ENVIRONMENTAL TECHNOLOGY
2021, VOL. 42, NO. 27, 4286–4295
https://doi.org/10.1080/09593330.2020.1754922

Taylor & Francis
Taylor & Francis Group

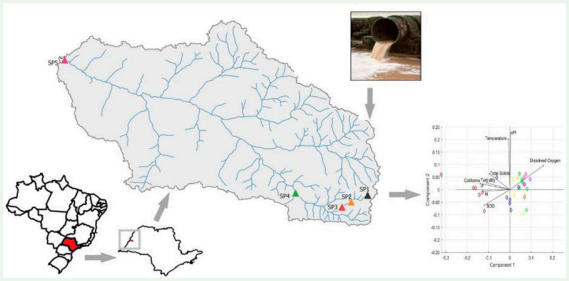
Assessment of water quality using principal component analysis: a case study of the Marrecas stream basin in Brazil

Alexandre Teixeira de Souza [a], Lucas Augusto T. X. Carneiro[b], Osmar Pereira da Silva Junior[c], Sérgio Luís de Carvalho[c] and Juliana Heloisa Pinê Américo-Pinheiro[c,d]

[a]Department of Environmental and Sanitary Engineering, Western São Paulo University (UNOESTE), Presidente Prudente, Brazil; [b]Mechanical Engineering Department, McGill University, Montreal, Canada; [c]Post-graduate Program in Civil Engineering, School of Engineering, São Paulo State University (UNESP), Ilha Solteira, Brazil; [d]Post-graduate Program in Environmental Sciences, Brazil University (UNIVBRASIL), São Paulo, Brazil

ABSTRACT

Monitoring water quality is a fundamental process to ensure proper anthropogenic usage and environmental protection of this resource. This study collected monthly measurements of 9 parameters (pH, Temperature, BOD, Total Solids, Thermotolerant Coliforms, Dissolved Oxygen, Total Nitrogen and Total Phosphorus) in 5 sampling stations along the Marrecas water stream, during a 1-year period. Temporal and seasonal variations were analyzed and interpreted for each element, explaining how specific geographical and anthropogenic factors affected the water body. Principal Component Analysis (PCA) was applied to evaluate each element's correlation and to reduce the number of parameters, easing the assessment of water quality for each location. Results were followed by the creation of an improved index for the region, which could better estimate the quality of water, only considering 4 of the original parameters. It was also recognized that each water body possesses several subtleties that impact on how its water quality should be measured and indexed into a single value, which validates the case for the creation of regional WQI's.



ARTICLE HISTORY
Received 12 November 2019
Accepted 6 April 2020

KEYWORDS
Dissolved oxygen; nitrogen; phosphorus; thermotolerant coliforms

1. Introduction

Aquatic ecosystems are extremely vulnerable to human activity, and their contamination impacts all kinds of life on the planet, causing ecological, social and economic damage to the affected region. Several studies analyzed the negative effects resulting from surface water pollution: diseases in both humans [1] and fishes [2], water shortages, decrease in fishing activities and increase in water treatment costs [3].

The current population growth fuels a constant increase in water pollution levels, and agricultural, industrial and domestic usage already accounts for more than one-third of renewable freshwater's consumption [4]; thus, constant supervision of water bodies is necessary to ensure sustainable usage. However, many countries, especially developing and sub-developed ones, lack an adequate monitoring system [5], concealing the damage extension and depriving the area of proper treatment.

Water resources have been extensively classified by Water Quality Indexes (WQI). A WQI assesses certain physical, chemical and biological parameters, quantitatively expressing the bulk of data into a single logical value, thus promoting effective communication to lawmakers and population [6]. Crescent urbanization and industrialization of the country in the past decades have compromised water bodies, and the goal of this WQI is to measure the water quality for public supply. However, this specific indicator disregards aquatic life

CONTACT Alexandre Teixeira de Souza alteiso@hotmail.com

protection and other characteristics. The present study follows the same premise, assessing the quality of water for human consumption.

The analyzed parameters (pH, Thermotolerant Coliforms, Biochemical Oxygen Demand, Total Nitrogen, Total Phosphorus, Temperature, Turbidity, Total Solids and Dissolved Oxygen) had been chosen by São Paulo's Environmental Company (CETESB) by virtue of their representability of wastewater effluent contamination [7]. DO and BOD are valid representatives of organic matter present in surface water, which may lead to the growth of undesired algae microbial life. Nitrogen and Phosphorus are common nutrients that in excess can also cause harmful life to develop. Total Solids and Turbidity are related to both natural causes, such as soil erosion particles, and anthropogenic action arising from effluent and litter discard. Temperature and pH affect chemical reaction rates and gas solubility, which in turn may create imbalances in the ecosystem; the latter is also associated with river eutrophication. At last, Thermotolerant Coliforms are a class of pathogens whose monitoring is essential to avoid human contamination. Thus, it is seen that this indicator tries to embrace multiple variables related different issues concerning groundwater quality.

