



Evaluation and monitoring of the water quality of an Argentinian urban river applying multivariate statistics

Jesica Alejandra Tello^{1,2} · Jorge Leandro Leporati³ · Patricia Laura Colombetti⁴ · Cynthia Gabriela Ortiz⁵ · Mariana Beatriz Jofré^{1,4} · Gabriela Verónica Ferrari^{1,2} · Patricia González^{1,2}

Received: 24 October 2023 / Accepted: 1 April 2024 / Published online: 10 April 2024
© The Author(s), under exclusive licence to Springer-Verlag GmbH Germany, part of Springer Nature 2024

Abstract

In this work, we present the water quality assessment of an urban river, the San Luis River, located in San Luis Province, Argentina. The San Luis River flows through two developing cities; hence, urban anthropic activities affect its water quality. The river was sampled spatially and temporally, evaluating ten physicochemical variables on each water sample. These data were used to calculate a Simplified Index of Water Quality in order to estimate river water quality and infer possible contamination sources. Data were statistically analyzed with the opensource software R, 4.1.0 version. Principal component analysis, cluster analysis, correlation matrices, and heatmap analysis were performed. Results indicated that water quality decreases in areas where anthropogenic activities take place. Robust inferential statistical analysis was performed, employing an alternative of multivariate analysis of variance (MANOVA), MANOVA.wide function. The most statistically relevant physicochemical variables associated with water quality decrease were used to develop a multiple linear regression model to estimate organic matter, reducing the variables necessary for continuous monitoring of the river and, hence, reducing costs. Given the limited information available in the region about the characteristics and recovery of this specific river category, the model developed is of vital importance since it can quickly detect anthropic alterations and contribute to the environmental management of the rivers. This model was also used to estimate organic matter at sites located in other similar rivers, obtaining satisfactory results.

Keywords Urban river · Water quality · Anthropogenic pollution · Statistical model · Multivariate analysis · Robust statistics · River monitoring

Responsible Editor: Marcus Schulz

✉ Jesica Alejandra Tello
tjesik32@gmail.com

¹ Instituto de Química San Luis (INQUISAL, CONICET), Almirante Brown 907, 5700 San Luis, Argentina

² Departamento de Química, Facultad de Química Bioquímica y Farmacia, Universidad Nacional de San Luis, Avenida Ejército de los Andes 950, 5700 San Luis, Argentina

³ Departamento de Ciencias Básicas, Facultad de Ingeniería y Ciencias Agropecuarias, Universidad Nacional de San Luis, Ruta Provincial 55 (Ex 148) - Extremo Norte, Villa Mercedes, San Luis, Argentina

⁴ Departamento de Biología, Facultad de Química Bioquímica y Farmacia, Universidad Nacional de San Luis, Avenida Ejército de los Andes 950, 5700 San Luis, Argentina

⁵ Departamento de Educación y Formación Docente, Facultad de Ciencias Humanas, Universidad Nacional de San Luis, Almirante Brown 951, 5700 San Luis, Argentina

Introduction

Rivers are one of the main natural sources used for drinking and utility water supply, aquaculture, agricultural irrigation, and energy production (Gupta and Gupta 2021a). Nevertheless, rivers are not simply biophysical phenomena that provide services to people. There is a connection between cities and rivers, a bi-directional or reciprocal relationship maintained by humans and rivers since they are social–ecological systems (Anderson et al. 2019). These systems contribute to improving the landscape of the city with parks and riverside forests, but they also become sources of waste and pollution, since they can be used as landfill for sewage and industrial effluents (Gatica et al. 2012). These freshwater resources also play important roles in urban areas, such as serving as carriers of water and suspended solids, providing habitats for a diverse and productive biota, and acting as social and cultural elements for the human inhabitants living in the

watershed (Walsh et al. 2005). The relationships between cities and water bodies, as well as the issues related to the quantity and quality of urban water, change in response to the expansion and development of cities (Schirmer et al. 2013; Walsh et al. 2005). In developing countries, urban waterfronts have been transformed into mixed residential and commercial areas with high-density infrastructure and services. Currently, rivers and their banks are considered natural urban ecosystems that provide a wide development potential in cities (Khairabadi et al. 2023).

When there are urban uses in the hydraulic public domain or protection zones of a section of a river and/or the section is immersed in an urban matrix, that section should be considered an urban river (Duran Vian et al. 2020). Urban rivers are important not only because they constitute the habitat of species that control and maintain the ecological state of the basin in which they are located but also because they represent a considerable resource that is part of the productive, social, and cultural activities of the surrounding community (Obisesan and Christopher 2018).

Urban rivers face their own challenges: they are in transition with numerous tensions (social, political, economic, ecological, and environmental tensions) leading to water scarcity, decreasing water quality, and climatic variability (Ghimire et al. 2022). The water quality of an urban river is affected by several variables, i.e., geography, topography, atmosphere, anthropogenic activities, etc. (Gupta and Gupta 2021b). The impact of anthropogenic activities on urban rivers (irregular urbanization and industrialization, population growth, use of fertilizers and pesticides in agriculture, discharge of domestic wastewater, removing sand-gravel from stream beds, etc.) has increased substantially in recent years, leading to the rapid loss of their natural character (Schirmer et al. 2013; Ustaoğlu et al. 2020; Das et al. 2023; Ghimire et al. 2022; Gupta et al. 2022). This has become a serious problem for many other related aquatic environments, such as streams, lakes, and estuaries (Walsh et al. 2005). This situation is one of the main environmental problems in the world, as the increase in human population raises the need for freshwater (Ustaoğlu et al. 2020).

In urban areas, streams receive groundwater, treated and untreated wastewater, and industrial waste, among others, and are often degraded by a multitude of stressors (Obisesan and Christopher 2018; Ghimire et al. 2022). This degradation is called the “syndrome of urban rivers” (Schirmer et al. 2013; Walsh et al. 2005). Urban rivers are very sensitive to changes in land use, and the response is a warning signal of potential downstream water deterioration (Kominkova 2013). Anthropogenic activities deteriorate the quality of river water and prevent their utilization for other purposes such as drinking water or irrigation (Gupta and Gupta 2021b).

As highly dynamic systems, water quality estimation in rivers is a complex assignment but very significant in

order to generate pollution mitigation strategies (Gupta et al. 2022). Regular monitoring of the water quality of rivers, including the control and evaluation of physicochemical and/or biological parameters, became extremely important to protect this resource (Ustaoğlu et al. 2020). These parameters can be combined and transformed into simple dimensionless numbers, called water quality indices, that are effective in monitoring river pollution by transforming multifaceted water variable information into usable and intelligible data (Gupta et al. 2022).

Currently, there are general water quality indices and specific indices tailored to particular purposes (Almeida et al. 2012; del Corigliano 2008; Unda-Calvo et al. 2020). For example, the Simplified Index of Water Quality (SIWQ) combines five physicochemical parameters to determine water quality (Queral 1982). The numerical value of the SIWQ defines five levels of use of water according to its quality: all uses, swimming and fishing, irrigation and industry, forest irrigation, navigation, and very restricted use (Bustamante 1989; Bustamante et al. 2002, Losada Benavides et al. 2020).

A complete assessment of water quality requires extensive long-term data collection over space and time, generating huge data sets. Evaluation of spatiotemporal variation using statistical approaches is an effective method for monitoring the water quality of a river (Gupta et al. 2022). Multivariate statistical techniques, chemometric or environmetric methods (e.g., hierarchical cluster analysis (HCA), principal component analysis (PCA), factor analysis (FA), multivariate analysis of variance (MANOVA), correlation matrices (CM), regression analysis (RA), artificial intelligence (AI), etc.) are the most widely used ones in the analysis and classification of water qualities (Gupta et al. 2022; Fletcher et al. 2013). They also allow the interpretation of complex data sets and the understanding of temporal and spatial variations (Gatica et al. 2012; Ouali et al. 2009; Unda-Calvo et al. 2020), which can help to visualize possible water pollution sources from raw analytical data and reduce the subjectivity in the use of water quality indexes (S. Gupta and Gupta 2021a). The multivariate exploratory statistical techniques (PCA, HCA, etc.) identify the natural clustering pattern and associate variables based on similarities between samples (Kannel et al. 2007). RA also allows the developing prediction models, classification models, and time series for monitoring river water quality (Gupta and Gupta 2021a). Multiple linear regression (MLR) allows for analyzing the relationship between variables, providing more user interaction and control over predictive analytics compared to other prediction models such as machine learning models (Yildiz et al. 2017).

MLR is the most widely used statistical technique to estimate relationships between dependent and independent variables in regression modeling. It is used considering

two main objectives: to interpret data and to predict future response values (Etemadi and Khashei 2021).

These models make it possible to accelerate and lower the costs of water quality evaluations (Ewaid et al. 2018), considering the expensive task of determining numerous parameters to monitor rivers. To develop a model, it is necessary to carry out a comprehensive study of the system to select, using statistical tools, those parameters indicative of contamination (Valentini et al. 2021). Regression models have been successfully used in recent years for modeling hydrological processes (S. Gupta and Gupta 2021b; Gupta et al. 2022).

Problem statement

Currently, there are numerous worldwide rivers affected by anthropogenic activities, and, therefore, researchers are dedicated to studying them through quality indices and multivariate statistics (Alvareda et al. 2020; Barakat et al. 2016; Brilly et al. 2006; Bu et al. 2019; Carrasco et al. 2019; Connor et al. 2014; Dimri et al. 2021; Edokpayi et al. 2015; Fan et al. 2010; Hernandez-Ramirez et al. 2019; Howladar et al. 2021; Keupers and Willems 2017; Pinto and Maheshwari 2011; Valentini et al. 2021; Varol 2020). In Argentina, studies of urban rivers in large cities seek to establish their water quality and the degree of anthropogenic impact (Casares and De Cabo 2018; Bonansea et al. 2013; Merlo et al. 2011; Mgelwa et al. 2020; Nimptsch et al. 2005; Rautenberg et al. 2015; Valdés et al. 2021; Gatica et al. 2012; Lupi et al. 2019; del Corigliano 2008; Cazenave et al. 2009).

Similar studies have been conducted in the province of San Luis (Almeida et al. 2012; González et al. 2014), using statistics to evaluate the water quality of mountain rivers affected by tourism. The San Luis River is an urban river that flows through two developing cities, San Luis and Juana Koslay, and is affected by anthropic activities. Scarce research has been performed in order to evaluate the water quality of this river, with studies conducted over short periods of time and focused mainly on the industrial section of the river without developing a quality monitoring model (Castro et al. 2021). However, anthropic activities are developed not only in the industrial zone but also in the entire surrounding area and the riverbank itself, in both San Luis and Juana Koslay cities. The anthropic disturbances that affect this river encompass effluent discharges, runoff from animal husbandry, land removal through sand mining, uncontrolled riverbank vegetation management, and the accumulation of solid wastes (Borgatello 2014; Calderon et al. 2014; Ortiz 2017; Castro et al. 2021). Moreover, in recent years, the population of San Luis City has significantly grown, resulting in the discharge of inadequately treated wastewater into the San Luis River, as reported by the municipality (Giorda 2021). Simultaneously, the water of this river infiltrates

and replenishes underground aquifers that eventually flow into the Salinas del Bebedero. In this area, there is a salt exploitation company that provides a significant portion of Argentina with this compound. The water quality of this river is relevant not only from an ecological point of view but also from social, cultural, and economic approaches, which makes its monitoring relevant.

