






Multivariate water quality analysis of Lake Cajititlán, Mexico

Misael Sebastián Gradilla-Hernández  · José de Anda  ·
Alejandro Garcia-Gonzalez  · Demetrio Meza-Rodríguez  ·
Carlos Yebra Montes  · Yocanóchitl Perfecto-Avalos 

Received: 17 April 2019 / Accepted: 13 November 2019 / Published online: 3 December 2019
© Springer Nature Switzerland AG 2019

Abstract Lake Cajititlán is a shallow body of water located in an endorheic basin in western Mexico. This lake receives excess fertilizer runoff from agriculture and approximately 2.3 Hm³ per year of poorly treated wastewater

from three municipal treatment plants. Thirteen water quality parameters were monitored at five sampling points within the lake over 9 years. The objective of this work was to characterize the spatial and temporal variations of the water quality and to identify the sources of data variability in order to assess the influence and the impact of different natural and anthropogenic processes. One-way ANOVA tests, principal component analysis (PCA), cluster analysis (CA), and discriminant analysis (DA) were implemented. The one-way ANOVA showed that biochemical oxygen demand and pH present statistically significant spatial variations and that alkalinity, total chloride, conductivity, chemical oxygen demand, total hardness, ammonia, pH, total dissolved solids, and temperature present statistically significant temporal variations. PCA results explained both natural and anthropogenic processes and their relationship with water quality data. The CA results suggested there is no significant spatial variation in the water quality of the lake because of lake mixing caused by wind. The most significant parameters for spatial variations were pH, NO₃⁻, and NO₂⁻, consistent with the configuration of point and nonpoint sources that affect the lake's water quality. The temporal DA results suggested that conductivity, hardness, NO₂⁻, pH, and temperature were the most significant parameters to discriminate between seasons. The temporal behavior of these parameters was associated with the transport pathways of seasonal contaminants.

M. S. Gradilla-Hernández (✉) · A. Garcia-Gonzalez ·
Y. Perfecto-Avalos
Escuela de Ingeniería y Ciencias, Tecnológico de Monterrey, Av.
General Ramon Corona 2514, Nuevo México, CP,
45138 Zapopan, Jalisco, Mexico
e-mail: msgradilla@tec.mx

e-mail: sebastian.gradilla@jalisco.gob.mx

A. Garcia-Gonzalez
e-mail: alexgargo@tec.mx

Y. Perfecto-Avalos
e-mail: yocan@tec.mx

J. de Anda · D. Meza-Rodríguez
Centro de Investigación y Asistencia en Tecnología y Diseño del
Estado de Jalisco, A. C. Normalistas 800, CP, 44270 Guadalajara,
Jalisco, Mexico

J. de Anda
e-mail: janda@ciatej.mx

D. Meza-Rodríguez
e-mail: dmezar@ciatej.edu.mx

C. Yebra Montes
ENES-León, Universidad Nacional Autónoma de México, Blvd.
UNAM 2011, Predio el Saucillo y El Potrero, CP, 37684 León,
Guanajuato, Mexico
e-mail: carlosyebra@comunidad.unam.mx

Keywords Multivariate statistical analysis · Endorheic basin · Subtropical lakes · Shallow lakes · Anthropogenic contamination

Introduction

Surface water pollution

Surface waters are vulnerable to pollution due to urbanization and modernization trends worldwide. Although studies have shown that both natural and anthropogenic processes determine the surface water quality in a given area (Yang et al. 2010), agricultural, industrial, and urban activities are considered to be significant sources of aquatic ecosystem pollution (Ouyang et al. 2006). Municipal and industrial wastewater discharge are constant polluting sources for surface waters (Singh et al. 2004). Additionally, the extensive use of fertilizers and agrochemicals affects water quality due to runoff of these substances from soil to surface water. Nonpoint-source loads from agriculture are the main drivers of eutrophication in aquatic systems (Hempel et al. 2018). As a consequence, water quality parameters observed in surface water bodies, such as biological oxygen demand, chemical oxygen demand, nutrients, and conductivity, may gradually increase, the dissolved oxygen concentration may start to decrease, and nuisance algae blooms may appear as the first signs of advanced contamination. Low- or no-oxygen conditions in surface waters may lead to the death of fish and other organisms (Matson et al. 1997).

Multivariate techniques for water quality analysis

The regular monitoring of water conditions yields a complex matrix of numerous physical, chemical, and biological parameters. The patterns in these parameters are often difficult to interpret or extract meaningful conclusions (Costa et al. 2006; Dillon and Rigler 1974; Kittiwanih et al. 2007; Murphey 2006; Ryther and Dunstan 1971). Statistical analysis of water quality is used to understand the patterns in water quality measurements. The application of different multivariate statistical techniques allows the identification of possible sources of pollution that affect water resources as well as possible solutions to pollution problems (Vega et al. 1998).

Multivariate statistical analysis is used to extract the most representative information from extensive water quality datasets, and for proving special and temporal variations caused by natural, anthropogenic, or seasonal factors (Liebhold et al. 2004; Lofis et al. 1991; Potapova and Charles 2007; Thornton et al. 1990). The multivariate statistical techniques commonly used in water quality

datasets are principal component analysis (PCA), cluster analysis (CA), and discriminant analysis (DA). The PCA technique evaluates the correlations among the water quality parameters. Different strategies are adopted to apply PCA depending on the data subsets given by spatial and temporal factors (Bengraïne and Marhaba 2003; Ouyang et al. 2006; Pejman et al. 2009; Shrestha and Kazama 2007; Singh et al. 2004; Vega et al. 1998). CA is customarily applied to water quality datasets to group similar sampling sites, thus characterizing the spatial variability (Pejman et al. 2009; Shrestha and Kazama 2007; Vega et al. 1998). The result of a CA is a dendrogram, frequently obtained by Ward's method using squared Euclidean distances as a measure of similarity. In the studies performed by Singh et al. (2004) and Shrestha and Kazama (2007), DA was used to find the most significant parameters to classify the samples into temporal groups (seasons) and spatial groups (sampling sites). Next, an interpretation of the variability between the groups was given for each parameter, to obtain better information about the water quality, as well as to identify pollution sources, and understand the temporal/spatial variations in water quality for effective river water quality management and monitoring.

The objective of this work was to characterize the spatial and temporal variations of the water quality of Lake Cajititlán and to identify the sources of data variability in order to assess the influence and the impact of different natural and anthropogenic processes. As a contribution to the literature on the multivariate statistical analysis of water quality, the graphical analysis of PCA was improved using biplots in this study, describing the importance and correlations of the parameters with a higher influence on the two first principal components. For the DA analysis, linear discriminant functions were included, for which the coefficient of each variable is related to the contribution importance to classify an observation. Scatterplots were included to visualize the temporal DA results and to identify the classification patterns between temporal groups.

Material and methods

Study area and monitored parameters

Lake Cajititlán is a subtropical shallow body of water located in an endorheic basin in western Mexico (Fig. 1) at 1551 m above sea level. It has a surface area of

1744 ha, a maximum storage volume of 70.89 Hm³, and a maximum depth of 5.4 m at maximum capacity. This lake is found in a municipality with an elevated population growth rate. As a consequence, it receives a significant amount of discharge from wastewater treatment plants located in the vicinity of the lake, in addition to discharge of untreated wastewater from some small towns located around the lake and the rainfall runoff from large agriculture areas surrounding the lake. During the rainy season, excess fertilizer runoff to the lake from low-basin agricultural lands (see Fig. 1) and sediment resuspension in shallow lake waters increase the excess nutrients and organic matter in the water column (de Anda et al. 2019a).

In general terms, the lake has alkaline waters, an average diurnal dissolved oxygen concentration of about 8.9 mg/L, a biological oxygen demand (BOD₅) mean concentration of 18.6 mg/L, a chemical oxygen demand mean concentration of 185.9 mg/L, and total dissolved solids reaching 575.1 mg/L. Nutrient concentrations are also relatively high, with total phosphorus

reaching about 1.0 mg/L and total nitrogen mean concentrations of 8.5 mg/L. As a result of a mean annual temperature of 24 °C, low wind velocity, and an enrichment of nutrients in surface waters and sediments, the lake contains extremely high amounts of blue-green algae and high concentrations of chlorophyll that maintain an intense green color in its waters throughout the year. In previous works, this lake has been considered in the eutrophic state (de Anda et al. 2019a).

During the last decade, there have been several episodes of massive fish mortality. These episodes have occurred only during or immediately after the rainy season (Gradilla-Hernández et al. 2018). Due to this recurrent massive fish mortality, the State Water Commission of Jalisco (CEA, in Spanish) started a monitoring program, and water quality data involving multiple parameters has been obtained from 5 monitoring stations (see Fig. 1) since September 2009. The measurements have been made at a depth of 0.8 m for all five sampling stations. The coordinates of the five sampling points are shown in Table 1. Thirteen (13) water quality

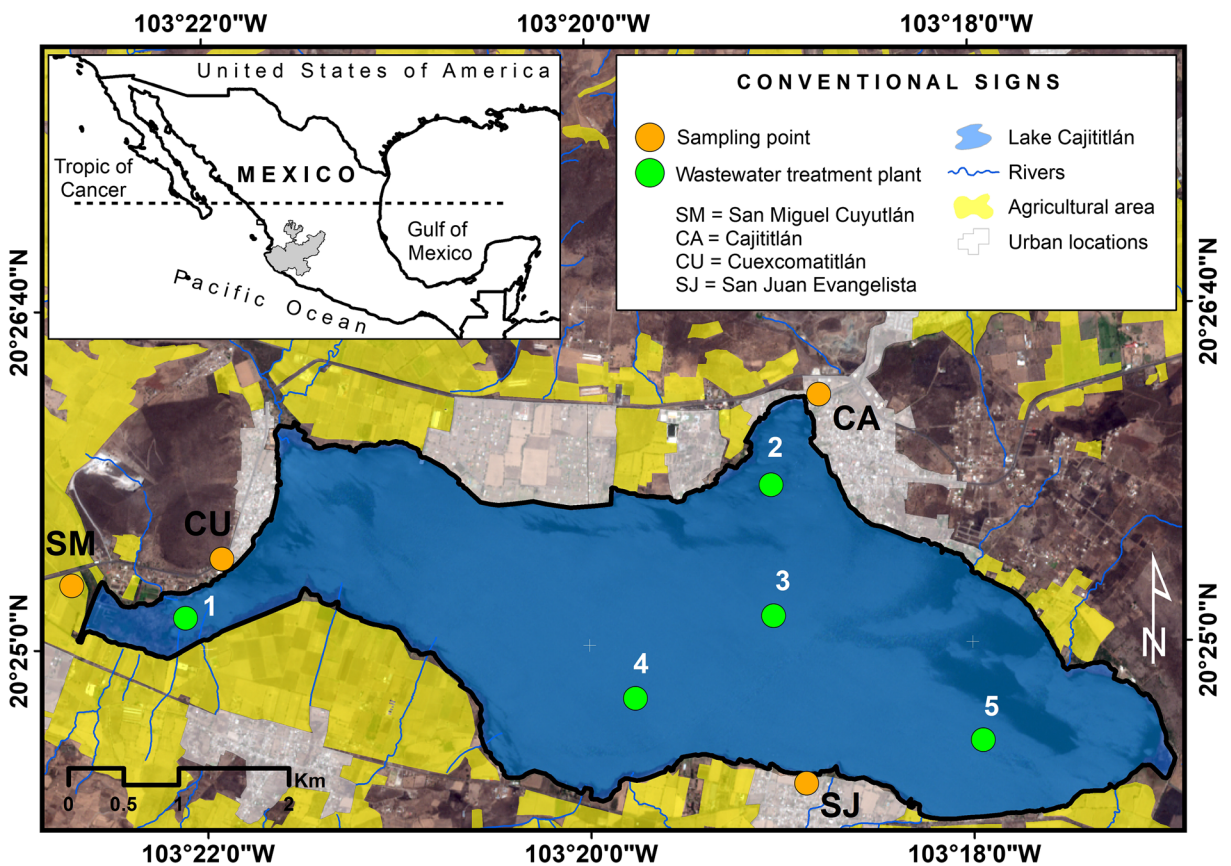


Fig. 1 Geographical location of Lake Cajititlán and locations of the sampling points from the Water Commission of Jalisco (CEA)