Accounting for temporal and seasonal changes in the environment demands for several consecutive measurements, resulting in large volumes of data. Additionally, different regulators around the world select and weigh their parameters based on specific geographical characteristics and constraints [8]. These practices hinder both collection and analysis of data, so certain statistical tools have been widely used to handle inputs in the environmental science field [9–11].

One of such methods is Principal Component Analysis (PCA), a statistical tool of feature extraction, in which fewer independent variables are created from a combination of the original parameters. This process facilitates the visualization of obtained data, and how they interrelate [12]. Zeinalzadeh and Rezaei [11] employed PCA to extract the most important indicators in water samples of the Shahr Chai River in Iran and conceived a two-parameter index that better detected variations in river conditions when compared to the standard NHSWQI. Simeonov [10] assessed water quality in Northern Greece by using PCA to explore the relationship between several parameters. It was indicated that PCA was a valid method for processing large data sets and effective for creating analytical protocols. Similarly, Mahapatra et al. [13] applied PCA to classify water samples into different categories for subsequent management planning. Toledo and Nicolella [14] applied Factor Analysis to discuss parameter selection to evaluate changes in water quality in Brazil. Shrestha and Kazama [9] applied different statistical techniques in the Fuji river in Japan to evaluate temporal/spatial changes and was able to achieve a parameter reduction from 12 to 6 indicators used for quality analysis.

Alternative statistical methods are available and have been used for evaluating water quality. Cluster Analysis (CA) groups homogeneous objects/patterns based on pre-selected group-building criteria. Azhar et al (2015) and Fathi et al. [15] used CA to classify similar sampling stations according to their system characteristics. Factor Analysis (FA) approaches data reduction by identifying latent variables (factors) to explain covariance, thus describing the original parameters in terms of a linear combination. Amadi [16] used FA to reveal sources of groundwater pollution in a dumpsite, differentiating between natural and anthropogenic causes. Joung [17] introduced FA for developing a weight scale based on the correlation between studied parameters.

Results were used to produce a water quality index. Another multivariate technique is Discriminant Analysis (DA), which constructs a discriminant function to evaluate objects in a population and allocates them in groups known a priori. Shrestha and Kazama [9] selected groups based on spatial and seasonal elements to evaluate the causes of parameter variation. Although all methods permit dimensional reduction, often FA, CA, and DA are used when there is greater interest in examining and interpreting the association among the variables, while PCA is practiced when the objective is to emphasize data reduction despite losing some perception [1].

This study makes use of Principal Component Analysis to remove highly correlated parameters from the WQI. Expanding from previous literature, we aim to advance on the application of PCA to validate the formulation of a local water quality index that better reflects the intricacies of a particular water body, simplifying water quality analysis (in terms of costs and time) while retaining similar efficiency. Furthermore, this research provides a spatial classification relating the region's activity zones to their predominant indicators.

Aside from focusing on parameter reduction, PCA was chosen for this study due to its relatively straightforward implementation that can be better used for standardized procedures when compared to other methods which may have many different algorithms.

2. Methods

2.1. Study area and sampling points

The water basin under study (Figure 1) is located in the Western region of São Paulo state and it has an area of

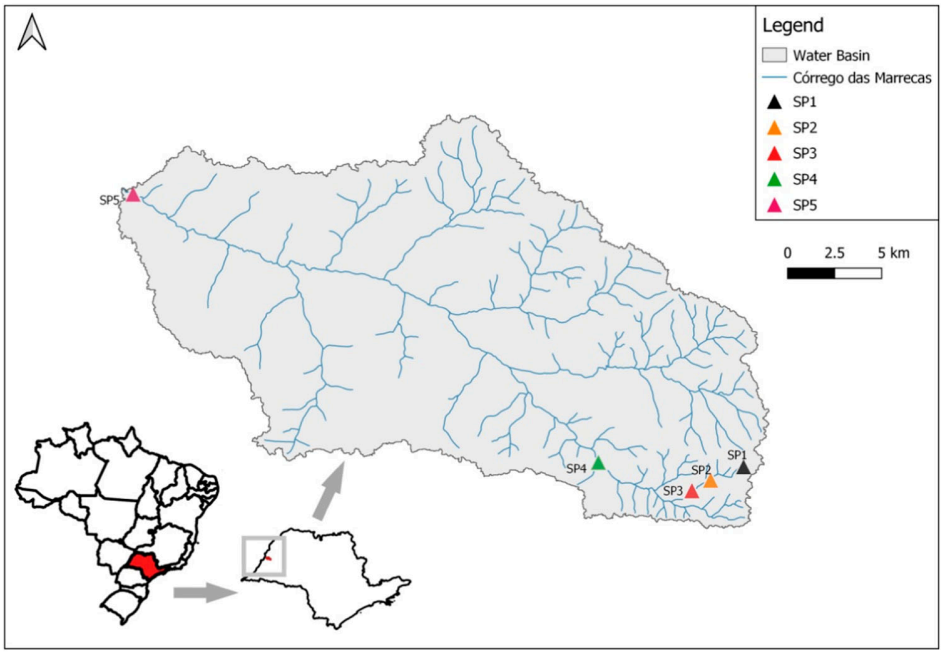


Figure 1. Map of the studied region. Satellite Image retrieved from ASTER GDEM V2 and further manipulated on QGIS software. ASTER GDEM is a product of Japan's Ministry of Economy, Trade, and Industry (METI) and NASA.