Considering the context outlined above, this study has three primary objectives:

- 1) To estimate the SIWQ index and classify the quality of river water for different uses.
- 2) To assess the spatiotemporal variations of water quality of the San Luis River using multivariate statistical analysis.
- 3) To develop and evaluate a multiple linear regression model to monitor an urban river.

Material and methods

Study area

The study area comprises the San Luis River, an urban river located in the province of San Luis, Argentina. This river, also called the Chorrillos River, courses through two developing cities: Juana Koslay and San Luis, where significant urbanization has occurred over the past decade. Data obtained from the Instituto Nacional de Estadística y Censos show that the Department of Pueyrredón—an area encompassing San Luis, Juana Koslay, and La Punta cities and some other smaller localities—increased its population from 204,019 inhabitants in 2010 (Instituto Nacional de Estadística y Censos (INDEC) 2010; Población Provincial Por Localidades Años 1869-2010 n.d.) to 260,295 in 2022 (Instituto Nacional de Estadística y Censos (INDEC) 2023), which implies a population growth of 25%.

This river originates at the foot of the sierras of San Luis, specifically in Juana Koslay City (33° 16' 59,32" south latitude, 66° 14' 15,03" west longitude), at the confluence of the Las Chacras and Cuchi Corral streams. It receives contributions from smaller basins located on the southern flank of Sierra de los Venados, as well as from an unnamed stream that collects water from its eastern sector. The San Luis River is a third-order urban stream, and it is part of the Bebedero Basin. It meanders through Juana Koslay City and is impounded at Dique Chico. From there, it flows southwestward through the city of San Luis, encompassing the industrial zone, and ultimately reaches an area where effluents from the municipal sewage treatment plant are discharged. Subsequently, the river traverses flat terrain, and during periods of low flow, its volume decreases significantly, with infiltration into the subsurface occurring

approximately 15 km away from the city of San Luis. The San Luis River ultimately drains into the Bebedero Basin, approximately 42 km southwest of the capital city, near Salinas del Bebedero (Calderon et al. 2014). This body of water can be regarded as one of the most severely impacted fluvial systems in the province of San Luis.

Sampling sites and regime

Eleven sampling sites were selected considering the level of disturbance of the river along its course through the cities of Juana Koslay and San Luis. The first two sites, represented by JA_w and JB_w, were considered the lowest disturbance level sites since they were located at a short distance from the river origin in the city of Juana Koslay. The next six sites, represented by JC_w, JD_w, JE_w, JF_w, JG_w, and JPO_w, are located in San Luis City, and the last three sites, JZ1_w, JZ2_w, and JZ3_w, are located near the end of the river course, prior to infiltration. These sites are related to different anthropogenic impacts: rainwater (JE_w), truck washes (JG_w) or water waste (JPO_w) discharges, recreational purposes (JD_w, JE_w, and JF_w), river sand mining (JC_w and JG_w), or animal husbandry (JZ1_w, JZ2_w and JZ3_w). Figure 1 shows the sampling sites that were monitored, and Table 1 shows the location and description of studied sampling sites. Selected sites represented a gradient of urbanization and human disturbance, ranging from areas with minimal disturbance—where the environmental baseline was determined—to sites with increasing levels of human impact.

The climate of the region is continental, with an annual rainfall regime of 595.4 mm and a monthly average of 49.7 mm (Ledesma and Arrellano 2022). Sampling was conducted during the 2015–2017 period based on rainfall measurements that defined dry and wet seasons. The period from May to October was defined as the dry season, and the period from November to April was the wet season. A total of 80 samples were collected.

Water sample analysis

Water samples were collected, preserved, transported, and analyzed according to Standard Methods for Water and Wastewater (APHA 2017). Parameters such as pH, conductivity (CONDUCT.), total dissolved solids (TDS, SM 45040), temperature (T), and dissolved oxygen (DO) were measured in situ with portable equipment. Other five parameters were determined in the laboratory: turbidity (TURBIDITY) (S.M. 2130-B); total suspended solids (TSS, S.M. S.M. 2540-F); organic matter (OM) (DQO S.M.5220-B); phosphorus (P, S.M. 4500-PO₄³⁻E); NITRATE (S.M.-4500-NO₃⁻E). The parameters were expressed in milligrams per liter, except conductivity ($\mu\text{S cm}^{-1}$), pH, and turbidity (NTU).

Simplified index of water quality

SIWQ is obtained from a simple formula that combines five physicochemical parameters: OM, TSS, DO, conductivity, and T to provide a quick and intuitive idea of water quality. SIWQ varies between 0 (minimum water quality)

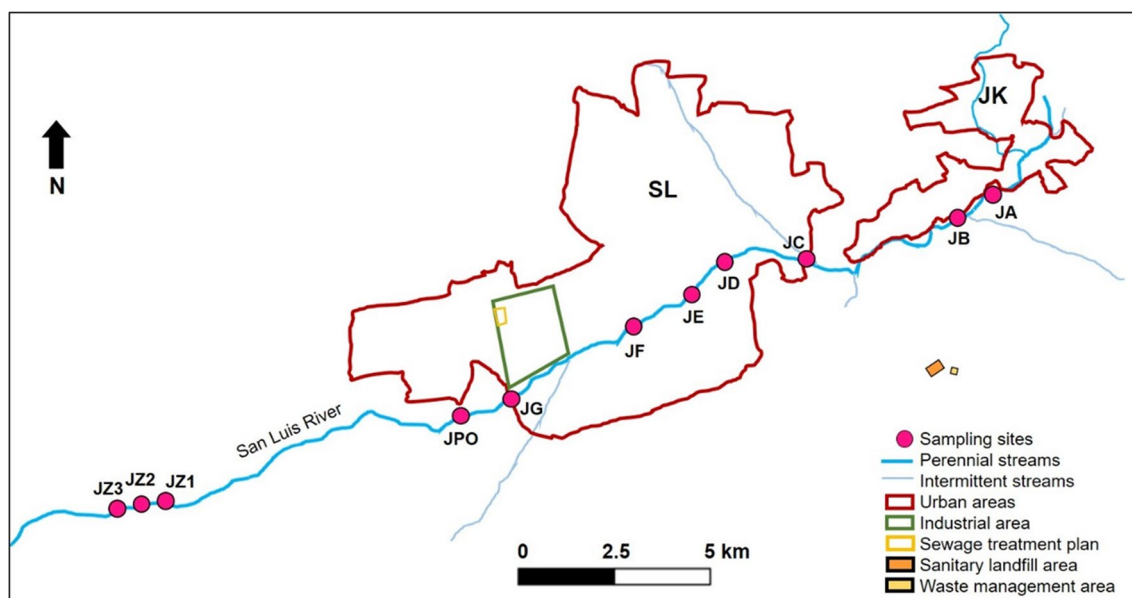


Fig. 1 Sampling sites along the San Luis River. SL, San Luis City; JK, Juana Koslay City

Table 1 Location and description of studied sampling sites from the San Luis River

Sampling Sites	Location	Description
JA_w	Juana Koslay City: near-river origin	Site close to houses under construction, new neighborhoods
JB_w	Juana Koslay city	Hard to access the site due to vegetation. A neighborhood and a school are at 100 m approx.
JC_w	San Luis–Juana Koslay cities' transition zone	Aggregate extraction quarry. Located before <i>Dique Chico</i>
JD_w	San Luis City	Riverside recreation area. It is located in front of a park cemetery
JE_w	San Luis City	Located 100 m from the city shopping mall and 50 m from a sports center. Recreation area. River drain discharge
JF_w	San Luis City	Access to the urbanized southern area. Zone of difficult access to the shore. Hardly disturbed
JG_w	San Luis City	Near to a truck wash. The site behind an aggregate quarry
JPO_w	San Luis City	Discharge of the oxidation pools of the municipality of San Luis City. Close to the City Racetrack
JZ1_w	Vicinity of San Luis City	The site is located a few kilometers from the municipal wastewater discharge, the final stretch of the river. Agricultural-livestock production area
JZ2_w	Vicinity of San Luis City	Area of agricultural-livestock production before river infiltration
JZ3_w	Vicinity of San Luis City	Fork of the river in its last section

and 100 (maximum water quality) and was proposed by Queralt (1982). SIWQ can be calculated according to the equation: $SIWQ = E(A + B + C + D)$; where E is a factor that depends on temperature and its values range from 0.8 to 1, A is a factor that depends on OM, its values range from 0 to 30, B is a factor that depends on TSS, its values range from 0 to 25, C is a factor that depends on DO, its values range from 0 to 25, and D is a factor that depends on conductivity, its values range from 0 to 20 (Queralt 1982). Other authors (Alonso Duré 2013; I. Bustamante et al. 2002; Bustamante 1989; Losada Benavides et al. 2020) modified the order and number of categories, and currently, water quality can be classified into five categories according to its potential uses: (a) $0 \leq SIWQ \leq 30$ Navigation, very restricted use; (b) $30 \leq SIWQ \leq 45$ water only suitable for forest irrigation; (c) $45 \leq SIWQ \leq 60$ water suitable for irrigation and industry; (d) $60 \leq SIWQ \leq 85$ water suitable for swimming and fishing; and (e) $85 \leq SIWQ \leq 100$ is considered all-purpose water (Losada Benavides et al. 2020). A SIWQ of 60 is the minimum acceptable value for the water quality of a river (López Fernández et al. 1998).

Statistical analysis

Exploratory/descriptive multivariate statistical analysis

Exploratory multivariate analyses were conducted with the opensource software R (version 4.1.0). Multivariate statistics were used to analyze environmental data (Aldás and Uriel 2017; Dormann 2020) and compare the spatial and temporal values of the studied physical–chemical variables.

Principal component analysis

PCA (Aldás and Uriel 2017; Dormann 2020) was performed aiming to extract significant components, reducing the contribution of variables with lesser significance. This technique reduces information provided by the variables to their essential features and transforms them into components (linear combinations of the original variables) that represent the proportion of variance explained by all variables (Greenacre et al. 2022).

Hierarchical cluster analysis

This method helps to identify natural associations in the dataset of river water quality and uses dendrograms to represent the similarity pattern between sampling sites (Aldás and Uriel 2017; Dormann 2020; Gupta et al. 2022). This descriptive statistical technique was applied by using Ward's method (Ward 1963; Barakat et al. 2016) and the Manhattan distance technique in order to maximize homogeneity within the groups. Ward's method considers the increase in the squared error as the proximity between two clusters and is the most common method to categorize groups. A dendrogram represents the clusters and their proximity with a reduction in the dimensionality of the original data (Barakat et al. 2016).

Heatmaps

Heatmaps were also used to simultaneously visualize groups of samples and their characteristics. A heatmap is another way to show hierarchical clustering, in which data values are transformed into a color scale. Dendrograms were combined

with a heatmap to allow visual identification of possible characteristic patterns within each cluster.

Inferential statistical analysis

Correlation matrix

Inferential statistical techniques and individual and multiple correlation analyses were performed to explore the relationships between physical–chemical and elemental variables, resulting in correlation matrices. Pearson's correlation coefficient (r) was employed to measure the extent of linear association between the parameters, which is useful for assessing the degree of dependence of one variable on others (Singh et al. 2020). Independent variables were selected by eliminating those that showed a strong correlation with each other (collinearity).

