoring station and for each pollution.

Table 2. Most representative water-quality/hydrologic variables at each monitoring site.

| Monitoring station | Variables with significant loading                                  |
|--------------------|---------------------------------------------------------------------|
| SLC01              | T, NO₂⁺                                                             |
| SLC02              | NO₂⁺, NH₄⁺                                                           |
| PS01               | Turbidity, NO₂⁺, BOD₅, DO                                           |
| PS03               | T, TN, TP, Q, NO₂⁺                                                 |
| PS04               | Turbidity, T, NH₄⁺, NO₂⁺, NO₃⁻                                     |
| PS02               | Turbidity, TN, NO₂⁺, DO, NH₄⁺                                     |

4. Conclusions

Multivariate exploratory analyses were used in this study to assess variations in the surface-water quality of SLC watershed. The following main conclusions can be summarized:

- The PCA/k-means categorized the hydrologic/water-quality attributes in Winter and Summer Cluster, which respectively represent fall-winter and spring-summer periods.
- The HCA revealed spatial similarity during the cold season (fall-winter) between PS04 and PS03, the two monitoring stations positioned in the PS lake, and among PS01, SLC01, and SLC02, that are located upstream the reservoir, whose contributing watersheds are characterized by natural and cultivated pastures.
• The PCA loadings values were useful to detect the most representative water-quality attributes. PCI represents the physical water-quality variables (T and turbidity); PC2 is contributed by TN, in its dissolved and particle-bound form.

The results of this study have demonstrated that the combination of large-scale studies and multivariate statistical techniques can adequately describe the behavior in time and space of water-quality. Furthermore, the outcomes can satisfactorily explain water-quality fluctuations due to land use. An enhanced design of temporal and spatial sampling, along with further studies, may better define the complex land use-water quality relationship.

References
[1] Giri S and Qiu Z 2016 Understanding the relationship of land uses and water quality in Twenty First Century: A review J. Environ. Manage. 173 41-8
[2] Rodriguez-Romero A J, Rico-Sánchez A E, Mendoza-Martínez E, Gómez-Ruiz A, Sedeño-Díaz J E, López-López E 2018 Impact of changes of land use on water quality, from tropical forest to anthropogenic occupation: a multivariate approach Water 10 1518
[3] Zhang W, Li H, Sun D, Zhou L. 2012 A statistical assessment of the impact of agricultural land use intensity on regional surface water quality at multiple scales Int. J. Env. Res. Pub. He. 9 4170-86
[4] Gorgoglione A, Bombardelli F A, Pitton B J L, Oki L R, Haver D L, Young T M 2018 Role of sediments in insecticide runoff from urban surfaces: analysis and modeling Int. J. Env. Res. Pub. He. 15 1464
[5] Gorgoglione A, Crisci M, Kayser R H, Chreties C, Collischonn W 2019 A New Scenario-Based Framework for Conflict Resolution in Water Allocation in Transboundary Watersheds Water 11 1174
[6] Li P, Tian R, Liu R 2019 Solute Geochemistry and Multivariate Analysis of Water Quality in the Guohua Phosphorite Mine, Guizhou Province, China Exposure and Health 11 81-94
[7] Wu J, Li P, Qian H, Duan Z, Zhang X 2014 Using correlation and multivariate statistical analysis to identify hydrogeochemical processes affecting the major ion chemistry of waters: Case study in Laoheba phosphorite mine in Sichuan, China Arab. J. Geosci. 7 3973-82
[8] Navas R, Alonso J, Gorgoglione A, Vervoort R W 2019 Identifying climate and human impact trends in streamflow: a case study in Uruguay Water 11 1433
[9] Goyenola G, Meierhoff M, Teixeira-de Mello F, González-Bergonzoni I, Graeber D, Fosalba C, Vidal N, Mazzeo N, Ovesen N B, Jeppesen E and Kronvang B 2015 Phosphorus dynamics in lowland streams as a response to climatic, hydrological and agricultural land use gradients Hydrol Earth Syst. Sci. Discuss. 12 3349–90
[10] MGAP 2018 Integrated map of land cover/use of Uruguay (Montevideo: Ministry of Livestock, Agriculture and Fisheries)
[11] DINAMA. OAN–Observatorio Ambiental Nacional. Available online: https://www.dinama.gub.uy/oan/geoportal/ (Accessed on 20 November 2019)
[12] Gorgoglione A, Gioia A, Iacobellis V 2019 A Framework for assessing modeling performance and effects of rainfall-catchment-drainage characteristics on nutrient urban runoff in poorly gauged watersheds Sustainability 11 4933
[13] Jain A K, Murty M N, Flynn P J 1999 Data Clustering: A Review ACM Comput. Surv. 31 264-323
[14] Massart D L, Vandeginste B G M, Deming S M, Michotte Y, Kaufman L 1988 Chemometrics-A Text Book (Elsevier: Amsterdam, The Netherlands) chapters 1–4 pp 14–21
[15] He Y, Gao B, Sophian A and Yang R 2017 Coil-Based Rectangular PEC Sensors for Defect Classification (Transient Electromagnetic-Thermal Nondestructive Testing) pp 55-90
[16] Dutta S, Dwivedi A, Kumar M S 2018 Use of water quality index and multivariate statistical techniques for the assessment of spatial variations in water quality of a small river Env. Monit. Assess. 190 718