Table 1 Geographical coordinates and altitude of the sampling points in Lake Cajititlán

Sampling point	Geographical coordinates	Altitude, m a.s.l observed on April 5, 2018
1	20° 25' 25.2900" N 103° 21' 37.4400" W	1548.1
2	20° 25' 52.0100" N 103° 18' 59.0400" W	1547.5
3	20° 25' 8.20000" N 103° 19' 11.2000" W	1546.4
4	20° 24' 37.4283" N 103° 20' 8.2248" W	1546.74
5	20° 24' 31.2379" N 103° 17' 28.4126" W	1547.2

parameters were included in this study (Table 2). Temperature, pH, dissolved oxygen, and conductivity were measured on site by the State Water Commission. For the determination of the remaining water quality parameters, the samples were preserved at 4 °C and transported to the laboratory of the Water Commission of the state of Jalisco. On-site measurements and the analysis of the transported samples were both made by a laboratory certified to analyze water quality in compliance with Mexican regulations that are based on internationally approved protocols (CNA 2016; AWWA 2017) and imposed by the National Water Commission.

An additional monitoring campaign was conducted during the month of July of 2018, with the objective of

Table 2 Water quality parameters, abbreviations, and units observed for surface water in Lake Cajititlán

Parameters	Abbreviations	Units
Alkalinity	ALK	mg CaCO ₃ /L
Total chloride	CL	mg/L
Conductivity	CON	μS/cm
Biochemical oxygen demand	BOD	mg/L
Chemical oxygen demand	COD	mg/L
Total hardness	HAR	mg CaCO ₂ /L
Ammonia nitrogen	NH ₃	mg/L
Nitrate nitrogen	NO ₃ ⁻	mg/L
Nitrite nitrogen	NO ₂ ⁻	mg/L
Dissolved oxygen	DO	mg/L
pH	pH	-
Total dissolved solids	TDS	mg/L
Temperature	T	°C

measuring in situ the concentration of dissolved oxygen at night. Measurements were made between 4 and 7 a.m. at the 0.8-, 2.0-, and 3.0-m depths using the YSI 6600 V2 probe (YSI 2010).

Data processing and descriptive statistics

The raw data generated by CEA from September 2009 to April 2018 were obtained from the state water information system of the State of Jalisco as a time series (CEA 2018). In this way, a time-series vector was generated for each water quality parameter (P1 to P13) and sampling point (see Fig. 1). In total, 64 vectors (each with a size of 67 or 68) were generated (13 for each sampling point except for sampling point 5, which included only 12 vectors because dissolved oxygen values were missing). A total of 4352 values were contained in the data matrix.

A raw dataset may contain a percentage of data objects (outliers), which are considerably dissimilar to the rest of the data based on some measurement. Outliers may merely be noisy observations. Alternatively, they may indicate abnormal behavior in the system. It is important to detect the kind of identified outliers in the dataset in order to make the decision to remove or maintain these observations (Díaz Muñoz et al. 2012; Robinson et al. 2005). Therefore, the first statistical work for this data matrix was to identify the outliers in the time-series vectors of each of the analyzed parameters.

The boxplot is a common graphical tool to visualize the distribution of continuous data. However, when the data are skewed, usually many points exceed the whiskers and are often erroneously declared as outliers. Hubert and Vandervieren (2008) proposed an adjustment of the boxplot including a robust measurement of skewness in the determination of the whiskers. In this study, an adjusted box and whisker diagram method was used to detect outliers for asymmetric distributions. This resulted in a more accurate representation of the data and the determination of possible noisy observations, instead of data indicating abnormal behavior in the system, as proposed by Hubert and Vandervieren (2008).

After the detected outliers were removed from the data set, a nonlinear curve for each parameter and each sampling point was fit to the remaining data set. New values were then created by interpolating over the curve as proposed by Gnauck (2004), who suggested that missing data in long-term water quality data time series

have to be replaced by “artificial” data to obtain records; this can be done by interpolation, approximation or filtering of data sets.

After data processing was completed, the average, range, and standard deviation for each time series during the study period were calculated.

Principal component analysis

PCA is a statistical method used to reduce the dimensions of a large group of data (Jolliffe 1986; Jolliffe et al. 2003; Mackey 2009). PCA is a good technique for selecting the most significant variables and discarding those that are redundant or highly correlated (Pinto da Costa and Soares 2005). This method recognizes the variance within a sum of correlated variables to create a smaller group of uncorrelated variables called principal components (PCs), which are weighted linear combinations of the novel variables (Hotelling 1933; Pearson 1901). Principal components can be understood as an interaction of different observed variables, which describe the behavior of a single process that causes the link between these variables (Jolliffe 2002). To perform PCA, a multivariate random vector $x = (x_1, x_2, \dots, x_p)$ with mean μ and covariance Σ is considered (Jolliffe et al. 2003; Mackey 2009). In this study, such a multivariate vector is given by the water quality parameters: $x = (\text{ALK}, \text{CL}, \text{CON}, \text{BOD}, \text{COD}, \text{HAR}, \text{NH}_3, \text{NO}_3^-, \text{NO}_2^-, \text{OD}, \text{pH}, \text{TDS}, \text{T})$ Eq. 1.

Thus, 13 different linear combinations of x were obtained as

$$\begin{aligned}
 y_i = w'_i x = & w_{i1}P1 + w_{i2}P2 + w_{i3}P3 + w_{i4}P4 \\
 & + w_{i5}P5 + w_{i6}P6 + w_{i7}P7 + w_{i8}P8 \\
 & + w_{i9}P9 + w_{i10}P10 + w_{i11}P11 \\
 & + w_{i12}P12 + w_{i13} P13
 \end{aligned} \tag{1}$$

for suitable multipliers w_{ij} , resulting in 13 new random variables (y_1, y_2, \dots, y_{13}) called the principal components of x . The weights w_{ij} are also called loadings because they explain how much each of the original observations x_i contributes to each of the principal components. The loadings w_i are chosen so that the y_i have the largest possible variances, are mutually orthogonal, and have a unit length so that $w'_i w_i = 1$ (Jolliffe et al. 2003; Mackey 2009).

Eigenvalues were calculated to measure the significance of the components. The criteria used to determine the number of components to retain was to consider a sufficient number of components to explain between 70% and 90% of the total variation of the original variables (Jolliffe 2002; Rencher 2002). In this study, 5 principal components were retained for each sampling point, accounting for approximately 79% of the total variance. (Zelterman 2015). A biplot was also used to further interpret the first two principal components (Jolliffe 2002). Each vector in the biplot represents a parameter of the water quality data set, the length of the vector from the origin to the coordinates reflects the variance of that variable, and the correlation of two variables is reflected by the angle between the two corresponding vectors for the two variables: the smaller the angle, the greater the correlation (Jolliffe 1986; Pinto da Costa and Soares 2005).

Principal component analysis was performed using the software RStudio 1.1.456 with the factoextra package.

Spatial and temporal statistical analysis

The procedure to detect spatial variations of water quality parameters consisted, at first, of a univariate ANOVA to determine if the differences in the mean of each variable between sampling points were statistically significant. Afterward, observations were grouped into 5 sampling points and a spatial discriminant analysis was implemented in order to determine if the spatial variations could be classified as belonging to a specific sampling point. Finally, cluster analysis was used to group the observations based on their characteristics. Observations within the same cluster exhibit high internal homogeneity, while observations from different clusters show high external heterogeneity.

An analogous procedure was used to analyze temporal variations (except for cluster analysis, which was not performed given that there were only three seasonal groups). A One-way ANOVA was carried out to determine if the difference in the mean of each variable between seasons was statistically significant. Subsequently, a temporal discriminant analysis was used to classify the observations in three different seasonal groups and determine if temporal variations were significant to classify the observations as belonging to a specific season.

Univariate statistical analysis

As a first approach, we tested for spatial and temporal variations in water quality using univariate statistical analysis. One-way ANOVA was performed for each water quality parameter to determine if the difference in the means between sampling points was significant using a cutoff value of $p < 0.05$. Subsequently, Tukey's honest significance tests were performed by multiple comparisons of variable means between any pair of sampling points using a cutoff value of $p < 0.05$. Similar analyses were carried out to compare the mean values of the variables between temporal seasons.

Cluster analysis

Cluster analysis is a technique for recognizing similar and near objects within a dataset, and groups these objects into clusters based on their characteristics (Andreopoulos 2017; Hennig et al. 2016; Murtagh 1983; Pollard 1981; Savaresi et al. 2002). Thus, objects within the same cluster should exhibit high internal homogeneity, while objects from different clusters would show high external heterogeneity (Duda et al. 2001; Guénoche et al. 1991). The resulting clusters indicate patterns useful for analyzing the similarity of water quality tendencies between the sampling points.

Agglomerative hierarchical clustering techniques were used to produce partitions by a series of successive fusions of the 5 sampling points into groups. In this case, the vector containing the mean values of the variables in each sampling point was considered to compare the distances and to merge those with a small degree of dissimilarity as follows: The first step considered 5 clusters, $C_1, C_2, C_3, C_4,$ and C_5 , each containing a single sampling point, from SP1 to SP5, respectively. Then, the nearest pair of distinct clusters, C_i and C_j , was found, which were then merged and C_j was deleted, decreasing the number of clusters by one. If the number of clusters then was equal to one, the process was stopped; otherwise, the previous step was repeated (Everitt and Hothorn 2011). To calculate distances or similarities between pairs of vectors of means, the squared Euclidean distance was used, as shown in Eq. 2:

$$d_{ij} = \sum_{k=1}^p (x_{ik} - x_{jk})^2 \quad (2)$$

where p is the number of variables, and the mean values for k th variables localized in vectors i and j are

represented by $x_{i,k}$ and $x_{j,k}$, respectively. Ward's method was used to evaluate the distances between clusters to attempt to minimize the sum of the squares (SS) of any two hypothetical clusters that could be formed at each step. Ward's method is the most widely used clustering algorithm; when used in combination with the hierarchical method, it can be a powerful technique to group cases. The spatial variability of water quality between the vector of means of the sampling points was determined from CA, using the linkage distance $Dlink/Dmax$, which represents the quotient between the linkage distances for a particular case divided by the maximal linkage distance. The quotient was multiplied by 100 to standardize the linkage distance represented on the x -axis (Shrestha and Kazama 2007).

Cluster analysis was performed with the software R-3.5.3 using the `ggdendro` package.

Discriminant analysis

Discriminant analysis is a method that classifies samples into categorical dependent values using linear discrimination functions. A linear discrimination function is a linear combination of the variables for each observation in the data set. The maximum number of functions that is estimated is either equal to the number of variables or the number of groups minus one, whichever is smaller. Each successive linear discriminant function contributes less to the overall discriminatory power (Cacoullos 1973; Fisher 1936; Hotelling 1936).