Multivariate analysis of variance

The MANOVA technique was carried out for an inferential study. However, not all parameters met the normality and homoscedasticity conditions for the application of MANOVA. Thus, the MANOVA.WIDE function in R was employed. This more robust technique calculates Wald-type statistics (WTS) and modified statistic ANOVA type (MATS), which are versions of these statistical tests designed for semiparametric multivariate data (Aldás and Uriel 2017; Dormann 2020).

Multiple linear regression models

Finally, multiple linear regression models (MLRM) (Harrel 2015) were applied in order to study the potential relationship between a dependent variable and a series of explanatory variables. The multivariate regression model can be expressed using the equation: $Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_m X_m$, where Y is the dependent variable, X_1, X_2, \dots, X_m are the independent variables, predictor variables, or regressors, $\beta_1, \beta_2, \dots, \beta_m$ are the coefficients of the model, and m is the number of independent parameters considered in the regression.

The selection of the regressor variables was carried out by progressive elimination, until a model with an adjusted coefficient of determination (R^2) greater than 50%, and a confidence level of 90% or 95% was obtained. R^2 determines the percentage of the variance of the dependent variable explained by the regression model. The corresponding residual tests were carried out: the Durbin–Watson test was used to measure independence, the Shapiro–Wilk test was applied to measure normality, the Breusch–Pagan test was used to measure homoscedasticity, and the Variance Inflation Factor (VIF) test was applied. The Grubbs test was applied to detect

outliers. The results of these tests were verified by graphs like the normal Q–Q plot and histogram to assess normality, residuals versus fitted values to test homoscedasticity, box-plots, and distance of Cook to detect outliers and influential values in the model. Standard residual error and adjusted R^2 were also taken into account to determine an optimal fit.

The accuracy of the model was tested using cross-validation methods, specifically the Validation Set Approach. This method evaluates the ability of the model to predict the outcome of new unseen observations not used to build the model. The validation set approach consists of dividing the data into two sets: the training set, used to train or build the model, and the testing or validation set, used to test the model by estimating the prediction error (new data from SLR and data from different rivers with similar characteristics). Subsequently, the prediction error is quantified as the mean squared difference between the observed and the predicted outcome value (Kassambara 2018). The statistics Root Mean Square Error (RMSE), R^2 , and error rate were used to predict error. RMSE measures the average error of the model in predicting the outcome of an observation and is selected because it is an absolute measure of goodness of fit; the lower the RMSE values, the better the model. R^2 was chosen based on the fact that it is a relative measure of the fit of the model to the dependent variables. It represents the correlation between the observed values and the predicted values. Higher R^2 values indicate a better model (James et al. 2017). The learning curves of the model were plotted as log RMSE vs. iterations. Log RMSE was used for a better appreciation of the differences between both curves (Viering and Loog 2021).

Results and discussion

Water samples and SIWQ

As described before, water samples were obtained from the San Luis River, and ten physicochemical variables were determined on each sample. The average values of all variables determined are shown in Table 2.

Average values of OM, TSS, DO, conductivity, and T were used to calculate an average SIWQ in order to classify water quality at each sampling site. Table 3 displays the obtained SIWQ values and the classification of water samples from each studied site.

SIWQ values indicate a greater deterioration in water quality at sites JPO_w, JZ1_w, JZ_w, and JZ3_w. Water quality degradation is mainly affected by variables related to the discharge of municipal effluents from the city of San Luis (OM, conductivity, P, and nitrate). Additionally, medium-quality site JG_w could be affected by specific anthropogenic activities such as sand mining, truck wash

Table 2 Average and standard deviation values of physicochemical variables determined in water samples from sites of the San Luis River

Site	pH	T (°C)	Conductivity (µS cm ⁻¹)	OM (mg L ⁻¹)	DO (mg L ⁻¹)	Turbidity (NTU)	Nitrate (mg L ⁻¹)	P (mg L ⁻¹)	TSS (mg L ⁻¹)	TDS (mg L ⁻¹)
JA_w	7.83 ± 0.50	16.6 ± 5.2	628.80 ± 187.86	3.36 ± 1.25	8.46 ± 1.10	12.79 ± 8.25	1.10 ± 0.51	0.042 ± 0.023	33.51 ± 21.61	422.30 ± 127.00
JB_w	7.77 ± 0.57	16.5 ± 6.1	592.25 ± 160.16	3.45 ± 1.15	8.69 ± 0.60	15.71 ± 10.58	1.08 ± 0.28	0.043 ± 0.037	41.16 ± 27.72	397.75 ± 108.44
JC_w ^a	7.71 ± 0.45	16.9 ± 5.0	795.90 ± 268.80	4.40 ± 3.52	8.52 ± 2.10	180.91 ± 322.07	1.72 ± 1.69	0.073 ± 0.105	473.97 ± 843.82	533.50 ± 192.94
JD_w	7.94 ± 0.37	16.7 ± 4.5	754.40 ± 231.21	3.17 ± 1.30	8.58 ± 0.88	28.99 ± 29.68	1.38 ± 0.75	0.044 ± 0.035	75.96 ± 77.77	506.10 ± 155.36
JE_w	7.87 ± 0.39	17.7 ± 4.7	735.27 ± 218.9	2.93 ± 1.47	8.11 ± 1.23	32.34 ± 21.43	1.19 ± 0.65	0.033 ± 0.030	84.74 ± 56.15	491.29 ± 147.47
JF_w	7.90 ± 0.30	17.7 ± 5.3	699.25 ± 224.31	2.86 ± 1.65	8.61 ± 0.95	25.05 ± 13.80	1.36 ± 0.79	0.035 ± 0.028	65.63 ± 36.17	469.13 ± 149.94
JG_w ^a	7.94 ± 0.68	17.7 ± 5.2	702.80 ± 193.13	3.93 ± 1.81	8.33 ± 12.68	184.47 ± 315.45	1.44 ± 0.98	0.064 ± 0.092	483.31 ± 826.49	476.00 ± 128.25
JPO_w	6.68 ± 0.30	22.2 ± 5.9	861.50 ± 258.09	40.75 ± 10.82	4.80 ± 0.57	61.00 ± 52.33	15.00 ± 3.88	4.360 ± 0.651	159.82 ± 137.09	581.00 ± 171.12
JZ1_w	7.24 ± 0.04	21.0 ± 4.2	1095.00 ± 66.46	32.65 ± 10.11	2.98 ± 0.32	31.95 ± 14.21	2.55 ± 1.63	4.00 ± 2.546	83.71 ± 37.24	734.50 ± 43.13
JZ2_w	7.03 ± 0.08	20.0 ± 2.8	1126.00 ± 42.43	28.15 ± 11.95	3.02 ± 0.54	32.30 ± 17.96	3.20 ± 1.84	4.600 ± 3.677	84.63 ± 47.06	754.50 ± 28.99
JZ3_w	7.13 ± 0.03	20.0 ± 1.4	1128.00 ± 14.14	27.05 ± 3.75	3.92 ± 0.25	29.70 ± 11.74	2.50 ± 1.84	3.450 ± 2.051	77.81 ± 30.75	732.00 ± 43.84

^aThe values of P, turbidity, and TSS in sites JC_w and JG_w have a high dispersion with respect to the mean due to the wide range of variability of the data determined at these sites

Table 3 Average SIWQ determined and classification of water for each studied sampling site of the San Luis River

Site	SIWQ	Quality	Recommended use
JA_w	78.8	Good	Swimming and fishing
JB_w	79.8	Good	Swimming and fishing
JC_w	60.7	Good	Swimming and fishing
JD_w	71.1	Good	Swimming and fishing
JE_w	70.8	Good	Swimming and fishing
JF_w	75.3	Good	Swimming and fishing
JG_w	59.6	Medium	Irrigation and industry
JPO_w	18.2	Very bad	Navigation, very restricted use
JZ1_w	17.6	Very bad	Navigation, very restricted use
JZ2_w	19.2	Very bad	Navigation, very restricted use
JZ3_w	21.9	Very bad	Navigation, very restricted use

discharges, and runoff (Borgatello 2014; Castro et al. 2021). The average SIWQ value for the JC_w site is barely higher than the medium quality limit, meaning that a small variation of any physicochemical variable can affect its water quality.

Considering the information provided by the SIWQ index and the recreational use of the San Luis River in some sections, it would be advisable to conduct further studies by applying the Recreational Water Quality Index (Almeida et al. 2012).

Physicochemical variables determined in water samples were analyzed using multivariate statistics to reduce the number of significant variables and develop a faster and more cost-effective predictive model for monitoring river water quality. Correlation matrix and exploratory analysis were performed using average data (Table 2), whereas inferential analysis was performed using the entire dataset (data not shown), not just average data.

Physicochemical variables correlation

A correlation matrix was calculated to explore associations among the different physicochemical variables in the dataset under study (Fig. 2). There was a positive and statistically significant correlation between the parameters P and OM (0.98), conductivity and P (0.89), and P and TDS (0.90) with *p* values < 0.001, between pH and DO (0.84), OM and TDS (0.83), and OM and conductivity (0.82) with *p* values < 0.01, between OM and nitrate (0.72) with *p* value < 0.05, and finally, between P and nitrate (0.60) with *p* values < 0.1. Figure 2 also depicts a negative and statistically significant correlation between the parameters DO and OM (−0.92), P and DO (−0.97), DO and conductivity (−0.95), MO and pH (−0.96), pH and P (−0.93), and OD and TDS (−0.95) all with *p* values < 0.001, and between pH and nitrate (−0.78), pH and conductivity (−0.74), and