approximately 488.8 km^2. The major stream in this basin is the Marrecas Stream, approximately 45 km long. It has its source in the urban perimeter of the municipality of Dracena and its mouth in the city of Panorama. The region has a tropical climate, in which rain prevails during summer; the average temperature is 22.1°C and the average annual rainfall is 1204 mm [18]. This watershed has a high index of anthropogenic action, consisting of urban areas, temporary crops and especially cattle farms, sugarcane, eucalyptus and rubber tree production. Most of the riparian forest of the main course is in an advanced stage of degradation [19].

Five sampling points were strategically selected to map and evaluate human action along the Marrecas stream. Sampling Point 1 (SP1) is located at the stream's source, in the urban perimeter of Dracena. Flooding is common during rainy season, as drained water converges to this point. This source is channeled through a gallery to the rural area of Dracena, also leading sewage to the wastewater treatment plant (WWTP) of the municipality. SP2 is located upstream of the treated effluent from the sewage treatment plant. This point was chosen to determine the water quality prior to the effluent discharge from the WWTP and to subsequently assess the changes in water quality caused by the effluent discharge. Sampling Point 3 (SP3) is situated at the launching site of treated effluent from the WWTP and is characterized by the frequent presence of solid waste and debris on the stream banks caused by improper trash discard. Sampling Point 4 is in a rural environment 7.5 km downstream of SP3. High livestock presence and lack of riparian forest cause soil erosion and siltation of the water stream in this area. In SP5 the Marrecas Stream flows into the Paraná River, which is widely used as a fishing and recreational area.

2.2. Data collection

A total of 60 water samples were collected from the 5 different stations during wet and dry seasons from April 2017 to March 2018. Parameter measurements were conducted according to the Standard Methods for the Examination of Water and Wastewater [20]. Specific equipment and method for each parameter are displayed on Table 1.

2.2. Water quality index

CETESB has adapted the standard Water Quality Index created by the National Sanitation Foundation in the USA to create their own indicator, which has become the main index used in Brazil since 1975. The CETESB-WQI, known as IQA, is calculated as a weighted product of nine parameters (pH, coliforms, BOD, Total Nitrogen, Total Phosphorus, Temperature, Turbidity, Total Solids and Dissolved Oxygen) qualities [7]. The

Table 1. Methods and equipment to collect results.

Parameters	Unit	Method	Equipment
Turbidity	N T U	Nephelometric	Turbidimeter
Temperature	°C	Electrometric	Multiparameter probe
pH	–	Electrometric	pH meter
Total phosphorus	mgL^{-1}	Acid persulfate digestion	PhosVer3 digestor/ spectrophotometer
Total nitrogen	mgL^{-1}	Acid persulfate digestion	Digestor/spectrophotometer
Biochemical oxygen demand (BOD)	mgL^{-1}	Respirometric/ manometric	OxiTop BOD incubator
Dissolved oxygen (DO)	mgL^{-1}	Electrometric	Multiparameter probe
Total solids	mgL^{-1}	Gravimetric	Muffle furnace/water bath/Kiln
Thermotolerant coliforms (*E. coli*)	M P N/ 100 mL	Lactic acid bacteria count	3M petrifilm *E. coli* count plate

mathematical expression is given as:

$$IQA = \prod_{i=1}^{n} q_i^{w_i}$$

In which q_i is the quality of the i-th parameter found in the variation curve, and w_i is the weight of this parameter. These values were extracted from the CETESB – Water Quality Index documentation.

The IQA is graded according to CETESB's classification scheme in Table 2.

2.3. Principal component analysis

The nine original parameters will be combined into orthogonal (independent) components to explain their relationships for both dry and rainy periods. Then, a few Principal Components are selected, reducing the space dimension of the multivariate set. All initial values are considered, and data is projected on a small number of core components. The parameters have contrasting units; so it is recommended to normalize these values to remove any substantial variance [21]. This standardization is performed so that each variable has an arithmetic mean of 0 and a standard deviation of 1. To obtain the normalized results, each parameter x is modified to its normal score (z):

$$z = \frac{x - \mu}{\sigma}$$

Where μ is the arithmetic mean and σ is the standard deviation of each parameter.