This technique is used to obtain a statistical classification of multiple samples when there is prior knowledge of their belonging to a specific group (Campbell 1978; Huberty and Olejnik 2006). In this study, the discrimination functions were used to analyze the spatial and temporal water quality variations based on three different processes. The first process used was the standard method that incorporates all parameters; the second was the forward stepwise process, in which parameters are added one by one, starting with the most meaningful, until no important variations are found. Finally, the backward stepwise process was used, by which variables were extracted one after another; starting with the least significant variable and continuing until no significant change appears.

There are two types of functions in discriminant analysis: classification functions (Cfs) and linear discriminant functions (LDFs). Classification functions can be used to determine to which group each case most

likely belongs. In the case of this study, there were 5 different groups for spatial analysis (one for each sampling point). The season-correlated parameter was assumed to represent the major source of temporal variations in water quality. Therefore, 3 seasonal groups were used for temporal analysis as suggested by Ibarra-Montoya et al. (2012, 2010) for another subtropical Mexican lake (Aguamilpa): (i) The hot-dry season (HDS) comprising February–May; The wet season (WS) comprising June–September; and the cold-dry season (CDS) comprising October–January.

Each function was used to compute classification scores for each group, by applying Eq. 3:

$$Cf_i = c_i + \sum_{j=1}^m w_{ij} x_j \quad (3)$$

where i denotes the respective group, m is the number of variables in the data set, c_i is a constant value for the i th group, w_{ij} is the weight of the j th variable when computing the classification score for the i th group, x_j is the observed value for the respective case of the j th variable, and Cf_i is the resultant classification score. Once the classification scores were computed, each case was assigned to the group for which it had the highest classification score. Wilk's lambda statistic was used to denote the statistical significance of the discriminatory power of the models; its value ranges from 1 (no discriminatory power) to 0 (perfect discriminatory power).

Discriminant analysis, the respective canonical analysis, and scatterplots of scores were conducted and generated using STATISTICA 9.

Spatial distribution models

Spatial distribution models for the water quality parameters that present the most spatial variation were generated. Unknown point values were estimated using a mathematical function that minimizes the overall curvature of the surface, resulting in a smooth surface that passes exactly through the sample values in the sampling points (Huang and Stone 2003; Stone et al. 1997). The spline method was used to adjust the sample data to a polynomial function (North and Livingstone 2013; Parker et al. 2016). This method is preferred for generating slightly varying surfaces, such as pollution concentrations in water bodies (Kazemi et al. 2017).

Figures showing spatial distribution models for selected water quality parameters were generated using

the Spatial Analyst option of the ArcToolbox of ArcGis. The Spline Regularized Interpolation method was chosen.

Results and discussion

Descriptive statistics

The overall behavior of 13 lake water quality parameters from September 2009 to April 2018 is presented in Table 3. The water of Lake Cajititlán was found to be highly alkaline, with a mean pH in the range of 8.87 to 9.19 for all sampling points, likely due to the weathering process of the rock and soil located in its own basin. The predominant rock in the lake area is tuff (35.1%), igneous rocks of explosive origin, formed by loose or consolidated volcanic material. The second most abundant type of rock corresponds to basalt (28.60%). The predominant types of soil are vertisol (34.3%) and feozem (33.0%), which have large structures and high clay content. The soil color varies between black, dark gray, and reddish brown (IEEG Jalisco 2018). Due to the mineralization process of carbonaceous rocks and due to the presence of soil material rich in calcium and magnesium in the lake basin, the hardness of the lake waters is relatively high. The presence of ions of calcium and magnesium also increases the conductivity of the lake waters. In previous works, it was demonstrated that heavy metals present in sediment and in the sediment-water interface in Lake Cajititlán are mostly the result of the local geology. Therefore, the characteristics of the lake's waters appear to be mostly influenced by urban wastewater discharges and agricultural activities rather than by industrial pollution (de Anda et al. 2019b).

The lake waters also showed elevated total dissolved solids with a mean value of 582 mg/L. There are several potential sources of total dissolved solids into the lake, such as rainfall-runoff, runoff from agricultural drains, raw sewage discharges, and discharges from the wastewater treatments. Additionally, the lake is located into a deforested basin close to the urban area of Guadalajara (de Anda et al. 2019a).

The content of different forms of nitrogen (ammonia, nitrates and nitrites) in the lake waters were also very high, suggesting the input of agricultural runoff from the farming areas near the shoreline (Fig. 1), as well as an ongoing process of nitrification (Guo et al. 2014). The

Table 3 Mean, standard deviation, and range of the monitored water quality parameters in Lake Cajititlán in the period of September 2009 to April 2018

Parameter		Sampling point 1	Sampling point 2	Sampling point 3	Sampling point 4	Sampling point 5
ALK (mg CaCO ₃ /L)	Range	249.85–448.03	258–474.98	258–432.28	263.16–418.48	258–441.31
	Mean	352.60	358.10	358.03	358.30	362.51
	S.D.	44.19	41.75	39.50	36.62	40.95
CL ⁻ (mg/L)	Range	45.46–77.38	34.84–75.78	41.42–75.95	40.55–74.92	41.42–78.81
	Mean	62.21	62.35	62.71	62.31	63.16
	S.D.	7.05	8.26	7.35	7.27	8.00
CON (mS/cm)	Range	674.43–1108	697.2–1198	1106–685.93	711.89–1113	676–1140
	Mean	826.72	844.79	843.50	841.56	839.73
	S.D.	91.84	90.59	83.06	80.79	89.32
DBO (mg/L)	Range	9.72–39.24	6.42–27.36	11.28–36.6	8.49–35.34	8.85–33.84
	Mean	19.65	16.74	19.74	17.88	17.46
	S.D.	6.14	4.10	6.20	6.15	5.30
COD (mg/L)	Range	130.18–336.07	62.05–329.12	92.88–322.74	66.24–305.07	68.14–288.55
	Mean	198.68	186.89	56.72	181.22	177.60
	S.D.	50.86	57.66	183.02	50.90	47.09
HAR (mg CaCO ₃ /L)	Range	115.64–213.44	111.9–208.64	107.51–208.82	101.94–221.51	99.65–210
	Mean	152.41	152.01	151.86	151.78	153.51
	S.D.	22.72	24.66	24.77	25.98	26.40
NH ₃ (mg/L)	Range	0–5.57	0–3.65	0–4.7	0–5.82	0–6.91
	Mean	0.96	0.75	0.84	0.89	0.86
	S.D.	1.16	0.82	0.94	1.11	1.23
NO ₃ ⁻ (mg/L)	Range	0–0.16	0–0.12	0–0.35	0–0.15	0–0.16
	Mean	0.07	0.07	0.09	0.07	0.07
	S.D.	0.05	0.05	0.08	0.05	0.05
NO ₂ ⁻ (mg/L)	Range	0–0.02	0–0.02	0–0.02	0–0.02	0–0.02
	Mean	0.01	0.01	0.01	0.01	0.01
	S.D.	0.01	0.01	0.01	0.01	0.01
DO (mg/L)	Range	1.52–16.78	0–19.54	1.7–18.16	1.26–15.47	–
	Mean	8.06	8.02	9.94	8.09	–
	S.D.	4.18	4.80	3.64	2.97	–
pH	Range	7.40–9.27	8.27–9.58	8.22–9.45	8.05–9.46	8.29–9.53
	Mean	8.87	9.14	9.14	9.09	9.19
	S.D.	0.42	0.28	0.27	0.32	0.26
TDS (mg/L)	Range	420–701	286–670	397–697	379–667	375–691
	Mean	573.28	574.37	580.45	573.09	581.95
	S.D.	72.96	76.03	67.92	63.13	67.23
T (°C)	Range	16.80–28.47	15.06–27.50	17–28.56	15.3–26.44	17.5–28.54
	Mean	22.67	23.26	23.66	22.76	23.14
	S.D.	2.71	2.54	2.75	2.51	2.32

high chemical oxygen demand concentration demonstrates that the treatment plants are not performing satisfactorily, as previously suggested by de Anda et al.

(2019a). The decomposition of organic matter accumulated in the lake sediments and the oxidation of inorganic chemicals, such as ammonia and nitrite, also

contribute to the increase of chemical oxygen demand values in a waterbody (Akan et al. 2012). The presence of chlorine ions in the lake waters can be attributed to the use of sodium hypochlorite (NaOCl) in the final disinfection process of municipal wastewater treatment plants that discharge their waters directly to the lake.

The mean dissolved oxygen concentrations measured by CEA (Table 3) were between the values of 8.02 and 9.94 mg/L. These are average values measured at 80 cm of depth during the day. The values measured in the nightly monitoring campaign were between 2.10 and 4.44 mg/L (Table 4). It can be noted that the dissolved oxygen values measured during the nightly monitoring program are significantly lower than those reported by CEA. This is a phenomenon that commonly occurs in eutrophic lakes where the dissolved oxygen concentrations are higher during the day due to high radiation intensity that increases the photosynthetic activity of a large amount of blue-green algae; the dissolved oxygen concentration then drops at night due to consumption of a large amount of dissolved oxygen via respiration by microorganisms and algae (Qin et al. 2013; Duc Viet et al. 2016; de Anda et al. 2019a). Eutrophic bodies of water with the presence of a high number of blue-green algae frequently show levels of dissolved oxygen above local saturation values during the day (Duc Viet et al. 2016). Fertilizer enrichment increases algal biomass and increased algal metabolism results in higher rates of DO consumption during the night (Qin et al. 2013). The presence of fertilizers in

surface waters also intensifies the activity of nitrifying bacteria that generate energy for growth and maintenance using NH₃ and NH₄⁺ while they contribute to the oxygen depletion of their surroundings. In nitrification, ammonia and the ammonium cation are oxidized to nitrite, which is in turn oxidized to nitrate (Bollmann and Laanbroek 2011).

Events of massive death of fish have occurred in Lake Cajititlán mainly in the months of August and September at the end of the wet season (Gradilla-Hernández et al. 2018). When fertilizer transport by superficial runoff is increased and algal biomass is augmented, it can be expected that the lake water would show the lowest dissolved oxygen concentrations during the night, which would explain the death of fish by anoxia (Qin et al. 2013). Additionally, higher temperature values (occurring during the wet season) yield higher nitrification consumption rates and the higher levels of nutrient uptake by primary producers. Besides consuming the dissolved oxygen in the water column, a diverse set of algal species may produce toxins that may be harming the fish and other organisms (Smith and Schindler 2009).

Although the concentrations of nitrifying bacteria and blue-green algae were not included in the data matrix of this study, previous work has reported that Lake Cajititlán has high blue-green algae cells and chlorophyll concentrations in the water surface all over the lake extension, indicative of a high level of eutrophication (de Anda et al. 2019a).