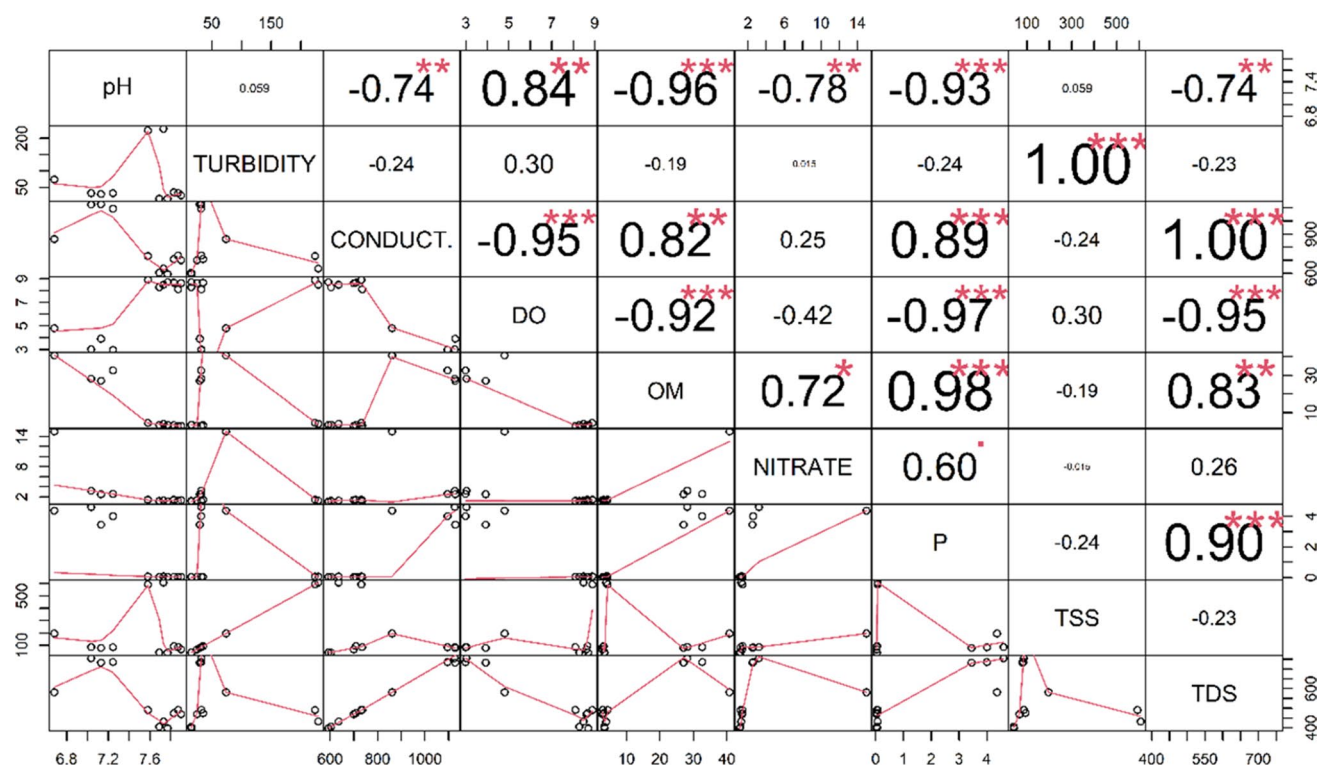


Fig. 2 Correlation matrix of physicochemical variables of water samples from the San Luis River: scatterplots, values, and statistical significance of each correlation coefficient (Pearson correlation coefficients [*r*]). Signification codes: 0 ***0.001; **0.01; *0.05; “.” 0.1; “ ’ ” 1

pH and TDS (−0.74) both with *p* values < 0.01. A correlation coefficient of 1 was observed between the variable conductivity and TDS, as well as between turbidity and SST.

Spatial and temporal similarity and site grouping

PCA was then applied using average data to associate the parameters that characterized the groups (Fig. 3).

The PCA resulted in the selection of two principal components, which represent 88.1% of the variability data. PC1 (Dim1) explains 72% of the variability. The linear combination of the original parameters that originated this component is:

$$PC1 = -0.4243 pH - 0.1099 Turbidity + 0.3880 Conductivity - 0.4251 DO + 0.4407 OM + 0.2940 Nitrate + 0.4427 P$$

Data show that conductivity, OM, P, DO, and pH are the variables with the greatest contribution to PC1. Figure 3 shows conductivity, OM, and P present a positive correlation with each other, with a higher correlation between OM and P, while DO and pH present a positive correlation with each other and a negative correlation with the parameters conductivity, OM, and P.

The second component, PC2 (Dim2), represents 16.1% of the variability of the data and is the result of the following linear combination of the original parameters:

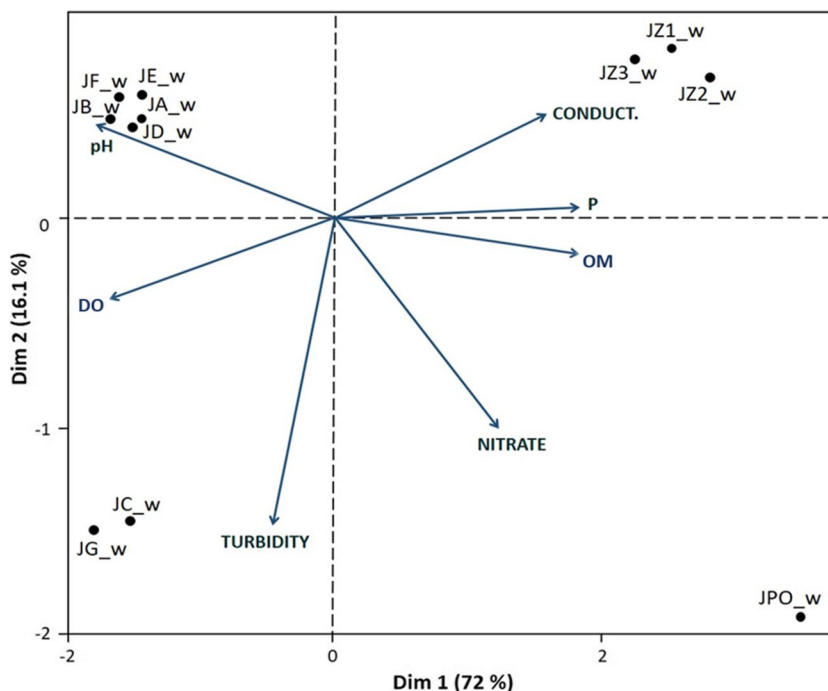
$$PC2 = 0.2388 pH - 0.7527 Turbidity + 0.2687 Conductivity - 0.1995 DO - 0.0846 OM - 0.5100 Nitrate + 0.0320 P$$

Turbidity and nitrate contribute significantly to this component. As indicated by the length of the vectors representing these parameters in Fig. 3, turbidity and nitrate exhibit a greater variability compared to the other variables. The parameters that contribute to PC1 are related to water pollution resulting from the organic load, while those contributing to PC2 are linked to pollution originating from land use in areas near the river (Rentier and Cammeraat 2022), with turbidity being the parameter with a major contribution to PC2.

The grouping of sites in Fig. 3 reveals that post-discharge sites (JZ1_w, JZ2_w, and JZ3_w) are characterized by OM, P, and conductivity. Conversely, JA_w, JB_w, JD_w, JE_w, and JF_w sites are associated with DO and pH, indicating a stronger connection to PC1. The discharge site JPO_w is related to the nitrate variable, while the JC_w and JG_w sites are linked to the turbidity parameter, representing PC2.

As part of the initial exploratory analysis, a dendrogram was created to determine site grouping. We used the Ward

Fig. 3 Biplot showing the projections of the variables in the first two PCs and the distribution of the sampling sites from the San Luis River



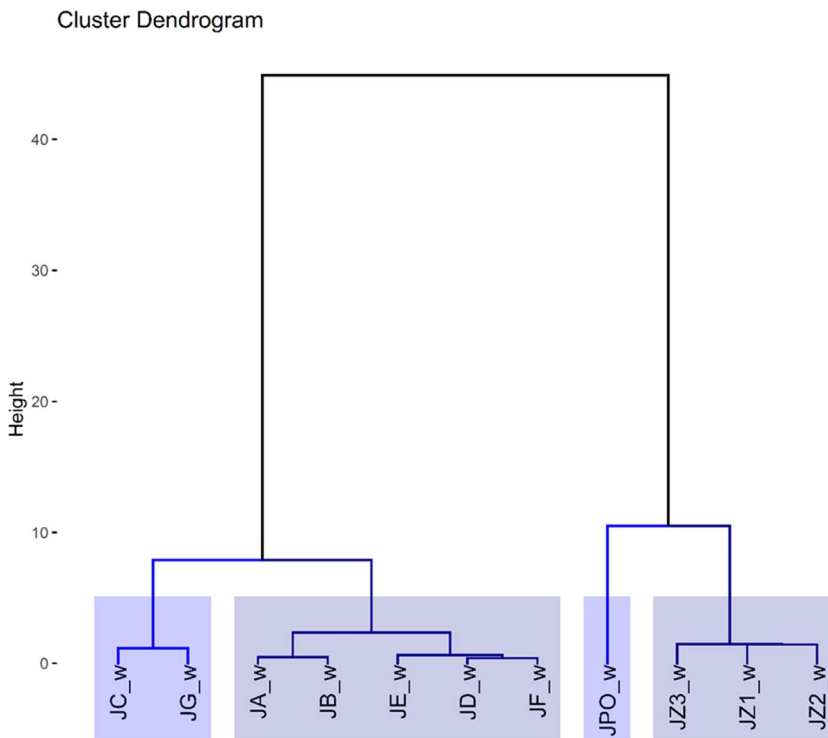
method with the Manhattan distance technique to produce the dendrogram shown in Fig. 4.

Figure 4 shows that all monitoring sites were classified according to four clusters: the first group includes JC_w and JG_w sites and is related to the second group, which consists of JA_w, JB_w, JE_w, JD_w, and JF_w sites. JPO_w site,

the only one in the third group, is related to the fourth group, which consists of JZ1_w, JZ2_w, and JZ3_w sites.

The first and second clusters constitute the first branch of this dendrogram. Both clusters present similarities, although to a lesser degree. Sites grouped in the first cluster present similar characteristics due to the presence of comparable

Fig. 4 Dendrogram based on hierarchical clustering (Ward’s method) for all studied sites in the San Luis River



anthropogenic disturbances in the upstream sampling sites, such as sand mining, truck washes, and household sewage effluent discharges. In the second cluster, JA_w and JB_w sites, located in the city of Juana Koslay, show greater similarity with each other rather than JE_w, JD_w, and JF_w sites located in a river segment that flows through a highly urbanized area of the city of San Luis.

The third and fourth clusters constitute the second branch of this dendrogram. The JPO_w site corresponds to an effluent discharge and integrates a conglomerate by itself. However, the municipal effluent discharge site JPO_w and sites downstream, JZ1_w, JZ2_w, and JZ3_w, exhibit a great similarity, suggesting that the river cannot naturally recover from the impact of organic pollution.

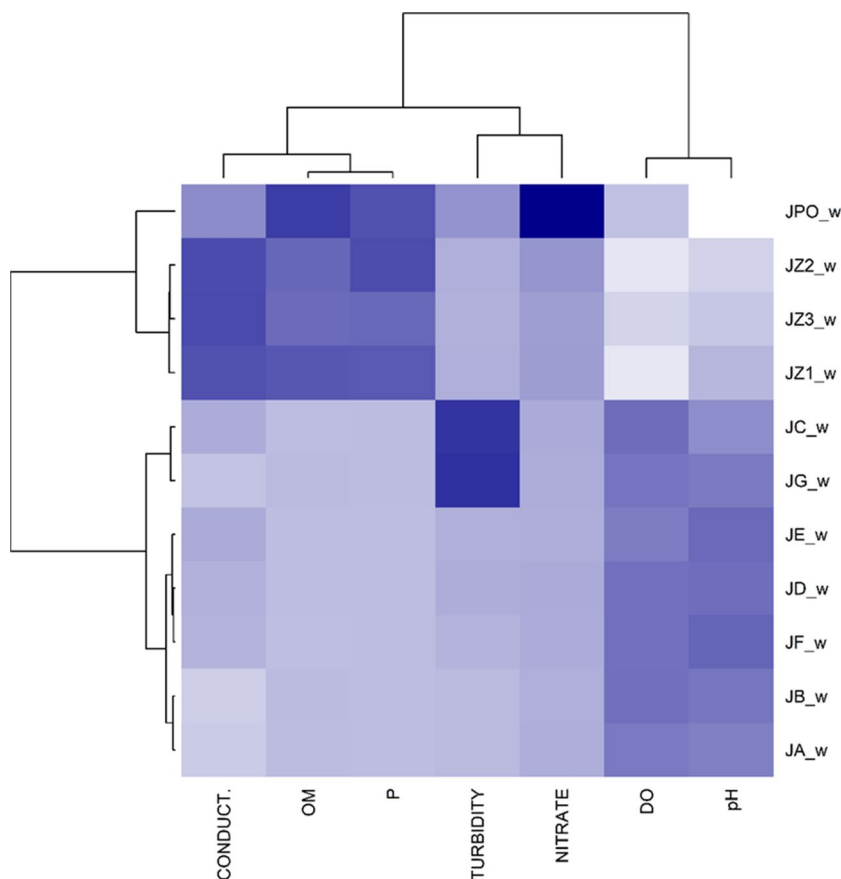
The clustering outcomes were illustrated using both a dendrogram tree and a heatmap. The heatmap technique combines information from dendrograms obtained for sites and variables. Results performed using average data are shown in Fig. 5, allowed us to determine which parameters characterized or gave rise to the clusters obtained in the dendrogram technique.