Table 2. CETESB's IQA classification.

Index	Classification
79<IQA≤100	Excellent
51<IQA≤79	Good
36<IQA≤51	Regular
19<IQA≤36	Bad
0<IQA≤19	Very bad

The correlation matrix is computed to check the relationship between each parameter, creating a normalized matrix inside the non-dimensional range [−1,1]. Values close to magnitude |1| have a strong correlation, while the positive or negative sign denotes the direction of this link. The Principal Component loadings are the eigenvectors extracted from the Correlation Matrix. The PC scores are a linear combination of the eigenvectors with the original data matrix [22], and represent the individual parameter's score on a given PC.

The Kaiser-Meyer-Olkin test is performed to ensure the suitability of the data set for factor analysis [23]. All processes were carried on MATLAB.

In this study, PCA is used not only to study and verify the relationship between each parameter, but also to discard a subset of redundant variables that does not affect much of the results, saving time, money and resources [24].

3. Results and discussion

3.1. Spatial and seasonal parameter variation

Human activities and natural cycles interfere in water quality, and to consider the water proper for human consumption, the Brazilian Ministry of the Environment created a resolution indicating the legal limits for each parameter considered in the IQA calculation. Temperature is an indirect agent in variation: an increase in temperature intensifies chemical reaction rates, decreases gas solubility [25], and in this region, is correlated to rain activities [26]. The surface water temperature presented acceptable values during the entire year.

Dissolved Oxygen (DO) is related to the presence of organic matter, in which higher quantities promote an increase of decomposing microorganisms, consequently increasing oxygen consumption. Hypoxia happens when the water body self-purifying capacity (from aeration and algae photosynthesis) is insufficient, directly damaging aquatic life [27]. DO concentration in SP3 presented

values below the minimum acceptable level; this condition may be associated with a low efficiency from the wastewater treatment plant in removing organic matter. SP1 is in an urban environment and likewise presented unsatisfactory figures, which could be related to improper discard of litter in the area. Seasonal variations can be spotted on all sampling points: lower concentrations occurred during dry season, since the higher temperatures decrease gas solubility. The results of this study are similar to other studies carried out in Brazil [28].

SP3 presented the highest values for total solids present in water, with half of the samples outside legal limits. Other points also presented values above legislation. SP5 (river mouth) is the only location that displays satisfactory results. The entrance of solids into water may be a natural process via erosion, or due to anthropogenic activities, such as wastewater and litter discard [29]. Additionally, along the Stream of Marrecas, there is little to no preservation of riparian forest, an efficient agent in retaining sediments from getting into water [30]. Surface runoff is the presumed factor that contributes to the greater presence of Total Solids during rainy season [31], varying from 168 mgL^{-1} in SP4 (after WWTP) to a high of 714 mgL^{-1} in SP3 (at WWTP), compared to148 mgL^{-1} in SP4 and 652 mgL^{-1} in SP3 during dry season.

All sampling points had acceptable pH values. SP4 had a minimum value of 6.18 during dry season and SP2 had the maximum of 8 during wet season. The lower numbers may be related to eutrophication caused by wastewater launch in SP4, resulting in abrupt proliferation of macrophytes and algae [32,33], whose decomposition contributes to oxygen deficit and hydrogen sulphide formation.

Turbidity is affected by the presence of solids in water [34]; the only sample that exceeded legal limits happened on SP3 (123 NTU) during September 2017, in which Total Solids level reached 652 mgL^{-1}. Seasonal changes were not noticed.

Thermotolerant coliforms presence is an indicator of enteric pathogens [35], reflecting the disposal of wastewater and/or animal waste into the water body [36,37]. Lower fecal contamination was noticed on SP5 (river mouth) during rainy season, in which a water volume increase may have further diluted its concentration, and values ranged from 2.0E02M P N/100 mL^{-1} to 1.1E03M P N/100 mlL^{-1}. The highest values for fecal coliforms were found on the wastewater treatment plant (SP3), ranging from 4.0E04M P N/100 mL^{-1} to 5.7E5M P N/100 mL^{-1}, also during rainy season. Contrary to what happened in SP5, flash floods were responsible for temporarily reducing the WWTP efficiency [38,39]. Sewage pipes pass through the same channel that canalizes the Stream of Marrecas source (SP1), and the large concentration of coliforms – that reached 3.8E5M P N/100 mL^{-1} may indicate leakages on site, especially during rain, and improper trash discard by locals.

Predominantly, all sampling points presented exceeding values for coliforms. Causes are related to the urban nature of the water stream source and the lack of riparian forest, which could help on re-establishing biological equilibrium [40].