Table 4 DO levels at night

Sampling point	Sampling depth (m)	DO (mg/L)
1	0.8	4.34
	2	2.28
2	0.8	3.43
	2	2.21
	≥ 3 (3.0)	2.86
3	0.8	2.95
	2	2.56
	≥ 3 (4.6)	2.39
4	0.8	2.48
	2	2.20
	≥ 3 (3.7)	2.10
5	0.8	3.46
	2	2.93
	≥ 3 (4.2)	2.72

Principal component analysis

Table 5 presents the results of the PCA analysis for each of the sampling points. Five significant components, making up more than 79% of the variance, were found for each sampling point. The first principal component (PC1) explained between 37.68 and 42.83% of the variability at all sampling points and was correlated with alkalinity, total chloride, conductivity, hardness, nitrate, nitrite, and total dissolved solids. Vega et al. (1998) and Bengraïne and Marhaba (2003) also found the presence of some of these water quality parameters in the first component of their PCA (with 27% and 37% of the explained variance, respectively) and linked these findings to the mineral and solute content of the water. PCA performed for another Mexican lake (Coyuca Lake) also found conductivity and total dissolved solids

Table 5 PCA analysis

Variable	SP1					SP2					SP3				
	PC1	PC2	PC3	PC4	PC5	PC1	PC2	PC3	PC4	PC5	PC1	PC2	PC3	PC4	PC5
	ALK	-0.39	0.12	-0.03	0.24	-0.18	-0.36	0.04	-0.08	0.33	-0.17	-0.39	0.02	-0.13	
CL	-0.35	0.30	-0.04	0.10	0.06	-0.36	0.18	-0.09	0.24	-0.01	-0.37	-0.21	-0.23		
CON	-0.35	0.27	-0.26	0.03	0.01	-0.31	0.19	-0.27	0.20	-0.01	-0.34	-0.04	-0.31		
BOD	0.08	0.15	0.22	0.75	0.02	-0.06	0.05	0.67	0.08	-0.50	0.04	-0.30	0.25		
COD	-0.16	0.40	0.03	-0.29	0.24	-0.20	0.39	-0.23	-0.23	-0.23	-0.22	-0.24	-0.30		
HAR	-0.41	-0.16	0.04	0.04	-0.02	-0.37	-0.18	-0.02	-0.17	-0.17	-0.39	0.08	0.11		
NH ₃	0.14	0.18	-0.42	0.38	0.05	0.28	0.16	-0.19	0.28	-0.51	0.19	-0.26	-0.35		
NO ₃ ⁻	0.30	0.46	-0.17	-0.33	-0.33	0.32	0.40	-0.11	0.08	0.13	0.29	-0.26	-0.34		
NO ₂ ⁻	0.34	0.42	-0.15	-0.33	-0.33	0.32	0.40	-0.11	0.08	0.10	0.32	-0.23	-0.36		
DO	0.10	0.25	0.54	0.20	-0.25	-0.02	0.51	0.22	-0.45	0.12	-0.04	-0.50	0.44		
pH	-0.04	0.25	0.59	-0.22	0.16	-0.22	0.26	0.15	-0.67	0.12	-0.13	-0.41	0.31		
TDS	-0.38	0.24	-0.06	-0.20	-0.20	-0.36	0.20	-0.10	0.07	0.14	-0.39	-0.08	-0.16		
T	0.12	0.13	-0.02	0.11	0.86	-0.06	0.21	0.53	0.43	0.54	0.01	-0.42	-0.02		
Eigenvalue	4.90	1.79	1.40	1.20	1.07	5.34	2.06	1.44	0.97	0.76	5.09	1.86	1.37		
% Total variance	37.68	13.79	10.77	9.23	8.20	41.11	15.84	11.09	7.48	5.86	39.17	14.33	10.53		
Cumulative % variance	37.68	51.48	62.25	71.48	79.68	41.11	56.95	68.04	75.52	81.38	39.17	53.50	64.03		

Variable	SP4					SP5				
	PC1	PC2	PC3	PC4	PC5	PC1	PC2	PC3	PC4	PC5
	ALK	0.22	-0.05	-0.38	0.11	-0.04	-0.37	-0.02	-0.18	-0.37
CL	0.12	0.04	-0.37	0.24	-0.16	-0.39	-0.22	0.10	-0.39	0.02
CON	-0.08	0.05	-0.32	0.29	-0.02	-0.35	-0.24	0.16	-0.35	0.10
BOD	0.60	-0.35	-0.01	-0.39	-0.44	0.01	0.25	-0.12	0.01	-0.59
COD	-0.38	-0.34	-0.24	0.40	-0.26	-0.24	-0.42	-0.10	-0.24	-0.13
HAR	0.11	-0.03	-0.39	-0.11	0.14	-0.38	0.15	-0.08	-0.38	0.08
NH ₃	0.34	-0.36	0.20	0.16	-0.19	0.20	0.15	-0.34	0.20	0.01
NO ₃ ⁻	-0.03	0.15	0.32	0.46	-0.21	0.27	-0.23	0.08	0.27	0.01
NO ₂ ⁻	-0.26	-0.26	0.34	0.42	-0.20	0.31	-0.48	0.04	0.31	0.01
DO	-0.28	-0.28	-0.02	-0.19	-0.58	-	-	-0.39	-	-
pH	-0.43	-0.01	-0.16	-0.12	-0.37	-0.16	-0.16	0.18	-0.16	-0.66
TDS	0.07	0.06	-0.36	0.18	-0.05	-0.08	-0.22	0.05	-0.38	0.00
T	0.24	0.75	0.04	-0.17	-0.32	0.30	-0.04	0.77	0.03	-0.41
Eigenvalue	1.13	0.93	5.02	1.69	1.53	1.25	1.82	1.00	5.14	1.33
% Total variance	8.68	7.16	38.63	13.01	11.75	9.64	15.17	7.67	42.83	11.08
Cumulative % variance	72.71	79.87	38.63	51.65	63.40	73.04	88.00	80.71	42.83	69.08

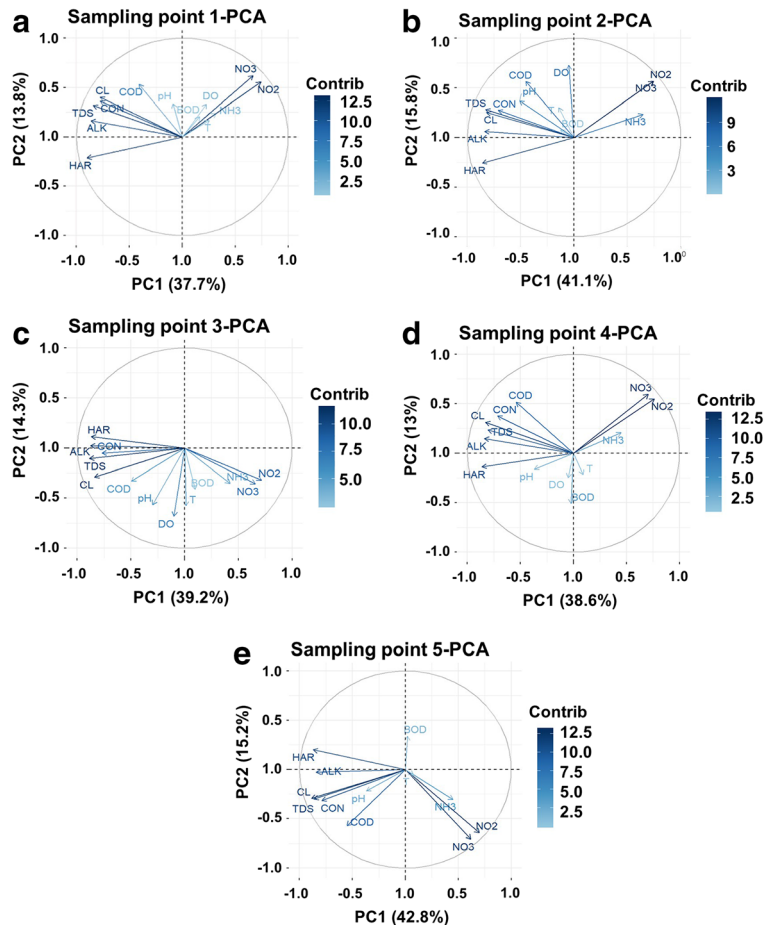
as the main elements of the first component (Ávila Pérez et al. 2015).

The second principal component (PC2) found in this study explained approximately 14% of the variance and was mainly composed of nitrate, nitrite, and chemical oxygen demand, except for sampling point 3, for which PC2 correlated with biochemical oxygen demand, dissolved oxygen, and pH. Badillo-Camacho et al. (2015) conducted a factor analysis of a tropical lake (Chapala), located just 18 km south of Lake Cajititlán, and found that nitrite and dissolved oxygen were related in one of the components, associating them with domestic wastewater and agricultural runoff. A previous study in Lake Cajititlán found the presence of direct discharges of raw wastewater along the shore of the lake, as well as a lack of measures to control the runoff from agricultural

areas (de Anda et al. 2019a). Several water quality studies (Bengraïne and Marhaba 2003; Ouyang et al. 2006; Vega et al. 1998; Pejman et al. 2009; Singh et al. 2004; Shrestha and Kazama (2007)) used principal component analysis to establish combinations of variables capable of describing the variability observed in the data sets. In this study, we improved the graphical analysis of PCA by means of biplots; this plot is the orthogonal projection of the data on the subspace spanned by the two first principal components (those with the most contribution to the total variance), describing the importance and correlations of the parameters with higher influence.

Together, the first two principal components explained approximately 54% of the data variability for each of the sampling points. The biplots in Fig. 2 show that the variables alkalinity, total chloride, conductivity,

Fig. 2 Biplots of PC1 and PC2; each vector represents a variable, and the correlation of two variables is reflected by the angle between the two corresponding vectors. The color scale and the length of each vector are related to the contribution to the total variance



total hardness, and total dissolved solids are highly correlated, as are nitrite and nitrate nitrogen; the two subsets of these variables are inversely correlated. In addition, the lengths of the vectors for dissolved oxygen, pH, and temperature denote the low contribution of these variables to the variance in the two first principal components.

For the remaining components, there are dissimilarities in the significant variables as it can be seen in Table 5. Although not all loadings for PC3, PC4, and PC5 are consistent for all five sampling points, these components consistently have significant loadings for pH as well as biochemical oxygen demand, chemical oxygen demand, nitrate, nitrite, and ammonia. Furthermore, the pH has the opposite sign than the other parameters, which suggests a negative correlation. Biochemical oxygen demand, chemical oxygen demand, nitrate, and nitrite are water quality parameters of concern related to municipal wastewater treatment systems not performing satisfactorily and agricultural runoff. Low pH values, on the other hand, are indicative of anaerobic bacterial environments that develop in reactors within treatment plants where wastes decompose (Akpór and Muchie 2011). Therefore, these principal components may also be associated with poorly treated municipal wastewater treatment plant effluents as well as agricultural runoff, which have been previously reported for Lake Cajititlán (de Anda et al. 2019a).

These principal components (PC1, PC2, and PC3) are related to both natural and anthropogenic processes. PC2 and PC3 can help describe the causes of the massive death of endemic and commercial fish species in the last years. The massive fish death events have occurred mainly at the end of the wet season (Gradilla-Hernández et al. 2018) when fertilizer transported by superficial runoff is increased and algal biomass is augmented. Furthermore, the elevated nutrient concentrations in Lake Cajititlán are increased by the effluents of treatment systems facilities which provide primary and secondary treatment but cannot remove nutrients from municipal wastewater (de Anda et al. 2019a).

Spatial and temporal statistical analysis

The one-way ANOVA showed that the variables with statistically significant mean variations between sampling points were biochemical oxygen demand (p value = 0.00451) and pH (p value = 3.58×10^{-8}). Tukey's honest significance test results indicate that pH varies

significantly between SP1 and the rest of the sampling points (SP1 has the lower mean for pH). The variables with statistically significant mean temporal variations were alkalinity (p value = 0.000306), total chloride (p value = 8.26×10^{-8}), conductivity (p value = 2.81×10^{-9}), chemical oxygen demand (valor $p = 3.92 \times 10^{-4}$), total hardness (p value = 7.66×10^{-9}), ammonia (p value = 0.0286), pH (p value = 7.63×10^{-6}), total dissolved solids (p value = 3.16×10^{-9}), and temperature (p value < 2×10^{-16}).