Figure 5 shows that JA_w and JB_w, reference sites with low-moderate anthropogenic activity, have nearly all the analyzed parameter values lower or equal to those of sites upstream of the municipal effluent discharge (JC_w,

JD_w, JE_w, JF_w, and JG_w). Sites JC_w and JG_w exhibit higher turbidity values compared to other sites located prior to the aforementioned effluent discharge. This can be attributed to activities related to sand mining (Da and Le Billon 2022; Rentier and Cammeraat 2022) that exist in areas upstream. The exploitation of sand has increased in the area due to the high demand for this resource, primarily driven by its various uses, especially in construction (Da and Le Billon 2022). This increased demand can be attributed to the growth in population (Kemgang Lekomo et al. 2021). The continued resuspension of sediments resulting from in-stream sand removal affects local and downstream water quality. The presence of a certain amount of suspended solids is beneficial for the river system since suspended sediment and nutrients contribute to the ecosystem of the river. However, a large concentration of suspended solids leads to increased turbidity in the main channel, negatively affecting the stream of the river. Furthermore, during sand mining, fuel spills and leaks from excavation machinery and transport vehicles can occur. These contaminants can be adsorbed by sediments and transported downstream (Juez and Franca 2022; Rentier and Cammeraat 2022).

Data show that land uses and anthropogenic activities (sand mining, effluent discharges, etc.) are related to water

Fig. 5 Heatmap with dendrogram tree represents physicochemical variables values among sampling sites in the San Luis River



quality, particularly with changes in variables such as turbidity (Lu et al. 2023; Nasrabadi et al. 2016; Obisesan and Christopher 2018; Yan et al. 2023), nutrient concentrations (P, nitrate, among others), etc. (Lu et al. 2023). Other authors who studied ponds receiving clandestine discharges (Fontanarrosa et al. 2023) proposed similar results.

The group of sites located downstream of the JPO_w site is associated with higher values of parameters like OM, conductivity, and P, inferring that the discharge of poorly treated municipal effluents (Giorda 2021) has an impact on water quality. Nitrate is the most relevant parameter at the JPO_w site. High concentrations of nitrate are mainly due to household effluents, according to various studies (Chilundo et al. 2008; Pisani et al. 2020; Schirmer et al. 2013; Sikakwe et al. 2020). Both urban areas and crops are important contributors of nitrogen and phosphorus in water bodies (Wang et al. 2023).

This same technique was used with average data for wet and dry season parameters separately (data not shown). We observed that the JPO_w site presented the highest concentration of nitrate compared to the other sites for both dry and wet seasons. P and MO variables exhibited higher concentrations for sites after JPO_w for both periods. In addition, the conductivity of sites downstream of JPO_w is higher in the wet season. This could be explained by considering land use in the area associated with livestock.

Inferential multivariate analysis

After conducting exploratory or descriptive multivariate analysis and inferring differences in site groupings and parameter influences, we proceeded with inferential multivariate analysis. This study was performed in order to define whether variations in the studied parameters for the sampled sites were statistically significant or not.

Inferential analysis (Dormann 2020) requires, as a necessary condition, to verify that the data have a normal distribution and homogeneity of variance. These requirements were verified using Royston's multivariate normality tests (Trujillo-Ortiz and Hernandez-Walls 2007), obtaining a p value of 1.17×10^{-36} . In addition, the Lilliefors (Kolmogorov–Smirnov) univariate normality test was applied for samples with $n > 50$, obtaining p values < 0.01 for conductivity, MO, OD, turbidity, nitrate, and P variables, and a p value of 0.39 for pH variable. The Cramer–von Mises test (Baringhaus and Henze 2017) was also applied, obtaining the same results and a p value of 0.226 for the variable pH. Considering this, we can conclude that only this variable exhibits a univariate normal distribution at the significance level of 0.01.

A multivariate homoscedasticity test was performed by applying Box's M -test for homogeneity of covariance

matrices, both for site and period, obtaining a p value $< 2.2 \times 10^{-16}$ for both cases. Considering p values are less than 0.05, they indicate there is no homogeneity in site or period.

The MANOVA.WIDE function from R software was used to calculate WTS and MATS. Both statistics reported a significant difference between period, site, and the site–period interaction, with a significance level of less than 0.001.

Multiple linear model to monitoring San Luis River

Based on the multivariate statistical analysis carried out, a model that allows monitoring the water quality of the San Luis River was proposed. This model, named the multiple linear model for monitoring San Luis River (MRSL), was developed considering OM as the dependent variable. Two reasons lead us to select this variable as a dependent one: on the one hand, it represents organic load; and on the other, this variable was clearly modified in river sections most affected by anthropogenic disturbances, as can be seen in PCA, heatmap, and inferential analysis.

The proposed multiple linear regression model evaluates the impact of each predictor (independent variable), considering the influence of all other predictors simultaneously. The categorical variable site was included in the model since it was significant. Values of categorical variable site are: 0 for site JA_w, 0.1429 for site JB_w, 0.9171 for site JC_w, -0.5170 for site JD_w, -0.4093 for site JE_w, -0.7866 , -0.7866 for site JF_w, 0.9135 for site JG_w, -0.9654 for site JPO_w, 10.6203 for site JZ1_w, 2.6424 for site JZ2_w, and 7.394 for site JZ3_w.

The achieved model is determined by the following equation:

$$OM = 1.631 + 1.47 \text{ Nitrate} + 4.204 \text{ P} - 0.005 \text{ Turbidity} + \text{Site}$$

The statistics of the multiple linear models obtained reveals that when all the physicochemical variables and the categorical variable site as predictors are included, the model can effectively account for 97.06% of the OM variance in the San Luis River. Furthermore, the F -test statistic of 178.8 indicates a high significance of the model, supported by a p value of 2.2×10^{-1} , which demonstrates a good association between the predictors and the OM variable. The physicochemical variables P, nitrate, turbidity, and the sites JZ1_w and JZ3_w have a statistically significant relationship with the response variable OM (p values < 0.001).

As evidenced by R^2 , the multiple linear regression model formulated by MRSL demonstrated that it effectively captured 97.06% of the overall variability of the variable explained by de-regression. The analysis of variance pertaining to the dependent variable OM (P: p value $< 2.2 \times 10^{-16}$, nitrate: p value $< 2.2 \times 10^{-16}$, turbidity: p value = 0.000596, and site: p value = 6.084×10^{-7}) indicated that

all physicochemical variables and the categorical variable site were statistically significant at the 0.01 level. Several authors have documented models that employ organic load as a basis for predicting parameters of water bodies, with the selection of dependent variables associated with organic load (Jiang et al. 2021) or parameters such as DO (Ahmed 2017; Ouma et al. 2020) or BOD (Vigiak et al. 2019), which serve as dependable indicators of water quality.

One noteworthy aspect of this work is that it helps bridge the gap in scientific data on the physicochemical parameters of the San Luis River, making this research a valuable study to initiate analysis of the area. The effort and costs related to sampling in order to obtain data from each site can be reduced by using this model capable of predicting parameter values such as OM.

The ability to predict OM is of great importance because it poses challenges for drinking water treatment plants. The main issues associated with OM include degradation of organoleptic quality, bacterial growth in the distribution network, and significant chlorine consumption during disinfection (LeChevallier 1990).

As shown in Fig. 6, the obtained model adheres to the necessary assumptions of normality, homoscedasticity, and independence of residuals. This is because it does not contain outliers, and there is no discernible pattern in the residual distribution. These assumptions were confirmed using the Kolmogorov–Smirnov test ($D = 0.51$, p value = 1), the studentized Breusch–Pagan test ($BP = 20.52$, $df = 13$, p value = 0.08), and the Durbin–Watson test ($DW = 1.86$, p value = 0.06). In all cases, the tests confirmed normality,

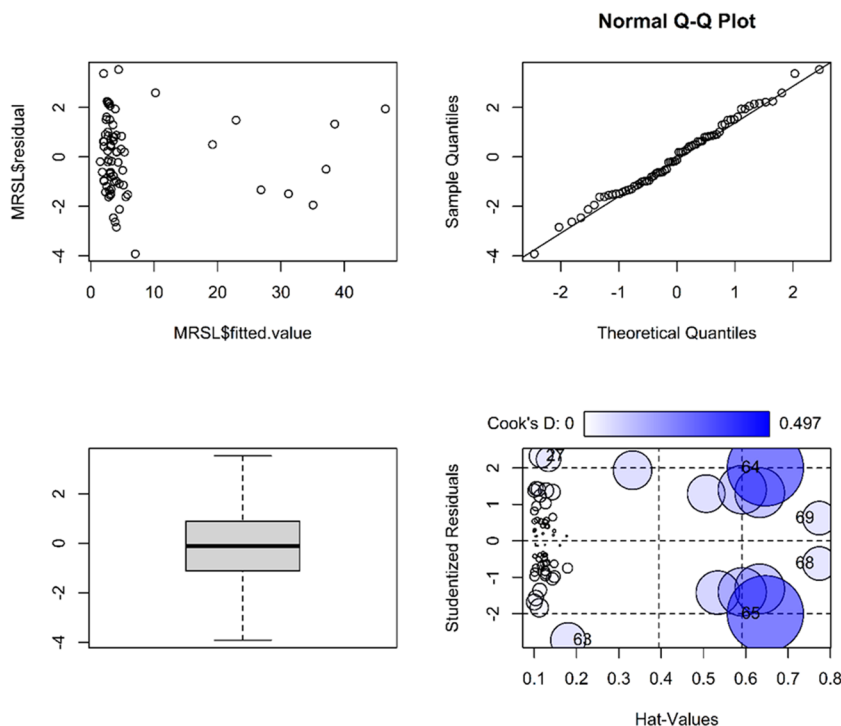
homogeneity of variance, and residual independence, with p values greater than the 0.01 significance level. The correlation matrix showed a correlation of 0.14 between nitrate and turbidity and a correlation of 0.50 between P and nitrate, confirming the absence of multicollinearity. The calculated VIF values were all below 10, providing further evidence against multicollinearity. Cook's distance values were also below 1.5, indicating no significant influence of the variables. Additionally, the Grubbs test did not detect any outliers (p value = 0.2845).

The MRSLS model obtained is parsimonious (few parameters, but it fits well, is easier to understand, and has strong predictive ability), and it satisfies all the necessary assumptions for this type of multiple linear regression. Table 3 shows the minimum and maximum values of each physicochemical variable that must be considered to apply the MRSLS model successfully.

The independent variables in the MRSLS model, nitrate, turbidity, and P, are associated with the anthropogenic activities discussed throughout the study. This suggests the model allows for the rapid identification of increases in the quantity or frequency of effluent discharge or activities related to soil removal impacting the river.