Established limits for Total Nitrogen are dependent on the sample's pH values, while Phosphorus has a single value limit. In SP1, urbanization may be regarded as the main cause in creating excessive N and P loads in groundwater and surface water, arising from industries and mobile transportation [41]; [42]. Effluent launching from the WWTP on SP3 is directly related to the higher nutrient concentration, since the local WWTP has no tertiary treatment, responsible for Nitrogen and Phosphorus removal [25]. The clustering of agricultural activities in the region is also a significant contributor to Nitrogen and Phosphorus retention in the water stream, derived from fertilizer use and manure [43,44]. Eutrophication due to excess nutrients can cause harmful algal growth [45], disturbance of other components from the local aquatic ecosystem and release of toxic compounds [46]. Again, the presence of riparian forest along river corridors has been proved to be an efficient filter for pollution [47].

Elevated BOD levels evidence a large presence of organic matter in the Stream of Marrecas. During drought, values ranged from 10 mgL^{-1} in SP4 to 96 mgL^{-1} in SP3. In rain season, the minimum concentration was 11 mgL^{-1} in SP2 and the maximum was 92 mgL^{-1} in SP3. Lower summer averages, notably in SP1 and SP3, may be explained due to rain dilution and lower flow during drought (Table 3).

The CETESB Water Quality Index was determined for each of the sampling points. SP1 presented a 'bad' index for both seasons; SP2 and SP4 had acceptable figures, while SP3 displays results ranging from 'very bad' to 'bad'. Sampling Point 5, in the river mouth, was the only place with 'good' water quality (Table 4).

3.2. Statistical analysis

Principal Component Analysis was carried separately for the two seasons. Dry season had 61.04% of its components explained by PC1, 13.59% by PC2 and 6 Principal Components were necessary to explain 95% of the total variability, as seen on the screen plot (a) of Figure 2. For Rain Season the first Principal Component was responsible for 62.40% of the total variance and PC2 represented 13.69% of the observations and 5 PCs

Table 3. Average of parameter measurements and legal limits according to CONAMA Legislation 357/05 [48] for class II.

Parameter	Sampling point	Dry season (Avg + SD)	Rain season (Avg + SD)	Legal minimum	Legal maximum
pH	1	7.37 ± 0.28	7.25 ± 0.44	6	9
	2	7.55 ± 0.40	7.48 ± 0.28		
	3	7.25 ± 0.32	7.25 ± 0.21		
	4	7.29 ± 0.55	7.29 ± 0.29		
	5	7.25 ± 0.29	7.25 ± 0.20		
Coliforms (MPN)	1	3.32E04 ± 2.35E04	9.03E04 ± 1.52E05	–	1000/1000 mL
	2	2.78E03 ± 1.85E03	1.82E03 ± 1.15E03		
	3	7.85E04 ± 7.65E04	1.86E05 ± 2.11E05		
	4	1.63E03 ± 1.51E03	7.00E02 ± 5.25E02		
	5	1.52E03 ± 2.33E03	6.33E02 ± 3.33E02		
BOD (mg L^{-1})	1	50.17 ± 21.53	35.17 ± 5.53	–	5
	2	18.00 ± 1.67	21.33 ± 6.38		
	3	83.33 ± 10.88	76.83 ± 11.67		
	4	17.50 ± 5.68	17.33 ± 3.20		
	5	16.17 ± 2.93	15.17 ± 4.17		
Total nitrogen (mg L^{-1})	1	7.5 ± 2.74	5.17 ± 2.93	–	3.7 for pH<7.5 2.0 for 7.5<pH<8.0 1.0 for 8.0<pH< 8.5
	2	1.83 ± 0.98	2.00 ± 0.89		
	4	5.00 ± 3.29	3.67 ± 1.97		
	5	2.50 ± 1.05	2.00 ± 1.10		
Total phosphorus (mg L^{-1})	1	1.21 ± 0.98	1.62 ± 1.17	–	0.1
	2	0.53 ± 0.31	0.65 ± 0.39		
	3	5.77 ± 1.77	6.40 ± 0.83		
	4	1.39 ± 0.89	0.97 ± 0.52		
	5	0.48 ± 0.21	0.54 ± 0.28		
Temperature (°C)	1	21.74 ± 0.94	23.15 ± 0.54	–	40
	2	22.38 ± 1.05	24.55 ± 0.84		
	3	23.21 ± 1.27	25.08 ± 0.86		
	4	22.95 ± 1.10	25.19 ± 0.77		
	5	24.41 ± 0.78	26.48 ± 0.47		
Turbidity (NTU)	1	25.52 ± 14.96	23.85 ± 6.52	–	100
	2	21.80 ± 16.53	22.00 ± 7.13		
	3	76.50 ± 25.67	63.33 ± 5.32		
	4	21.62 ± 9.92	14.50 ± 4.04		
	5	15.83 ± 2.48	16.00 ± 4.05		
Total solids (mg L^{-1})	1	318.00 ± 84.65	427.17 ± 90.59	–	500
	2	276.83 ± 84.78	421.67 ± 72.81		
	3	425.67 ± 134.86	588.83 ± 107.30		
	4	232.33 ± 54.45	394.50 ± 108.09		
	5	223.67 ± 56.69	288.67 ± 101.14		
Dissolved oxygen (mg L^{-1})	1	4.25 ± 0.33	4.84 ± 0.15	5	–
	2	5.59 ± 0.69	6.12 ± 0.47		
	3	2.22 ± 0.26	2.74 ± 0.39		
	4	4.91 ± 0.36	5.39 ± 0.73		
	5	5.97 ± 0.87	7.12 ± 0.39		