Spatial DA was performed with the data set comprising 12 parameters (since there was no available dissolved oxygen data for SP5 in the CEA dataset) after grouping into 5 sampling points. Classification functions (Cfs) and classification matrices (CMs) obtained from the standard, forward stepwise, and backward stepwise modes of DA are shown in Table 6 and Table 7, respectively. The standard stepwise mode CFs using 12 discriminant variables yielded the corresponding CMs, assigning 35.12% of the cases correctly (Table 6). The forward stepwise DA mode included 7 discriminant variables (alkalinity, biochemical oxygen demand, chemical oxygen demand, total hardness, nitrate, nitrite, and pH) in the classification function, with 31.66% cases assigned correctly. Backward stepwise mode DA gave CMs with 23.08% correct assignments using only the pH parameter. In the spatial DA, Wilk's lambda statistics were 0.663 for standard mode, 0.688 for forward mode, and 0.886 for backward mode. Thus, the spatial DA results suggest that a linear discriminant function does not assign the cases correctly.

The standardized coefficients for the four linear discriminant functions shown in Table 8 pertain to the standardized variables and therefore to comparable scales. The first function has a higher explained variance (72.8%). The considered parameters have the following order of significance: pH, chemical oxygen demand, alkalinity, total hardness, biochemical oxygen demand, nitrate, and nitrite. The most significant variable is pH with coefficient 0.956; thus, a positive relationship is suggested; observations with low pH will have low scores for the first discriminant function and vice versa. Additionally, the one-way ANOVA and the coefficients of the classification functions in backward mode suggest that there is a pH variation between sampling points. The plot of means for this variable (Fig. 3) shows that the mean pH in SP1 is lower than for the remaining sampling points, but this difference is nonsignificant to characterize the data of each specific sampling point. Out of the 7

Table 6 Classification functions for discriminant analysis of spatial variations

	Standard mode					Forward stepwise mode					Backward stepwise mode				
	SP1	SP2	SP3	SP4	SP5	SP1	SP2	SP3	SP4	SP5	SP1	SP2	SP3	SP4	SP5
Constant	-585.637	-623.873	-623.118	-618.044	-631.721	-511.246	-544.080	-544.417	-540.034	-553.408	-393.968	-417.872	-418.234	-413.526	-423.010
ALK	0.546	0.573	0.567	0.569	0.577	0.419	0.440	0.438	0.440	0.447					
CL	-1.217	-1.258	-1.263	-1.231	-1.226										
CON	0.105	0.110	0.109	0.110	0.107										
BOD	-0.413	-0.559	-0.464	-0.515	-0.554	-0.134	-0.275	-0.184	-0.236	-0.266					
COD	-0.199	-0.214	-0.215	-0.216	-0.221	-0.213	-0.228	-0.228	-0.229	-0.234					
HAR	0.215	0.193	0.197	0.190	0.188	0.034	0.003	0.012	0.006	0.003					
NH ₃	5.187	5.198	5.229	5.310	5.397										
NO ₃ ⁻	-49.738	-51.815	-25.683	-51.187	-46.246	-22.529	-23.775	2.551	-23.757	-18.587					
NO ₂ ⁻	1798.585	1815.099	1678.270	1801.464	1781.483	1653.054	1654.509	1522.555	1656.571	1638.605					
PH	106.508	110.690	110.337	110.220	111.582	100.440	104.386	104.167	103.845	105.238	91.172	90.654	91.692	91.692	91.692
TDS	-0.108	-0.116	-0.111	-0.114	-0.113										
T	3.162	3.262	3.252	3.156	3.208										

parameters for the second linear discriminant function (with explained variance of 25.7%), the most significant are nitrate and nitrite with coefficients of -1.784 and 1.456, respectively, indicating that observations with high nitrate values have low scores for this function. The first discriminant function mostly discriminates between SP1 and the others by means of the pH values; since SP1 observations have low pH, their scores for this function are low. The second function provides a discrimination for approximately 10 observations of SP3; since this sample point has the highest nitrate mean (Fig. 3), these observations have low scores.

The DA results indicate that there was no reliable classification for the water quality data for the different lake sampling points, indicating the lack of a significant spatial variation in the lake’s water quality. These results may be associated with the continuous mixing of the lake waters due to advection and diffusion processes driven predominantly by wind, which are exacerbated in shallow lakes with a mean depth < 3 m (Cajititlán Lake has a mean depth of 3.87 m) (de Anda et al. 2019a). Momentum transferred by wind via surface shear stresses generates waves, currents, and associated turbulence, which cause mixing of the lake water and diminishes spatial variations (Liu et al. 2018).

A CA was performed on the vector of means for each sampling point (see Table 3), and the resulting dendrogram is shown in Fig. 4. A useful criterion to select the number of statistically significant clusters is to consider the groups such that $(Dlink/Dmax)*100 < 60$. In this case, there would be only two clusters, one of which groups sampling points two to five, and the remaining group is made up of SP1. If the inequality $(Dlink/Dmax)*100 < 45$ is considered, as presented by Yang et al. (2010) for Lake Dianchi in China, there would be three clusters, one of which groups SP2, SP3, and SP4, and two groups (SP1 and SP5) with only one sampling point. These results are consistent with the lake configuration, as sites SP2, SP3, and SP4 are in the center of the lake, whereas SP1 is in the extreme west and SP5 is in the extreme east.

Figure 3 presents spatial distribution models and the plots of means of selected water quality parameters to give a graphical interpretation of the spatial variation of the means of these parameters (pH, nitrite, and nitrate). These parameters were selected since the DA results suggest they show the most spatial variation. Considering three clusters (C1 with SP1; C2 grouping SP2, SP3, and SP4; and C3 with SP5), the values of pH increase from C1 to C3 (C1 < C2 < C3).

These clusters have different characteristic pollution
Table 7 Summary of classification with cross-validation

Sampling points	%Correct	Sampling point assigned by DA				
		SP1	SP2	SP3	SP4	SP5
Standard DA mode						
SP1	57.35%	39	9	5	11	4
SP2	38.24%	8	26	9	9	16
SP3	26.47%	11	16	18	10	13
SP4	16.42%	11	17	8	11	20
SP5	37.31%	6	16	13	7	25
Total	35.21%	75	84	53	48	78
Forward stepwise DA mode						
SP1	55.88%	38	14	5	6	5
SP2	25.00%	14	17	7	9	21
SP3	22.06%	14	10	15	7	22
SP4	74.63%	16	20	6	5	20
SP5	47.76%	5	12	11	7	32
Total	31.66%	87	73	44	34	100
Backward stepwise DA mode						
SP1	42.65%	29	9	11	8	11
SP2	8.82%	15	6	5	8	34
SP3	2.94%	14	5	2	10	37
SP4	4.48%	18	5	10	3	31
SP5	56.72%	13	4	7	5	38
Total	23.08%	89	29	35	34	151

sources. Along the lake shoreline, there are four operational wastewater treatment plants. The largest plant treats approximately 60 L/s, and it is located closest to SP1 within the community of San Miguel Cuyutlán (see Fig. 1), which receives sewage from a significant number of users and has been reported to work only intermittently because of operation failures (de Anda et al. 2019a) and may be the reason why SP1 is separately clustered from the remaining sampling points. As mentioned earlier, low pH values may be indicative of anaerobic bacterial environments that develop in reactors within treatment plants.

Of the 7 parameters for the second linear discriminant function, the most significant were nitrate and nitrite. Most of the nitrate and nitrite in the lake surface waters result from runoff from agricultural land. Figure 1 shows that agricultural activity is intense and consistent around the lakeshore and that all of the regions of the lake are

connected to them, which might contribute to the fact that the spatial variations of NO_3^- and NO_2^- do not present as clear of trends as the spatial pH variations.

Temporal variations in water quality were further evaluated through DA. Temporal DA was performed after dividing the entire data set into three seasonal groups. Classification functions (Cfs) and matrices (CMs) obtained from the standard, forward stepwise, and backward stepwise modes of DA are shown in Tables 9 and 10, respectively. The standard stepwise mode Cfs using 12 discriminant variables yielded the corresponding CMs, assigning 77.2% of the cases correctly. The forward stepwise DA mode included 8 discriminant variables in the classification function, with 76.92% of the cases assigned correctly. However, in backward stepwise mode, DA gave CMs with 77.51% correct assignments using only five discriminant parameters (Table 10), with little difference in match for each season compared with the standard and forward stepwise modes. In the temporal DA, Wilk's lambda statistics were 0.342 for standard mode, 0.348 for forward mode, and 0.369 for backward mode. Thus, the temporal DA results suggest that conductivity, hardness, nitrite, pH, and temperature are the most significant parameters to discriminate between the three seasons, which means that these five parameters account for most of the expected temporal variations in the lake water quality. Table 8 presents the standardized coefficients for the linear discriminant functions of seasonal variations. In this case, two discriminant functions were estimated. For each stepwise mode, the significant variables were the same as for the classification functions, but in this analysis, the absolute value of each coefficient is related to the importance of the variable in classifying an observation. The following interpretations are given for the functions obtained from the backward stepwise mode. In the first discriminant function, temperature has the most significant coefficient (0.991); thus, observations with high temperature will have high scores for this function and vice versa. pH, total hardness, and nitrite contribute negatively for the function but are less significant than temperature. For the second discriminant function, total hardness and nitrite have significant comparable coefficients (-1.227 and -1.040 , respectively), such that an inverse relationship is suggested; that is, observations with high values for these variables will have low scores for the second discriminant function and vice versa. Figure 5 shows the scatterplot for the scores of the two linear discriminant functions; a pattern exists with overlapping zones for the data in the three different

Table 8 Standardized coefficients for linear discriminant functions of spatial and temporal variations

	Spatial variations												Temporal variations											
	Standard mode						Forward stepwise mode						Standard mode			Forward stepwise mode			Backward stepwise mode					
	LD-f1	LD-f2	LD-f3	LD-f4	LD-f3	LD-f4	LD-f1	LD-f2	LD-f3	LD-f4	LD-f3	LD-f4	LD-f1	LD-f2	LD-f3	LD-f1	LD-f2	LD-f3	LD-f1	LD-f2	LD-f3	LD-f1	LD-f2	LD-f3
ALK	0.769	-0.136	0.074	-0.142	0.698	0.091	-0.509	-0.057	-0.260	0.183	-0.223	-0.168	-0.019	0.259	-0.405	-0.294	-0.401	0.505	1.285	-0.431	0.207	0.269	0.216	-0.260
CL	-0.12	-0.19	-0.785	-0.204																				
CON	0.238	0.043	0.516	1.212																				
BOD	-0.486	0.444	-0.247	0.361	-0.475	-0.449	-0.533	0.000	0.066	-0.143														
COD	-0.672	-0.036	0.605	-0.338	-0.703	0.009	0.735	0.118	0.037	-0.136														
HAR	-0.408	0.104	0.249	-0.531	-0.513	-0.15	-0.188	1.436	0.518	-1.276														
NH ₃	0.09	-0.043	-0.515	-0.078																				
NO ₃ ⁻	0.177	1.715	0.093	-0.053	0.216	-1.784	0.617	0.029	0.015	-0.432	0.019	0.452												
NO ₂ ⁻	-0.133	-1.384	0.115	-0.196	-0.169	1.456	-0.801	0.873	0.165	-0.736	0.148	0.719												
PH	0.978	0.008	-0.097	-0.012	0.956	-0.048	0.118	0.046	0.496	-0.165	0.513	0.232												
TDS	-0.266	0.301	-0.436	-0.481					0.055	-0.253														
T	0.107	0.125	0.595	-0.501					-1.008	-0.149	-1.005	0.121												
Eigenvale	0.325	0.112	0.016	0.007	0.305	0.108	0.005	0.001	1.166	0.350	1.160	0.330												
Cum. Prop.	0.706	0.951	0.986	1.000	0.728	0.985	0.998	1.000	0.769	1.000	0.778	1.000												