The MRSLS model fulfilled the assumptions of normality, homoscedasticity, linearity, and independence; hence, it was used to predict OM in different rivers of the province. The evaluated rivers were San Luis (SLRL2022, SLRPO2022, SLRA2021, SLRI2021, SLRL2021, SLRAG2019), Los Molles (LMRHP2019), Rosario (RRLT2022, RRLT2019), Conlara (CRC2023, CRSR2023, CRPG2023, CRPG2022,

Fig. 6 Verification of residuals homoscedasticity and independence (top left), residuals normality: Q–Q plot (top right), residual boxplot (bottom left), and influence plot (bottom right)



CRPG2021, CRPG2019), Quines (QR2022B, QR2022A, QR2019, QR 2021), Nogolí (NR2022b NR2022a), and San Francisco (SFR2021, SFR2022). For this purpose, river sites with similar characteristics to those studied in this work were selected. The physicochemical parameters experimentally determined at each site were within the range stipulated in Table 4. OM values estimated using the model were compared to OM determined experimentally (Fig. 7). The results in the prediction of OM values were satisfactory.

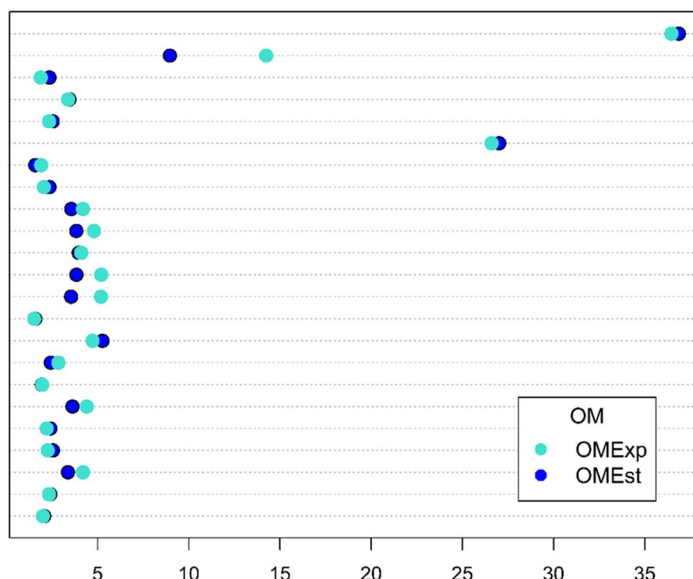
Normality and homoscedasticity of the data were evaluated in order to verify if the difference between the experimentally determined OM values (OMExp) and the model-estimated OM values (OMEst) (Fig. 7) was significant. Shapiro–Wilks test was used to test normality, obtaining a p value = 4.322×10^{-08} for estimated OM and a p value = 1.655×10^{-07} for the experimentally determined OM, indicating the data were not normal. Homoscedasticity was evaluated using the F -test, obtaining a p value = 0.995. This p value indicates the data are homoscedastic. Although the data present homogeneity of variances when graphing the boxplots (Fig. 8), the presence of outliers is observed.

Table 4 Statistics of each physicochemical variable used to develop the MRSL model

Statistics	OM (mg L ⁻¹)	Turbidity (NTU)	Nitrate (mg L ⁻¹)	P (mg L ⁻¹)
Min.	0.90	1.21	0.20	0.002
1st Qu.	2.45	11.55	0.80	0.019
Median	3.70	21.90	1.10	0.037
Mean	6.72	69.49	1.84	0.506
3rd Qu.	4.90	42.65	1.65	0.075
Max.	48.40	1050.00	17.70	7.200

Fig. 7 Dotchart comparing experimental values of OM (turquoise) and estimated values of OM using the MRSL model (blue) for sites located at different rivers in the Province of San Luis

- SLRPO2021
- SLRA2021
- SLRI2021
- SLRL2022
- SLRL2021
- SLRAG2019
- LMRSH2023
- RRLT2022
- RRLT2019
- CRC2023
- CRSR2023
- CRPG2023
- CRPG2022
- CRPG2021
- CRPG2019
- QR2022b
- QR2022a
- QR2021
- QR2019
- NR2022b
- NR2022a
- SFR2022
- SFR2021



The non-normal distribution and the presence of atypical data (outliers) required the application of a robust test to detect differences between the OMExp and OMEst groups. The Yuen test (Mair and Wilcox 2020) was used for paired samples, obtaining a p value = 0.179. A p value > 0.05 indicates that there are no significant differences between the trimmed means of OMExp and OMEst (trimmed means difference: -0.32; 95% confidence interval between -0.7955 and 0.1555). Based on this, we can conclude that the proposed multiple linear models satisfactorily predicts OM, not only for samples from the San Luis River but also for samples from different rivers with similar characteristics.

The accuracy of the model was tested using the cross-validation (set approach method). Two different train and test datasets were selected, and R^2 , RMSE, and error rate were

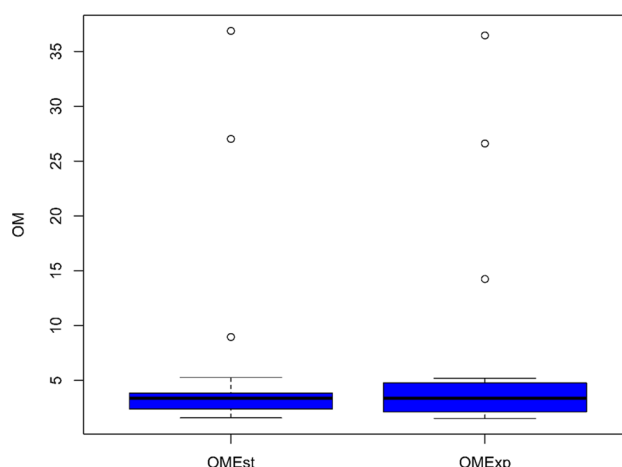


Fig. 8 Box plots comparing experimental values of OM and estimated values of OM using the MRSL model for different rivers in San Luis Province

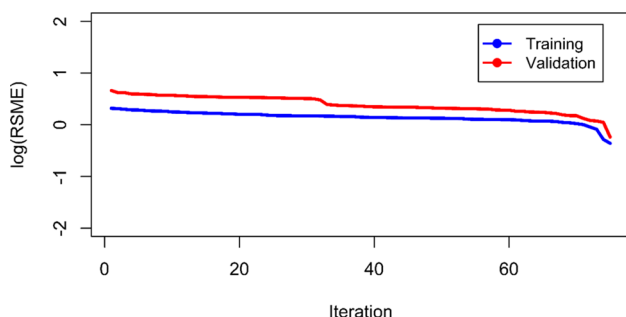


Fig. 9 Learning curves of MRSL model (logRMSE vs. iteration).

determined after 75 iterations. The values of these statistics were $R^2 = 0.999$, $RMSE = 0.44$, and error rate = 3.9% for the training dataset, and $R^2 = 0.99$, $RMSE = 0.584$, and error rate = 5.11% for the test dataset. The learning curves of the training and validation datasets (Fig. 9) indicate the model is neither under-fit nor over-fit. These statistics exhibit the accurate prediction power of the model for this urban river since an error of 0.58 mg L^{-1} of OM in a range of values from 2 to 50 mg L^{-1} is acceptable.

Conclusions

In this study, it was evidenced through the SIWQ that the quality of the water in the San Luis River is affected by anthropogenic activities. Furthermore, the application of multivariate statistics made it possible to detect significant differences in the temporal and spatial variation of the variables under study.

The multiple linear regression model developed demonstrates robustness in predicting and monitoring the water quality of the San Luis River and different rivers in semi-arid areas with similar characteristics affected by organic load pollution. Results provide reference information for authorities responsible for the environmental management of the San Luis River regarding how anthropogenic activities in the river and its banks affect water quality.

Acknowledgements We would like to express our gratitude to Graciela Lucero Arrúa from the Instituto de Lenguas (ILen - Universidad Nacional de San Luis) for revising the manuscript.

Author contribution Jessica Alejandra Tello: conceptualization, methodology, investigation, writing—original draft, and writing—review and editing. Jorge Leandro Leporati: data curation, visualization, and writing—review. Patricia Laura Colombetti: investigation and writing—review. Cynthia Gabriela Ortiz: investigation and writing—review. Mariana Beatriz Jofré: methodology and writing—review and editing. Gabriela Verónica Ferrari: investigation and writing—review and editing. Patricia González: conceptualization, methodology, writing—review and editing, and validation.

Funding The research leading to these results received funding from the Secretaría de Ciencia y Técnica, Universidad Nacional de San Luis, through grants to Projects PROICO UNSL N° 2-1914. “Calidad ambiental de Ecosistemas acuáticos del área de los humedales desaguadero - Salinas del Bebedero, Provincia de San Luis” (Resolución CS N°: 48/15), and PROICO UNSL N° 2-2418 “Calidad ambiental de ecosistemas acuáticos metodologías analíticas para la determinación de compuestos de interés ambiental” (Resolución CS N°: 126/18), and by the Instituto de Química San Luis (INQUISAL), Consejo Nacional de Investigaciones Científicas Y Técnicas (CONICET).

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author upon reasonable request.

Declarations

Ethical approval Not applicable.

Consent to participate Not applicable.

Consent to publish Not applicable.

Conflict of interest The authors declare no conflict of interest.

References

- Ahmed AAM (2017) Prediction of dissolved oxygen in Surma River by biochemical oxygen demand and chemical oxygen demand using the artificial neural networks (ANNs). *J King Saud Univ Eng Sci* 29(2):151–158. <https://doi.org/10.1016/j.jksues.2014.05.001>
- Aldás J, Uriel E (2017) Multivariate analysis applied with R, 2nd edn. Paraninfo S.A., Madrid (in Spanish)
- Almeida C, González SO, Mallea M, González P (2012) A recreational water quality index using chemical, physical and microbiological parameters. *Environ Sci Pollut Res* 19(8):3400–3411. <https://doi.org/10.1007/s11356-012-0865-5>
- Alonso Duré JA (2013) Quality evaluation of the waters of the Agua-pey stream (Paraguay) using macroinvertebrates as bioindicators. Universidad Nacional de Itapúa (In Spanish). <https://www.conacyt.gov.py/sites/default/files/ITES-BN-025.pdf>
- Alvareda E, Lucas C, Paradiso M, Piperno A, Gamazo P, Erasan V, Russo P, Saracho A, Banega R, Sapriza G, de Mello FT (2020) Correction to: water quality evaluation of two urban streams in Northwest Uruguay: are national regulations for urban stream quality sufficient? *Environ Monit Assess* 192:702. <https://doi.org/10.1007/s10661-020-08657-9>
- Anderson EP, Jackson S, Tharme RE, Douglas M, Flotemersch JE, Zwarteveen M, Lokgariwar C, Montoya M, Wali A, Tipa GT, Jardine TD, Olden JD, Cheng L, Conallin J, Cosens B, Dickens C, Garrick D, Groenfeldt D, Kabogo J, Arthington AH (2019) Understanding rivers and their social relations: a critical step to advance environmental water management. *Wiley Interdiscip Rev Water* 6(6):1–21. <https://doi.org/10.1002/WAT2.1381>
- APHA (2017) Standard methods. In: In Standard methods for the examination of water and wastewater. American Public Health Association, Washington DC
- Barakat A, El Baghdadi M, Rais J, Aghezzaf B, Slassi M (2016) Assessment of spatial and seasonal water quality variation of Oum Er Rbia River (Morocco) using multivariate statistical techniques. *Int Soil Water Conserv Res* 4(4):284–292. <https://doi.org/10.1016/j.iswcr.2016.11.002>