explained more than 95% of the variability. The elevated percentage of PC1 is already an indicator that the parameters are highly correlated and, in some circumstances, redundant.

Table 4. Average values of IQA for dry and rain season calculated for each sampling point.

	IQA dry	IQA rain
SP1	28.2	30
SP2	48.3	46.3
SP3	18.8	19.3
SP4	44.8	47.6
SP5	54	57.3

The Kaiser criterion [49] was applied to select Principal Components with eigenvalue > 1 for subsequent analysis. The eigenvalues on Table 5 provides this information, and based on it, PC1 and PC2 of each season will be used.

Biplots of both periods can be seen in Figure 3. PC1 mostly characterizes Dissolved Oxygen, Total Solids, Turbidity, Coliforms, P, N, and BOD. This organic component can be interpreted as a representation of direct and indirect anthropogenic action in the form of wastewater, industrial and agricultural pollution, as well as from the degradation of the riparian forest. The second PC mainly loads physicochemical parameters (pH and

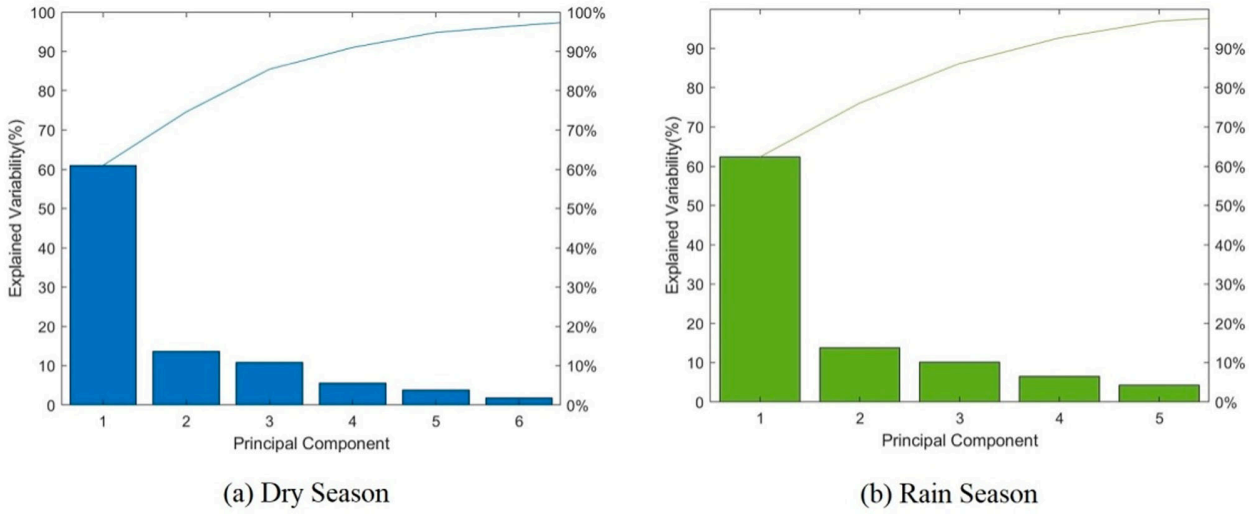


Figure 2. Scree plots of principal components and variability explained.

Table 5. Eigenvalues of each principal component, for both seasons.

	Eigen value (Dry)	Eigen value (Rain)
PC1	5.4935	5.6160
PC2	1.2230	1.2319
PC3	0.9748	0.9061
PC4	0.5026	0.5889
PC5	0.3419	0.3848
PC6	0.1580	0.1130
PC7	0.1386	0.0844
PC8	0.1032	0.0437
PC9	0.0640	0.0306

Temperature), and in its current form, is relatively independent of the non-point sources cited for PC1. pH is indeed affected by point sources and elements not identified/measured in this research; it is possible to notice a change in trends from Rain to Dry Season, in which the regular negative correlation between Temperature and pH is inverted. This phenomenon may happen due to increased algae photosynthesis in higher temperatures, which consequently increases the local pH [50] or, as cited before, due to variation in other compounds such as salts, heavy metals, and other industry effluents and/or soil leaching. The position of the points now clearly indicates what was previously implied. Sampling Point 3(after receiving WWTP effluents) has the lowest scores on PC1, staying closer to pollution indicators (Coliforms, Turbidity, P, N, BOD and Total Solids), while SP5 (river mouth) is distant