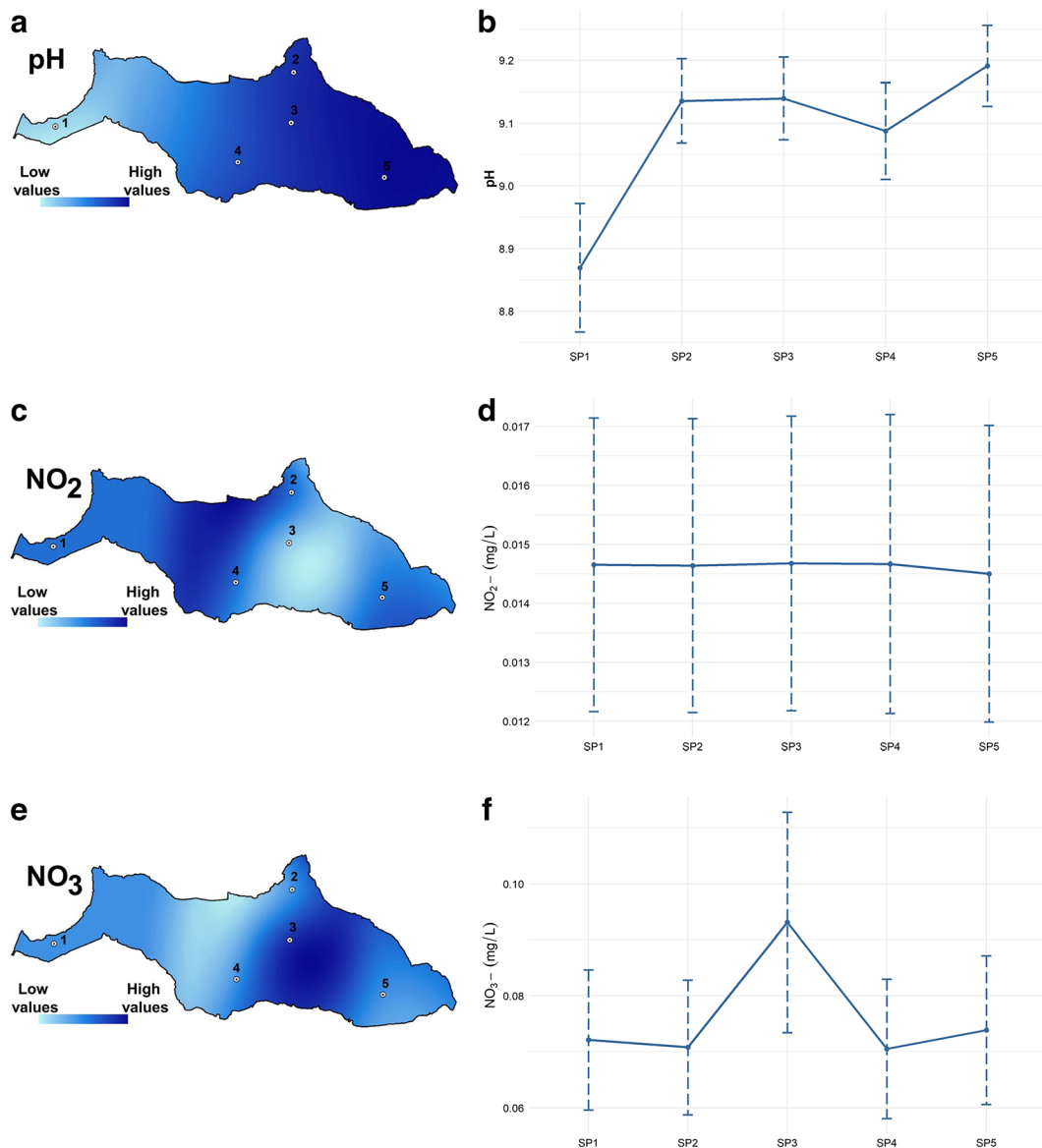


Fig. 3 Spatial distribution models and plot of the means showing spatial trends

seasons. Observations in the wet season have higher scores for the first discriminant function, followed by cases in the hot-dry season, and then the cold-dry season with lower scores (with an overlap during the last two seasons). This pattern is expected since the highest mean temperature occurs during the wet season (Fig. 6). On average, the observations during the hot-dry season have the lowest scores for the second discriminant function; this agrees with the interpretation that this season has the higher mean for total hardness.

The mean plots of selected parameters identified by DA are presented in Fig. 6. As mentioned above,

parameters showed different patterns during the year. A decrease in the average concentration of conductivity from the hot-dry season to the cold-dry season is observed. The average total hardness has the highest value for the hot-dry season. These trends in conductivity and hardness may be due to the effect of dilution of minerals and solute content during and after the rainy season. Because Cajitlán is a shallow lake located close to the Tropic of Cancer, water level variations between the dry and wet seasons are usually significant (de Anda et al. 2019a), and the dilution effects may be significant.

Table 9 Classification functions for discriminant analysis of temporal variations

	Standard mode			Forward stepwise mode			Backward stepwise mode		
	HDS	CDS	WS	HDS	CDS	WS	HDS	CDS	WS
Constant	- 575.740	- 549.666	- 558.721	- 530.007	- 503.456	- 512.091	- 504.842	- 475.582	- 485.211
ALK	0.408	0.411	0.425	0.221	0.223	0.235			
CL	- 1.325	- 1.274	- 1.286						
CON	0.107	0.098	0.112	0.034	0.025	0.039	0.074	0.065	0.081
BOD	0.221	0.192	0.176						
COD	- 0.110	- 0.113	- 0.113						
HAR	0.622	0.557	0.530	0.459	0.396	0.369	0.644	0.586	0.567
NH ₃	3.414	3.920	3.290	4.542	5.036	4.396			
NO ₃ ⁻	- 36.984	- 47.510	- 44.319	- 34.803	- 45.427	- 42.468			
NO ₂ ⁻	2381.109	2277.586	2274.520	1972.181	1874.019	1872.272	1958.124	1827.997	1807.614
PH	92.399	92.652	89.079	85.641	85.707	82.079	82.179	81.916	78.675
TDS	- 0.066	- 0.071	- 0.071						
T	3.037	2.591	3.914	2.588	2.152	3.464	2.524	2.097	3.360

The nitrite average has slightly increasing variations from the hot-dry season to the cold-dry season, even though it is not statistically significant (shown by the bars' overlap). The nitrite increase in the water column may be caused by increased water runoff in the wet season. Superficial runoff is a seasonal pathway that may transport fertilizers (Ouyang et al. 2006), which

can significantly increase the ammonium cation and ammonia (present in most fertilizers) and nitrite and nitrate due to the process of nitrification.

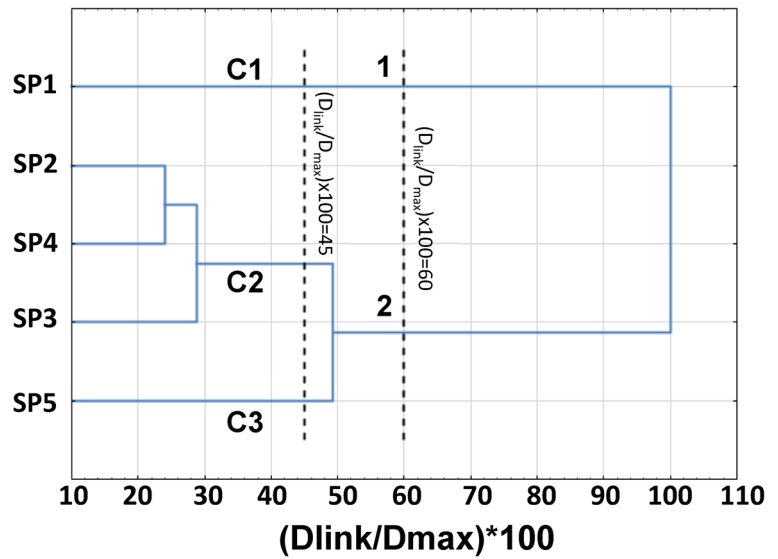
The average water temperature is higher in the wet season compared to the hot-dry season and the cold-dry season. At the same time, the average pH decreases from the hot-dry season to the wet season, and then increases in the cold-dry season. pH may increase due to the dragging of soil from the basin to the lake during the wet season. The predominant soil type is vertisol, which is alkaline because of its high content of clays (IEEG 2018). At the same time, higher temperature during the wet season would also result in elevated rates of nutrient uptake and oxygen production by blue-green algae, which would also increase the pH. As temperature increases, algae density levels may also increase, together with photosynthetic processes, which may reduce the water carbon dioxide levels in the water column and thus increase its pH (Qin et al. 2013).

In the studies performed by Singh et al. (2004) and Shrestha and Kazama (2007), discriminant analysis was used to find the most significant parameters to classify the samples in temporal groups (seasons) and spatial groups (sampling sites). Then an interpretation of the variability between the groups was given for each parameter, but the authors did not determine the importance of each water quality parameter to determine the membership of water quality data to some of the groups. In this study, the

Table 10 Summary of classification with cross-validation

Monitoring seasons	%Correct	Season assigned by DA		
		HDS	CDS	WS
Standard DA mode				
HDS	65.22%	75	19	21
CDS	72.04%	17	67	9
WS	91.54%	11	0	119
Total	77.22%	103	86	149
Forward stepwise DA mode				
HDS	63.48%	73	21	21
CDS	72.04%	16	67	10
WS	92.31%	10	0	120
Total	76.92%	99	88	151
Backward stepwise DA mode				
HDS	64.35%	74	23	18
CDS	70.97%	14	66	13
WS	93.85%	8	0	122
Total	77.51%	96	89	153

Fig. 4 Dendrogram for the vector of the means for each sampling point



analysis was improved by adding the standardized coefficients for linear discriminant functions and the scatterplot of the scores for these functions, providing an interpretation of the influence of some variables to classify the observations.

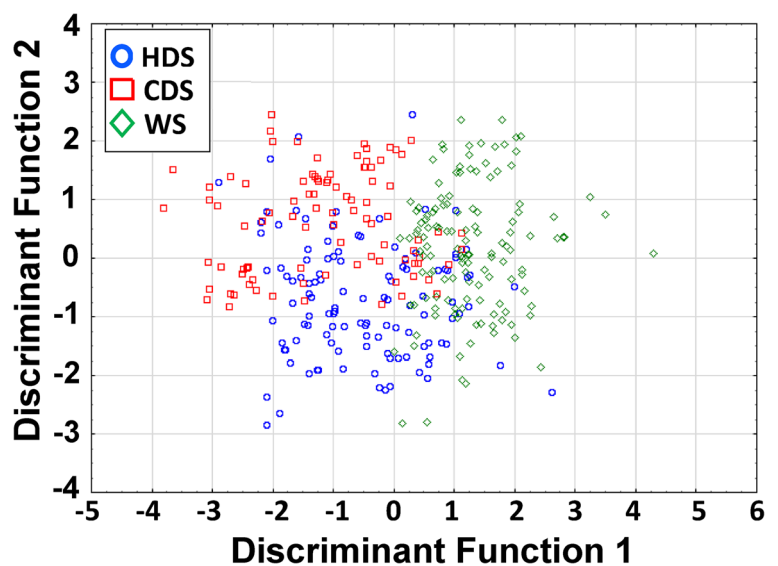
Conclusions

Water quality monitoring in many Mexican rivers and lakes is relatively new, and the data generated are very rarely analyzed and interpreted to generate more effective monitoring and management strategies. This

study contributes to the literature by providing a better understanding of the temporal and spatial variations of Lake Cajititlán to improve the monitoring strategies so that better decisions can be made and measures can be implemented to improve the lake’s water quality and protect its esthetic, social, environmental, and economic value. Further multivariate water quality studies of Lake Cajititlán should include other important water quality parameters, such as blue-green algae, chlorophyll, fecal coliforms, and heavy metals.