- Baringhaus L, Henze N (2017) Cramer – Von Mises distance: probabilistic interpretation, confidence intervals, and neighborhood of model validation. *J Nonparametric Stat* 29(2):1–22. <https://doi.org/10.1080/10485252.2017.1285029>
- Bonansa RI, Amé MV, Wunderlin DA (2013) Determination of priority pesticides in water samples combining SPE and SPME coupled to GC-MS. A case study: Suquia River basin (Argentina). *Chemosphere* 90(6):1860–1869. <https://doi.org/10.1016/j.chemosphere.2012.10.007>
- Borgatello NG (2014) Hydrogeochemical determination in the San Luis River and the influence of polluting elements in the Bebedero Basin. Thesis, Universidad Nacional de San Luis (in Spanish)
- Brilly M, Rusjan S, Vidmar A (2006) Monitoring the impact of urbanization on the Glinscica stream. *Phys Chem Earth* 31(17):1089–1096. <https://doi.org/10.1016/j.pce.2006.07.005>
- Bu H, Song X, Zhang Y (2019) Using multivariate statistical analyses to identify and evaluate the main sources of contamination in a polluted river near to the Liaodong Bay in Northeast China. *Environ Pollut* 245:1058–1070. <https://doi.org/10.1016/j.envpol.2018.11.099>
- Bustamante IDE (1989) Methodological aspects in water quality studies. *Henares Revista de Geología* 36:25–36
- Bustamante I, Sanz J, Goy JFG, Encabo J, Mateos J (2002) Study of the quality of surface waters in the natural spaces at the south of the provinces of Salamanca and Ávila. Applications of the ISQA index. In *Geogaceta*, pp. 103–106. In Spanish
- Calderon MR, González P, Moglia M, Oliva Gonzáles S, Jofré M (2014) Use of multiple indicators to assess the environmental quality of urbanized aquatic surroundings in San Luis, Argentina. *Environ Monitor Assess* 186(7):4411–4422. <https://doi.org/10.1007/s10661-014-3707-8>
- Carrasco G, Molina JL, Patino-Alonso MC, Castillo MDC, Vicente-Galindo MP, Galindo-Villardón MP (2019) Water quality evaluation through a multivariate statistical HJ-Biplot approach. *J Hydrol* 577(July). <https://doi.org/10.1016/j.jhydrol.2019.123993>
- Casares MV, De Cabo LI (2018) Trend analysis of water quality monitoring data for El Riachuelo (Matanza-Riachuelo Basin, Argentina). *Rev Int Contamin Ambiental* 34(4):651–665. <https://doi.org/10.20937/RICA.2018.34.04.08>
- Castro MF, Almeida CA, Bazán C, Vidal J, Delfini CD, Villegas LB (2021) Impact of anthropogenic activities on an urban river through a comprehensive analysis of water and sediments. *Environ Sci Pollut Res* 28(28):37754–37767. <https://doi.org/10.1007/s11356-021-13349-z>
- Cazenave J, Bacchetta C, Parma MJ, Scarabotti PA, Wunderlin DA (2009) Multiple biomarkers responses in *Prochilodus lineatus* allowed assessing changes in the water quality of Salado River basin (Santa Fe, Argentina). *Environ Pollut* 157(11):3025–3033. <https://doi.org/10.1016/j.envpol.2009.05.055>
- Chilundo M, Kelderman P, Ókeeffe JH (2008) Design of a water quality monitoring network for the Limpopo River Basin in Mozambique. *Phys Chem Earth* 33(8–13):655–665. <https://doi.org/10.1016/j.pce.2008.06.055>
- Connor NP, Sarraino S, Frantz DE, Bushaw-Newton K, MacAvoy SE (2014) Geochemical characteristics of an urban river: influences of an anthropogenic landscape. *Appl Geochem* 47:209–216. <https://doi.org/10.1016/j.apgeochem.2014.06.012>
- del Corigliano MC (2008) Indexes to assess environmental quality in urban rivers. *Rev Univ Nac Río Cuarto* 28(1–2):33–54 (in Spanish)
- Da S, Le Billon P (2022) Sand mining: stopping the grind of unregulated supply chains. *Extract Indust Soc* 10(March):101070. <https://doi.org/10.1016/j.exis.2022.101070>
- Das BK, Kumar V, Chakraborty L, Swain HS, Ramteke MH, Saha A, Das A, Bhor M, Upadhyay A, Jana C, Manna RK, Samanta S, Tiwari NK, Ray A, Roy S, Bayen S, Gupta SD (2023) Receptor model-based source apportionment and ecological risk assessment of metals in sediment of river Ganga, India. *Marine Pollut Bull* 195(May):115477. <https://doi.org/10.1016/j.marpolbul.2023.115477>
- Dimri D, Daverey A, Kumar A, Sharma A (2021) Monitoring water quality of River Ganga using multivariate techniques and WQI (Water Quality Index) in Western Himalayan region of Uttarakhand, India. *Environ Nanotechnol Monitor Manag* 15:100375. <https://doi.org/10.1016/j.enmm.2020.100375>
- Dormann C (2020) *Environmental data analysis: an introduction with examples in R*. Springer, Freiburg. <https://doi.org/10.1007/978-3-030-55020-2>
- Duran Vian F, Pons Izquierdo JJ, Serrano Martínez M (2020) What is an urban river? A methodological approach for its delimitation in Spain. *Architect City Environ* 15(44):1–30. <https://doi.org/10.5821/ace.15.44.9035>
- Edokpayi JN, Odiyo JO, Msagati TAM, Potgieter N (2015) Temporal variations in physico-chemical and microbiological characteristics of Mvudi River, South Africa. *Int J Environ Res Public Health* 12(4):4128–4140. <https://doi.org/10.3390/ijerph120404128>
- Etemadi S, Khashei M (2021) Etemadi multiple linear regression. *Measurement: J Int Measure Confeder* 186(August):110080. <https://doi.org/10.1016/j.measurement.2021.110080>
- Ewaid SH, Abed SA, Kadhum SA (2018) Predicting the Tigris River water quality within Baghdad, Iraq by using water quality index and regression analysis. *Environ Technol Innov* 11:390–398. <https://doi.org/10.1016/j.eti.2018.06.013>
- Fan X, Cui B, Zhao H, Zhang Z, Zhang H (2010) Assessment of river water quality in Pearl River Delta using multivariate statistical techniques. *Procedia Environ Sci* 2(5):1220–1234. <https://doi.org/10.1016/j.proenv.2010.10.133>
- Fletcher TD, Andrieu H, Hamel P (2013) Understanding, management and modelling of urban hydrology and its consequences for receiving waters: a state of the art. *Adv Water Resour* 51:261–279. <https://doi.org/10.1016/j.advwatres.2012.09.001>
- Fontanarrosa MS, Gómez L, Avigliano L, Lavarello A, Zunino G, Sinistro R, Vera MS, Allende L (2023) Land uses in cities and their impacts on the water quality of urban freshwater blue spaces in the Pampean region (Argentina). *Environ Monit Assess* 195(6). <https://doi.org/10.1007/s10661-023-11216-7>
- Gatica EA, Almeida CA, Mallea MA, Del Corigliano MC, González P (2012) Water quality assessment, by statistical analysis, on rural and urban areas of Chocancharava River (Río Cuarto), Córdoba, Argentina. *Environ Monitor Assess* 184(12):7257–7274. <https://doi.org/10.1007/s10661-011-2495-7>
- Ghimire S, Pokhrel N, Pant S, Gyawali T, Koirala A, Mainali B, Angove MJ, Paudel SR (2022) Assessment of technologies for water quality control of the Bagmati River in Kathmandu Valley, Nepal. *Groundwater Sustain Dev* 18(March):100770. <https://doi.org/10.1016/j.gsd.2022.100770>
- Giorda EC (2021) *Sustentable use of water*. Dissertation. I Jornada del Día Mundial del Agua UNSL <https://www.youtube.com/watch?v=VD3hGcfMj2E&t=1902s> (in Spanish)
- González SO, Almeida CA, Calderón M, Mallea MA, González P (2014) Assessment of the water self-purification capacity on a river affected by organic pollution: application of chemometrics in spatial and temporal variations. *Environ Sci Pollut Res* 21(18):10583–10593. <https://doi.org/10.1007/s11356-014-3098-y>
- Greenacre M, Groenen PJF, Hastie T, D'Enza AI, Markos A, Tuzhilina E (2022) Principal component analysis. *Nat Rev Methods Primers* 2(1):100. <https://doi.org/10.1038/s43586-022-00184-w>