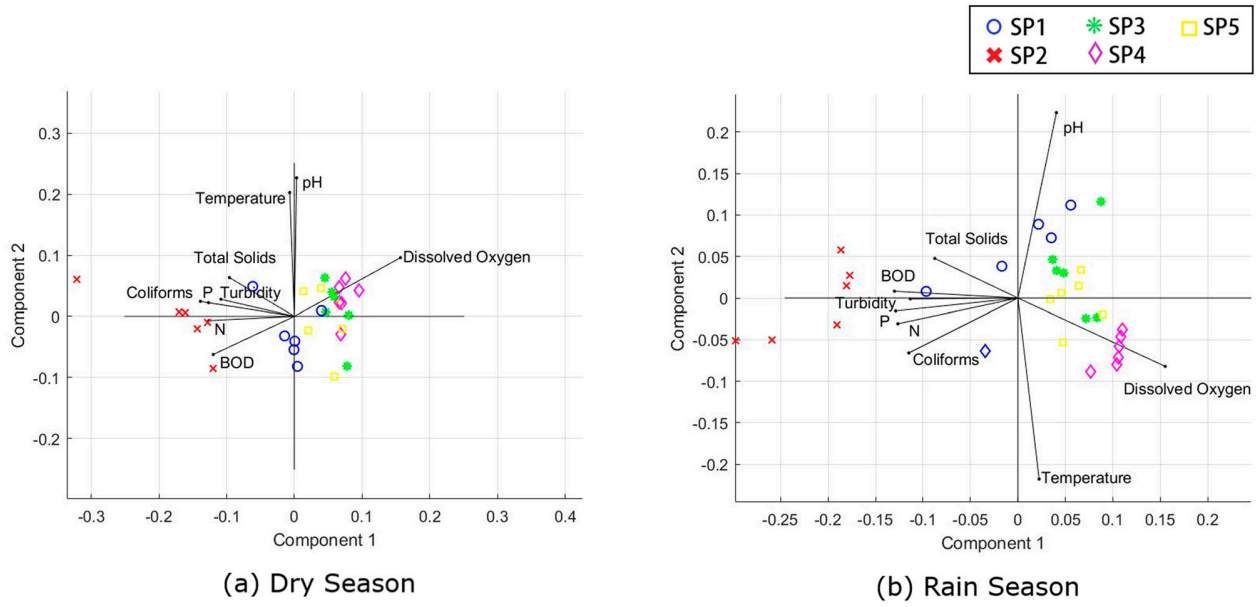


Figure 3. Biplots of PC1 × PC2 for SP1, SP2, SP3, SP4 and SP5 separated by season.

Table 6. Correlation matrix between the nine parameters for dry season.

	Turbidity	Temperature	Nitrogen	Phosphorus	Total Solids	Dissolved Oxygen	BOD	Coliforms	pH
Turbidity	1	0.0452	**0.7792**	**0.8429**	**0.6733**	**−0.7303**	**0.7215**	**0.8201**	0.0928
Temperature		1	0.0549	0.1355	0.1948	0.1611	0.0199	−0.0219	0.0832
Nitrogen			1	**0.9049**	**0.6904**	**−0.8846**	**0.8460**	**0.6868**	−0.0077
Phosphorus				1	**0.6840**	**−0.8070**	**0.8015**	**0.7579**	0.0163
Total Solids					1	**−0.5873**	**0.6284**	**0.6981**	0.0649
Dissolved Oxygen						1	**−0.8550**	−0.6400	0.0925
BOD							1	**0.6249**	−0.2228
Coliforms								1	0.0387
pH									1

Highly correlated values (>0.5) are indicated in bold.

Table 7. Correlation matrix between the nine parameters for rain season.

	Turbidity	Temperature	Nitrogen	Phosphorus	Total Solids	Dissolved Oxygen	BOD	Coliforms	pH
Turbidity	1	−0.0373	**0.8887**	**0.9294**	**0.7109**	**−0.8196**	**0.9446**	**0.6065**	−0.1531
Temperature		1	−0.0533	−0.0542	−0.1610	0.3330	−0.1459	−0.1699	−0.1544
Nitrogen			1	**0.9586**	**0.5986**	**−0.8162**	**0.9185**	**0.6445**	−0.2889
Phosphorus				1	**0.6393**	**−0.8509**	**0.9396**	**0.6355**	−0.2246
Total solids					1	**−0.6386**	**0.6926**	0.3246	−0.1074
Dissolved oxygen						1	**−0.8578**	−0.4768	0.1668
BOD							1	**0.6702**	−0.1998
Coliforms								1	−0.3522
pH									1

Highly correlated values (>0.5) are indicated in bold.

from the latter and positively associated with Dissolved Oxygen levels. SP1, located in an urban area, is related to the unfavourable organic parameters. SP2 and SP4 overlap, as they share environmental similarities of the rural zone.