The fact that Lake Cajititlán is a subtropical shallow endorheic body of water that receives a sustained

Fig. 5 Scatterplot for the scores of the two first linear discriminant functions using the stand



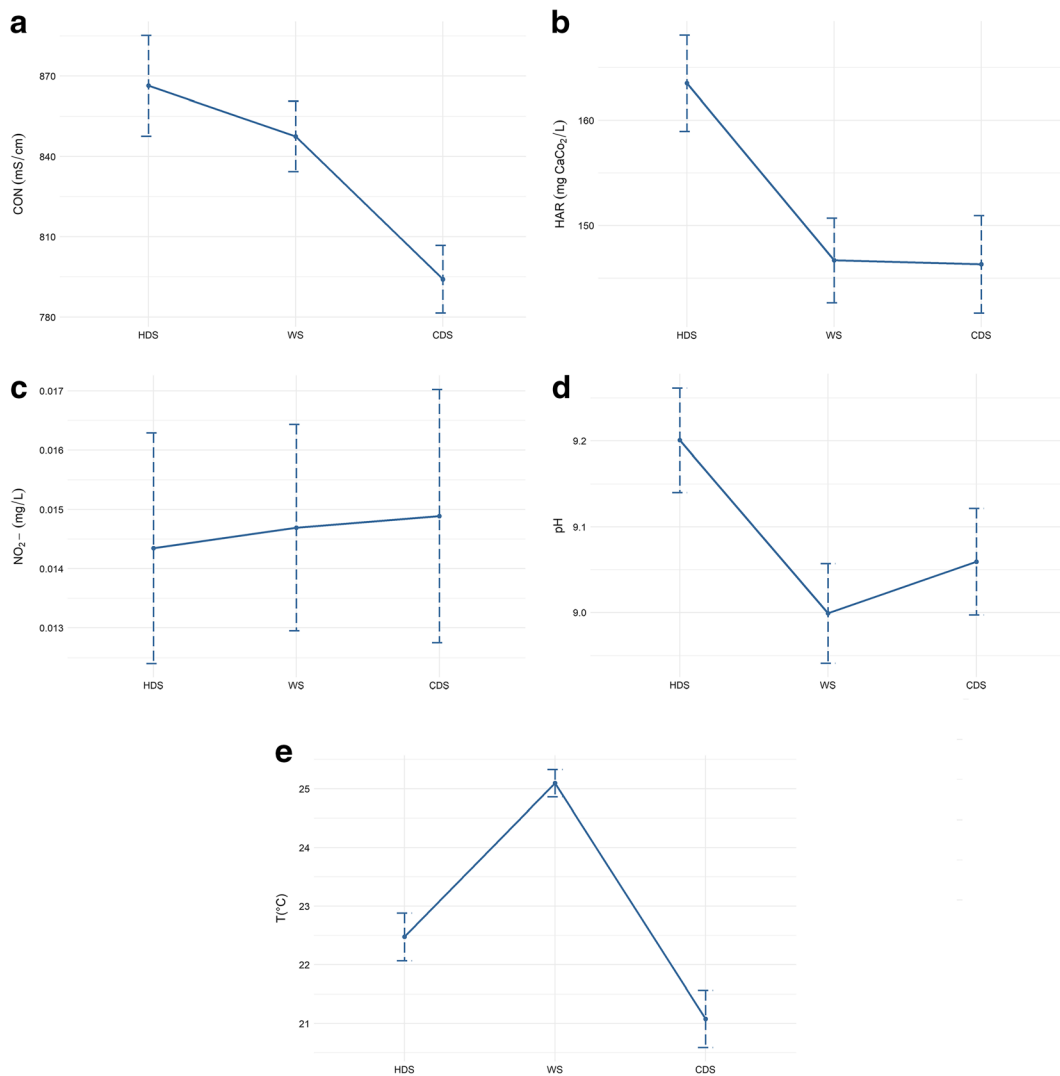


Fig. 6 Plots of means showing temporal trends

and significant amount of poorly treated municipal wastewaters and other discharges of agricultural drains and agricultural runoff during the rainy season, results in important temporal variations of water quality parameters. No significant spatial variations were identified in the water quality of the lake because of lake mixing caused by wind, which may be a significant momentum transfer process for shallow lakes. Variables such as biological oxygen demand, chemical oxygen demand and nutrient concentration are strongly associated with the phenomenon of blue-green algae growth in the lake. The presence of high blue-green algae populations is the main cause of important variations in the measured dissolved oxygen concentrations of the surface lake waters. When

the dissolved oxygen measurements are made during the first hours of the morning, the concentrations are usually low due to the respiration of blue-green algae. As the intensity of the light increases during the day, the process of photosynthesis begins to dominate and high concentrations of dissolved oxygen can be measured.

In order to improve the analysis carried out, time-series modeling could be used to detect trends and to predict the quality of water. To provide a quick way to assess the water quality of Lake Cajititlán, a widely used water quality index (WQI) could be implemented (such as the National Sanitation Foundation Water Quality Index, NSF-WQI). This index is a performance measurement that combines the information from

significant physical, chemical, and biological parameters into a functional form and it is a very practical method to take into account the critical quality parameters of a body of water and to reduce large amounts of data to a single number. Modified versions of the NSF-WQI could be developed to be applied for local conditions of Lake Cajititlán, to identify the change of trends and reflect seasonal variations of water quality as well as reduce the costs associated with monitoring water quality parameters.

References

- Akan, J. C., Abbagambo, M. T., Chellube, Z. M., & Abdulrahman, F. I. (2012). Assessment of pollutants in water and sediment samples in Lake Chad, Baga, North Eastern Nigeria. *Journal of Environmental Protection*. <https://doi.org/10.4236/jep.2012.311161>.
- Akpor, O. B., & Muchie, B. (2011). Environmental and public health implications of wastewater quality. *African Journal of Biotechnology*, *10*, 2379–2387.
- Andreopoulos, B. (2017). Clustering categorical data, *Wiley StatsRef: Statistics Reference Online*, <https://doi.org/10.1002/9781118445112.stat07907>.
- Ávila Pérez, H., García Ibañez, S., & Rosas-Acevedo, J. L. (2015). Análisis de Componentes Principales, como herramienta para interrelaciones entre variables fisicoquímicas y biológicas en un ecosistema léntico de Guerrero, México. *Revista Iberoamericana de Ciencias*, *2*, 43–53.
- AWWA. (2017). Standard methods for the examination of water and wastewater (23rd ed.). In E. W. Rice, R. B. Baird, & A. D. Eaton (Eds.), American Public Health Association, American Water Works Association, Water Environment Federation. ISBN: 9780875532875
- Badillo-Camacho, J., Reynaga-Delgado, E., Barcelo-Quintal, I., del Valle, P. F. Z., López-Chuken, U. J., Orozco-Guareño, E., Álvarez-Bobadilla, J. I., & Gómez-Salazar, S. (2015). Water quality assessment of a tropical Mexican lake using multivariate statistical techniques. *Journal of Environmental Protection*. <https://doi.org/10.4236/jep.2015.63022>.
- Bengraïne, K., & Marhaba, T. F. (2003). Using principal component analysis to monitor spatial and temporal changes in water quality. *Journal of Hazardous Materials*. [https://doi.org/10.1016/S0304-3894\(03\)00104-3](https://doi.org/10.1016/S0304-3894(03)00104-3).
- Bollmann, A., & Laanbroek, H. J. (2011). Nitrification in inland waters. In M. G. Klotz, B. B. Ward, & D. J. Arp (Eds.), *Nitrification* (pp. 385–403). Washington, DC: American Society for Microbiology Press.
- Cacoullos, T. (1973). *Discriminant analysis and applications*. Kent: Elsevier Science.
- Campbell, N. A. (1978). The influence function as an aid in outlier detection in discriminant analysis. *Applied Statistics*. <https://doi.org/10.2307/2347160>.
- CEA. (2018). Sistema de Calidad del Agua. Comisión Estatal del Agua del Estado de Jalisco, México [Resource Document]. <http://info.ceajalisco.gob.mx/sca/>. Accessed 25 Nov 2018.
- CNA. (2016). Normas Mexicanas Vigentes del Sector Hídrico. Comisión Nacional del Agua. <https://www.gob.mx/conagua/acciones-y-programas/normas-mexicanas-83266>. Accessed 10 Feb 2019.
- Costa, E., Pérez, J., & Krefft, J.-U. (2006). Why is metabolic labour divided in nitrification? *Trends in Microbiology*. <https://doi.org/10.1016/j.tim.2006.03.006>.
- de Anda, J., de J Díaz-Torres, J., Gradilla-Hernández, M. S., & de la Torre-Castro, L. M. (2019a). Morphometric and water quality features of Lake Cajititlán, Mexico. *Environmental Monitoring and Assessment*. <https://doi.org/10.1007/s10661-018-7163-8>.
- de Anda, J., Gradilla-Hernández, M. S., Díaz-Torres, O., de Jesús Díaz-Torres, J., & de la Torre-Castro, L. M. (2019b). Assessment of heavy metals in the surface sediments and sediment-water interface of Lake Cajititlán, Mexico. *Environmental Monitoring and Assessment*. <https://doi.org/10.1007/s10661-019-7524-y>.
- Díaz Muñoz, C., García Nieto, P. J., Alonso Fernández, J. R., Martínez Torres, J., & Taboada, J. (2012). Detection of outliers in water quality monitoring samples using functional data analysis in San Esteban estuary (Northern Spain). *Science of the Total Environment*. <https://doi.org/10.1016/j.scitotenv.2012.08.083>.
- Dillon, P. J., & Rigler, F. H. (1974). The phosphorus-chlorophyll relationship in lakes 1,2: phosphorus-chlorophyll relationship. *Limnology and Oceanography*. <https://doi.org/10.4319/l.o.1974.19.5.0767>.
- Duc Viet, N., Anh Bac, N., & Hoang, T. H. (2016). Dissolved oxygen as an indicator for eutrophication in freshwater lakes. *Proceedings of International Conference on Environmental Engineering and Management for Sustainable Development*. 47.
- Duda, R. O., Hart, P. E., & Stork, D. G. (2001). *Pattern classification*. New York: Wiley.
- Everitt, B., & Hothorn, T. (2011). *An introduction to applied multivariate analysis with R, Use R!* New York: Springer-Verlag.
- Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*. <https://doi.org/10.1111/j.1469-1809.1936.tb02137.x>.
- Gnauck, A. (2004). Interpolation and approximation of water quality time series and process identification. *Analytical and Bioanalytical Chemistry*. <https://doi.org/10.1007/s00216-004-2799-3>.
- Gradilla-Hernández, M. S., de Anda-Sanchez, J., Ruiz-Palomino, P., Barrios-Piña, H., Senés-Guerrero, C., Del Toro Barboza, M., & Vázquez-Toral, M. P. (2018). Estudio Preliminar del Índice de Calidad de Agua en el Lago de Cajititlán y su Potencial Predictivo de la Mortandad Masiva de Peces, In *Memorias del congreso nacional de hidráulica 2018*.
- Guénoche, A., Hansen, P., & Jaumard, B. (1991). Efficient algorithms for divisive hierarchical clustering with the diameter criterion. *Journal of Classification*. <https://doi.org/10.1007/BF02616245>.
- Guo, W., Fu, Y., Ruan, B., Ge, H., & Zhao, N. (2014). Agricultural non-point source pollution in the Yongding River Basin. *Ecological Indicators*. <https://doi.org/10.1016/j.ecolind.2013.07.012>.
- Hampel, J. J., McCarthy, M. J., Gardner, W. S., Zhang, L., Xu, H., Zhu, G., & Newell, S. E. (2018). Nitrification and