- Gupta AK, Kumar A, Maurya UK, Singh D, Islam S, Rathore AC, Kumar P, Singh R, Madhu M (2022) Comprehensive spatio-temporal benchmarking of surface water quality of Hindon River, a tributary of river Yamuna, India: adopting multivariate statistical approach. *Environ Sci Pollut Res*:0123456789. <https://doi.org/10.1007/s11356-022-24507-2>
- Gupta S, Gupta SK (2021a) A critical review on water quality index tool: genesis, evolution and future directions. *Eco Inform* 63(April):101299. <https://doi.org/10.1016/j.ecoinf.2021.101299>
- Gupta S, Gupta SK (2021b) Development and evaluation of an innovative Enhanced River Pollution Index model for holistic monitoring and management of river water quality. *Environ Sci Pollut Res* 28(21):27033–27046. <https://doi.org/10.1007/s11356-021-12501-z>
- Harrel FEJ (2015) Regression models strategies. Springer, Handbooks, Nashville. https://doi.org/10.1007/978-1-84628-288-1_21
- Hernandez-Ramirez AG, Martinez-Tavera E, Rodriguez-Espinosa PF, Mendoza-Pérez JA, Tabla-Hernandez J, Escobedo-Urías DC, Jonathan MP, Sujitha SB (2019) Detection, provenance and associated environmental risks of water quality pollutants during anomaly events in River Atoyac, Central Mexico: a real-time monitoring approach. *Sci Total Environ* 669:1019–1032. <https://doi.org/10.1016/j.scitotenv.2019.03.138>
- Howladar MF, Chakma E, Jahan Koley N, Islam S, Numanbakth MA, Al A, Z., Chowdhury, T. R., & Akter, S. (2021) The water quality and pollution sources assessment of Surma river, Bangladesh using, hydrochemical, multivariate statistical and water quality index methods. *Groundw Sustain Dev* 12:100523. <https://doi.org/10.1016/j.gsd.2020.100523>
- Instituto Nacional de Estadística y Censos (INDEC). (2010). Final results of the 2010 census. <http://www.censo2010>. Accessed: 7 August 2023 (in Spanish)
- Instituto Nacional de Estadística y Censos (INDEC) (2023) National census of population, households and housing 2022. Provisional results. Accessed: 7 August 2023 https://censo.gob.ar/index.php/mapa_poblacion2/
- James G, Witten D, Hastie T, Tibshirani R (2017) An introduction to statistical learning. Springer, New York
- Jiang Y, Li C, Sun L, Guo D, Zhang Y, Wang W (2021) A deep learning algorithm for multi-source data fusion to predict water quality of urban sewer networks. *J Clean Prod* 318(August):128533. <https://doi.org/10.1016/j.jclepro.2021.128533>
- Juez C, Franca MJ (2022) Physical impacts of sand mining in rivers and floodplains. In: *Encyclopedia of inland waters*, vol 4, 2nd edn. Elsevier Inc., Philadelphia. <https://doi.org/10.1016/B978-0-12-819166-8.00211-5>
- 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. <https://doi.org/10.1016/j.aca.2006.09.006>
- Kassambara A (2018) Machine learning essentials: practical guide in R. CreateSpace Independent Publishing Platform
- Kemgang Lekomo Y, Mwebi Ekengoue C, Douola A, Fotie Lele R, Christian Suh G, Obiri S, Kagou Dongmo A (2021) Assessing impacts of sand mining on water quality in Toutsang locality and design of waste water purification system. *Clean Eng Technol* 2(January):100045. <https://doi.org/10.1016/j.clet.2021.100045>
- Keupers I, Willems P (2017) Development and testing of a fast conceptual river water quality model. *Water Res* 113:62–71. <https://doi.org/10.1016/j.watres.2017.01.054>
- Khairabadi O, Shirmohamadi V, Sajadzadeh H (2023) Understanding the mechanism of regenerating urban rivers through exploring the lived experiences of residents: a case study of Abbas Abad river in Hamadan. *Environ Dev* 45(January):100801. <https://doi.org/10.1016/j.envdev.2023.100801>
- Kominkova D (2013) The urban stream syndrome – a mini-review. *The Open Environ Biol Monitor J* 5(1):24–29. <https://doi.org/10.2174/1875040001205010024>
- LeChevallier MW (1990) Coliform Regrowth in drinking water. *A Rev J/Am Water Works Assoc* 82(11):74–86. <https://doi.org/10.1002/j.1551-8833.1990.tb07054.x>
- Ledesma JA, Arrellano N (2022) Climate. Provincial Directorate of Statistics and Censuses. Min Sci Technol <http://www.estadistica.sanluis.gov.ar/wp-content/uploads/El-Clima-2022.pdf> (In spanish)
- López Fernández G, González Huecas C, López Lafuente A (1998) The quality of the waters of a river in the Duero basin: the Aguijoso. *Ingeniería Del Agua* 5:33–40 (In spanish)
- Losada Benavides LC, Rueda Sanabria CA, Martínez Silva P (2020) Evaluation of water quality in the El Quimbo hydroelectric reservoir. *Entre Ciencia e Ingeniería* 14(27):107–116. <https://doi.org/10.31908/19098367.1800> (In spanish)
- Lu Y, Chen J, Xu Q, Han Z, Peart M, Ng CN, Lee FYS, Hau BCH, Law WWY (2023) Spatiotemporal variations of river water turbidity in responding to rainstorm-streamflow processes and farming activities in a mountainous catchment, Lai Chi Wo, Hong Kong, China. *Sci Total Environ* 863. <https://doi.org/10.1016/j.scitotenv.2022.160759>
- Lupi L, Bertrand L, Monferrán MV, Amé MV, del Diaz MP (2019) 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(March):403–413. <https://doi.org/10.1016/j.jhydrol.2019.03.019>
- Mair P, Wilcox RR (2020) Robust statistical methods in R using the WRS2 package. *J Stat Softw* 52:464–488. <https://doi.org/10.3758/s13428-019-01246-w>
- Merlo C, Abril A, Amé MV, Argüello GA, Carreras HA, Chiappero MS, Hued AC, Wannaz E, Galanti LN, Monferrán MV, González CM, Solís VM (2011) Integral assessment of pollution in the Suquía River (Córdoba, Argentina) as a contribution to lotic ecosystem restoration programs. *Sci Total Environ* 409(23):5034–5045. <https://doi.org/10.1016/j.scitotenv.2011.08.037>
- Mgelwa AS, Hu YL, Ngaba MJY (2020) Patterns of nitrogen concentrations and their controls in two southern China urban river ecosystems. *Global Ecology and Conservation* 23:e01112. <https://doi.org/10.1016/j.gecco.2020.e01112>
- Nasrabadi T, Ruegner H, Sirdari ZZ, Schwientek M, Grathwohl P (2016) Using total suspended solids (TSS) and turbidity as proxies for evaluation of metal transport in river water. *Appl Geochem* 68:1–9. <https://doi.org/10.1016/j.apgeochem.2016.03.003>
- Nimptsch J, Wunderlin DA, Dollan A, Pflugmacher S (2005) Antioxidant and biotransformation enzymes in Myriophyllum quitense as biomarkers of heavy metal exposure and eutrophication in Suquía River basin (Córdoba, Argentina). *Chemosphere* 61(2):147–157. <https://doi.org/10.1016/j.chemosphere.2005.02.079>
- Obisesan KO, Christopher P (2018) Statistical models for evaluating water pollution: the case of Asejire and Eleyele Reservoirs in Nigeria. *J Environ Stat* 8(5):1–16
- Ortiz C (2017) Quality of the water and the riverside forest of the Chorrillos River: estimation of the ecological state using indices. Thesis. Universidad Nacional de San Luis (in Spanish)
- Ouali A, Azri C, Medhioub K, Ghrabi A (2009) Descriptive and multivariable analysis of the physico-chemical and biological parameters of Sfax wastewater treatment plant. *Desalination* 246(1–3):496–505. <https://doi.org/10.1016/j.desal.2008.04.058>
- Ouma YO, Okuku CO, Njau EN (2020) Use of artificial neural networks and multiple linear regression model for the prediction of dissolved oxygen in rivers: case study of hydrographic basin of River Nyando, Kenya. *Complexity* 2020. <https://doi.org/10.1155/2020/9570789>

- Pinto U, Maheshwari BL (2011) River health assessment in peri-urban landscapes: an application of multivariate analysis to identify the key variables. *Water Res* 45(13):3915–3924. <https://doi.org/10.1016/j.watres.2011.04.044>
- Pisani O, Bosch DD, Coffin AW, Endale DM, Liebert D, Strickland TC (2020) Riparian land cover and hydrology influence stream dissolved organic matter composition in an agricultural watershed. *Sci Total Environ* 717. <https://doi.org/10.1016/j.scitotenv.2020.137165>
- Provincial population by localities years 1869–2010. (n.d.). Dirección Provincial de Estadísticas y Censos. Ministerio de Ciencia y Tecnología. Gobierno de San Luis. Retrieved August 7, 2023, from <http://www.estadistica.sanluis.gov.ar/localidad/> (in Spanish)
- Queral R (1982) The quality of water in rivers. *Tecnología Del Agua* 4:49–57 (in Spanish)
- Rautenberg GE, Amé MV, Monferrán MV, Bonansea RI, Hued AC (2015) A multi-level approach using *Gambusia affinis* as a bioindicator of environmental pollution in the middle-lower basin of Suquia River. *Ecol Indic* 48:706–720. <https://doi.org/10.1016/j.ecolind.2014.09.025>
- Rentier ES, Cammeraat LH (2022) The environmental impacts of river sand mining. *Sci Total Environ* 838(May):155877. <https://doi.org/10.1016/j.scitotenv.2022.155877>
- Schirmer M, Leschik S, Musolff A (2013) Current research in urban hydrogeology - a review. *Adv Water Resour* 51:280–291. <https://doi.org/10.1016/j.advwatres.2012.06.015>
- Sikakwe GU, Nwachukwu AN, Uwa CU, Abam Eyong G (2020) Geochemical data handling, using multivariate statistical methods for environmental monitoring and pollution studies. *Environ Technol Innov* 18:100645. <https://doi.org/10.1016/j.eti.2020.100645>
- Singh G, Patel N, Jindal T, Srivastava P, Bhowmik A (2020) Assessment of spatial and temporal variations in water quality by the application of multivariate statistical methods in the Kali River, Uttar Pradesh, India. *Environ Monit Assess* 192(6). <https://doi.org/10.1007/s10661-020-08307-0>
- Trujillo-Ortiz A, Hernandez-Walls R (2007) Roystest: Royston's multivariate normality test. Mathworks <http://www.mathworks.com/matlabcentral/fileexchange/17811-roystest> Date accessed: 7 October 2023
- Unda-Calvo J, Ruiz-Romera E, Martínez-Santos M, Vidal M, Antigüedad I (2020) Multivariate statistical analyses for water and sediment quality index development: a study of susceptibility in an urban river. *Sci Total Environ* 711:135026. <https://doi.org/10.1016/j.scitotenv.2019.135026>
- Ustaoglu F, Tepe Y, Taş B (2020) Assessment of stream quality and health risk in a subtropical Turkey river system: a combined approach using statistical analysis and water quality index. *Ecol Indic* 113(May):105815. <https://doi.org/10.1016/j.ecolind.2019.105815>
- Valdés ME, Santos LHMLM, Rodríguez Castro MC, Giorgi A, Barceló D, Rodríguez-Mozaz S, Amé MV (2021) Distribution of antibiotics in water, sediments and biofilm in an urban river (Córdoba, Argentina, LA). *Environ Pollut* 269:116133. <https://doi.org/10.1016/j.envpol.2020.116133>
- Valentini M, dos Santos GB, Muller Vieira B (2021) Multiple linear regression analysis (MLR) applied for modeling a new WQI equation for monitoring the water quality of Mirim Lagoon, in the state of Rio Grande do Sul—Brazil. *SN Appl Sci* 3(1):1–11. <https://doi.org/10.1007/s42452-020-04005-1>
- Varol M (2020) Use of water quality index and multivariate statistical methods for the evaluation of water quality of a stream affected by multiple stressors: a case study. *Environ Pollut* 266:115417. <https://doi.org/10.1016/j.envpol.2020.115417>
- Viering T, Loog M (2021) The shape of learning curves: a review. *IEEE Trans Pattern Anal Mach Intell* 45:7799–7819. <https://doi.org/10.1109/TPAMI.2022.3220744>
- Vigiak O, Grizzetti B, Udias-Moinelo A, Zanni M, Dorati C, Bouraoui F, Pistocchi A (2019) Predicting biochemical oxygen demand in European freshwater bodies. *Sci Total Environ* 666:1089–1105. <https://doi.org/10.1016/j.scitotenv.2019.02.252>
- Walsh CJ, Roy AH, Feminella JW, Cottingham PD, Groffman PM, Morgan RP (2005) The urban stream syndrome: Current knowledge and the search for a cure. *J N Am Benthol Soc* 24(3):706–723. <https://doi.org/10.1899/04-028.1>
- Wang W, Yang P, Xia J, Huang H, Li J (2023) Impact of land use on water quality in buffer zones at different scales in the Poyang Lake, middle reaches of the Yangtze River basin. *Sci Total Environ* 896(April):165161. <https://doi.org/10.1016/j.scitotenv.2023.165161>
- Ward JH (1963) Hierarchical grouping to optimize an objective function. *J Am Stat Assoc* 58(301):236–244. <https://doi.org/10.1080/01621459.1963.10500845>
- Yan R, Yao J, Tian F, Gao J (2023) A novel framework for turbidity source apportionment of the urban lakeside river network. *Ecol Indic* 154(July):110561. <https://doi.org/10.1016/j.ecolind.2023.110561>
- Yildiz B, Bilbao JI, Sproul AB (2017) A review and analysis of regression and machine learning models on commercial building electricity load forecasting. *Renew Sust Energ Rev* 73:1104–1122. <https://doi.org/10.1016/j.rser.2017.02.023>

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.