- ammonium dynamics in Taihu Lake, China: seasonal competition for ammonium between nitrifiers and cyanobacteria. *Biogeosciences*. <https://doi.org/10.5194/bg-15-733-2018>.
- Hennig, C. M., Meilä, M., Murtagh, F., & Rocci, R. (Eds.). (2016). *Handbook of cluster analysis, Chapman & Hall/CRC handbooks of modern statistical methods*. Boca Raton: CRC Press, Taylor & Francis Group.
- Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*. <https://doi.org/10.1037/h0071325>.
- Hotelling, H. (1936). Relations Between Two Sets of Variates. In S. Kotz & N. L. Johnson (Eds.), *Breakthroughs in statistics: methodology and distribution* (pp. 162–190). New York: Springer.
- Huang, J. Z., & Stone, C. J. (2003). Extended linear modeling with splines. In D. D. Denison, M. H. Hansen, C. C. Holmes, B. Mallick, & B. Yu (Eds.), *Nonlinear estimation and classification, lecture notes in statistics* (pp. 213–233). New York: Springer.
- Hubert, M., & Vandervieren, E. (2008). An adjusted boxplot for skewed distributions. *Computational Statistics & Data Analysis*. <https://doi.org/10.1016/j.csda.2007.11.008>.
- Huberty, C. J., & Olejnik, S. (2006). *Applied MANOVA and discriminant analysis*. New Jersey: Wiley.
- Ibarra-Montoya, J. L., Rangel-Peraza, G., González-Farías, F. A., Anda, J. D., Zamudio-Reséndiz, M. E., Martínez-Meyer, E., & Macías-Cuellar, H. (2010). Modelo de nicho ecológico para predecir la distribución potencial de fitoplancton en la Presa Hidroeléctrica Aguamilpa, Nayarit, México. *Ambiente & Agua - An Interdisciplinary Journal of Applied Science*. <https://doi.org/10.4136/ambi-agua.154>.
- Ibarra-Montoya, J. L., Rangel-Peraza, G., González-Farías, F. A., Anda, J. D., Martínez-Meyer, E., & Macías-Cuellar, H. (2012). Uso del modelado de nicho ecológico como una herramienta para predecir la distribución potencial de *Microcystis* sp (cianobacteria) en la Presa Hidroeléctrica de Aguamilpa, Nayarit, México. *Ambiente & Agua - An Interdisciplinary Journal of Applied Science*. <https://doi.org/10.4136/ambi-agua.607>.
- IIEG Jalisco (2018). Municipal diagnosis: Tlajomulco de Zúñiga. [Resource Document]. <https://iieg.gob.mx/contenido/Municipios/TlajomulcodeZuniga.pdf>. Accessed 10 Apr 2019.
- Jolliffe, I. T. (1986). Principal component analysis and factor analysis. In I. T. Jolliffe (Ed.), *Principal component analysis* (pp. 115–128). New York: Springer.
- Jolliffe, I. T. (2002). *Principal component analysis*. New York: Springer-Verlag.
- Jolliffe, I. T., Trendafilov, N. T., & Uddin, M. (2003). A modified principal component technique based on the LASSO. *Journal of Computational and Graphical Statistics*. <https://doi.org/10.1198/1061860032148>.
- Kazemi, E., Karyab, H., & Emamjome, M.-M. (2017). Optimization of interpolation method for nitrate pollution in groundwater and assessing vulnerability with IPNOA and IPNOC method in Qazvin plain. *Journal of Environmental Health Science and Engineering*. <https://doi.org/10.1186/s40201-017-0287-x>.
- Kittiwonich, J., Yamamoto, T., Kawaguchi, O., & Hashimoto, T. (2007). Analyses of phosphorus and nitrogen cyclings in the estuarine ecosystem of Hiroshima Bay by a pelagic and benthic coupled model. *Estuarine, Coastal and Shelf Science, Biodiversity and Ecosystem Functioning in Coastal and Transitional Waters*. <https://doi.org/10.1016/j.ecss.2007.04.029>.
- Liebhold, A., Koenig, W. D., & Bjørnstad, O. N. (2004). Spatial synchrony in population dynamics. *Annual Review of Ecology, Evolution, and Systematics*. <https://doi.org/10.1146/annurev.ecolsys.34.011802.132516>.
- Liu, Z., Hu, J., Zhong, P., Zhang, X., Ning, J., Larsen, S. E., Chen, D., Gao, Y., He, H., & Jeppesen, E. (2018). Successful restoration of a tropical shallow eutrophic lake: strong bottom-up but weak top-down effects recorded. *Water Research*. <https://doi.org/10.1016/j.watres.2018.09.007>.
- Loftis, J. C., Taylor, C. H., Newell, A. D., & Chapman, P. L. (1991). Multivariate trend testing of Lake Water quality. *Journal of the American Water Resources Association*. <https://doi.org/10.1111/j.1752-1688.1991.tb01446.x>.
- Mackey, L. W. (2009). Deflation methods for sparse PCA. In D. Koller, D. Schuurmans, Y. Bengio, & L. Bottou (Eds.), *Advances in Neural Information Processing Systems* (Vol. 21, pp. 1017–1024). New York: Curran Associates, Inc..
- Matson, P. A., Parton, W. J., Power, A. G., & Swift, M. J. (1997). Agricultural intensification and ecosystem properties. *Science*. <https://doi.org/10.1126/science.277.5325.504>.
- Murphey, S. F. (2006). *Water quality of Boulder Creek, Colorado*. Reston: U.S. Geological Survey.
- Murtagh, F. (1983). A survey of recent advances in hierarchical clustering algorithms. *The Computer Journal*. <https://doi.org/10.1093/comjnl/26.4.354>.
- North, R. P., & Livingstone, D. M. (2013). Comparison of linear and cubic spline methods of interpolating lake water column profiles: interpolation of lake profiles. *Limnology and Oceanography: Methods*. <https://doi.org/10.4319/lom.2013.11.213>.
- Ouyang, Y., Nkedi-Kizza, P., Wu, Q. T., Shinde, D., & Huang, C. H. (2006). Assessment of seasonal variations in surface water quality. *Water Research*. <https://doi.org/10.1016/j.watres.2006.08.030>.
- Parker, S. J., Butler, A. P., & Jackson, C. R. (2016). Seasonal and interannual behaviour of groundwater catchment boundaries in a Chalk aquifer: seasonal groundwater catchment dynamics. *Hydrological Processes*. <https://doi.org/10.1002/hyp.10540>.
- Pearson, K. (1901). On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*. <https://doi.org/10.1080/14786440109462720>.
- Pejman, A. H., Bidhendi, G. R. N., Karbassi, A. R., Mehrdadi, N., & Bidhendi, M. E. (2009). Evaluation of spatial and seasonal variations in surface water quality using multivariate statistical techniques. *International journal of Environmental Science and Technology*. <https://doi.org/10.1007/BF03326086>.
- Pinto da Costa, J., & Soares, C. (2005). A weighted rank measure of correlation. *Australian & New Zealand Journal of Statistics*. <https://doi.org/10.1111/j.1467-842X.2005.00413.x>.
- Pollard, D. (1981). Strong consistency of K-means clustering. *Ann. Statist.* <https://doi.org/10.1214/aos/1176345339>.
- Potapova, M., & Charles, D. F. (2007). Diatom metrics for monitoring eutrophication in rivers of the United States.

- Ecological Indicators*, 7, 48–70. <https://doi.org/10.1016/j.ecolind.2005.10.001>.
- Qin, B., Gao, G., Zhu, G., Zhang, Y., Song, Y., Tang, X., & XU Hail., Deng, J. (2013). Lake eutrophication and its ecosystem response. *Chinese Science Bulletin*. <https://doi.org/10.1007/s11434-012-5560-x>.
- Rencher, A. C. (2002). *Methods of multivariate analysis*. New York: Wiley.
- Robinson, R. B., Chris, C., & Odom, K. (2005). Identifying outliers in correlated water quality data. *Journal of Environmental Engineering*. [https://doi.org/10.1061/\(ASCE\)0733-9372\(2005\)131:4\(651](https://doi.org/10.1061/(ASCE)0733-9372(2005)131:4(651).
- Ryther, J. H., & Dunstan, W. M. (1971). Nitrogen, phosphorus, and eutrophication in the coastal marine environment. *Science*. <https://doi.org/10.1126/science.171.3975.1008>.
- Savaresi, S. M., Boley, D. L., Bittanti, S., & Gazzaniga, G. (2002). Cluster selection in divisive clustering algorithms. In *Proceedings of the 2002 SIAM International Conference on Data Mining* (pp. 299–314). Philadelphia: Society for Industrial and Applied Mathematics.
- Shrestha, S., & Kazama, F. (2007). Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan. *Environmental Modelling & Software*. <https://doi.org/10.1016/j.envsoft.2006.02.001>.
- Singh, K. P., Malik, A., Mohan, D., & Sinha, S. (2004). Multivariate statistical techniques for the evaluation of spatial and temporal variations in water quality of Gomti River (India)—a case study. *Water Research*. <https://doi.org/10.1016/j.watres.2004.06.011>.
- Smith, V. H., & Schindler, D. W. (2009). Eutrophication science: where do we go from here? *Trends in Ecology & Evolution*. <https://doi.org/10.1016/j.tree.2008.11.009>.
- Stone, C. J., Hansen, M. H., Kooperberg, C., & Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling. *The Annals of Statistics*, 25, 1371–1425.
- Thornton, K. W., Kimmel, B. L., & Payne, F. E. (1990). *Reservoir limnology: ecological perspectives*. New York: Wiley.
- Vega, M., Pardo, R., Barrado, E., & Debán, L. (1998). Assessment of seasonal and polluting effects on the quality of river water by exploratory data analysis. *Water Research*. [https://doi.org/10.1016/S0043-1354\(98\)00138-9](https://doi.org/10.1016/S0043-1354(98)00138-9).
- Yang, Y.-H., Zhou, F., Guo, H.-C., Sheng, H., Liu, H., Dao, X., & He, C.-J. (2010). Analysis of spatial and temporal water pollution patterns in Lake Dianchi using multivariate statistical methods. *Environmental Monitoring and Assessment*. <https://doi.org/10.1007/s10661-009-1242-9>.
- YSI. (2010). YSI 6600 V2 Sonde. YSI Incorporated. <https://www.ysi.com/File%20Library/Documents/Specification%20Sheets/E52-6600V2.pdf>. Accessed 31 Oct 2018.
- Zelterman, D. (2015). *Applied multivariate statistics with R, Statistics for biology and health*. Cham: Springer.

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