3.3. Parameter reduction

The correlation matrices arising from PCA are used to perform a preliminary parameter reduction. Discarding variables may present several methods not only in a pragmatic standpoint (reducing costs, equipment and time) but also by excluding redundant factors that may over-emphasize a determinate result. Tables 6 and 7 acknowledge the interaction between the same sets of parameters (values in bold).

To create a minimized Water Quality Index, 4 global parameters were chosen. It is not the intent of this study to determine the most important variables, but to elect a few based on their representatives. pH was chosen due to its independence from other parameters, simplicity of measurement and as an indicator of biomass nature; Dissolved Oxygen as an indicator of aquatic life health and eutrophication, while also having strong correlation to Turbidity, Nitrogen and Phosphorus; Total Solids as a reference to erosion and effluent discharge, and Coliforms as a biological indicator and disease control. Although BOD would provide great representability for other parameters, it was opted out because its results take longer (5 days on average) to be determined, weakening the straightforward purpose of this reduced index. The IQA_{min} is created based on the same formula, but with equal weights (0.25) for each parameter. The results for the year analyzed are shown on Table 8.

Table 8. Reduced water quality Index considering four parameters: pH, dissolved oxygen, total solids and coliforms.

	IQA_{min} dry	IQA_{min} rain
SP1	26.8	24.6
SP2	40.2	39.3
SP3	18.9	17.7
SP4	41.2	41.4
SP5	44.7	49.9

The results presented above contain a more conservative behaviour; in general, the lower values express worse water quality than as indicated before. IQA_{min} proved itself as a better indicator of excessive nutrients (TN, TP) and coliforms, extremely important when considering the water for human consumption. Contrary to the regular IQA, the reduced IQA showed consistency at illustrating a general quality decrease during rainy season for the urban point (SP1) and the WWTP effluent area (SP3), as noticed when evaluating individual parameter values. It also highlights the inverted situation when nature is present: during rain periods, SP4 and SP5 raised in quality. Added benefits from a tailored quality index demonstrate that local researchers and organizations may customize their indicators based on anthropogeographical aspects of the region; however, it is relevant to remark that other non-utilized parameters play a critical role in many aspects for the study,

monitoring and conservation of water bodies, and should not be disregarded.

4. Conclusion

The present study addresses the necessary steps for the creation of a water quality index to evaluate water bodies around the world. Using Principal Component Analysis, the nine parameters were divided into 'organic' and 'physicochemical' factors. Within this statistical technique, the most representative parameters were extracted from the original WQI to create a reduced index IQA_{min} consisting of Dissolved Oxygen, pH, Total Solids and Coliforms. It was successfully demonstrated that advantages might arise from a custom-built index, better assessing anthropogenic impact. The method could be applied especially in developing countries where resources may be lacking, and/or during exceptional circumstances in which an index must be quickly calculated, leaving out time-consuming components (such as BOD).

Narrowing down the number of parameters used to asses water quality can be accurate and resource-saving, but this task is highly dependent on the studied region and requires knowledgeable judgment when selecting indicators.

Lastly, this research emphasizes how human pollution affects the environment, contemplating seasonal and spatial changes in a range that involves urban areas, agricultural fields, wastewater treatment facilities and rural zones.

Acknowledgements

The authors would like to thank the research funding agency CAPES for the scholarships granted to the post-graduate student participating in the study.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Alexandre Teixeira de Souza http://orcid.org/0000-0003-0357-0925

References

[1] Härdle WK, Simar L. Applied multivariate statistical analysis. 4th ed. Berlin: Springer Berlin Heidelberg; 2015.
[2] Sindermann CJ. Pollution-associated diseases and abnormalities of fish and shellfish: a review. Fish Bull. 1979;76:4.
[3] Carvalho AR, Schlittler FHM, Tornisielo VL. Relações da atividade agropecuária com parâmetros físicos químicos da água. Quím Nova. 2000;23(5):618–622.
[4] Schwarzenbach RP, Escher BI, Fenner K, et al. The challenge of micropollutants in aquatic systems. Science. 2006;313(5790):1072–1077.
[5] UNESCO. Water in a changing world: the United Nations world water development report, volume 3. Paris: UNESCO; 2009.
[6] Brown RM, Mcclelland NI, Deininger RA, et al. A water quality index – crashing the psychological barrier. Ind Environ Qual. 1972;1(1):173–178.
[7] Companhia de Tecnologia de Saneamento Ambiental - CETESB. Índices de qualidade das Águas São Paulo, 2017. https://cetesb.sp.gov.br/aguas-interiores/wp-content/uploads/sites/12/2017/11/Apêndice-D-Índices-de-Qualidade-das-Aguas.pdf.
[8] Tyagi S, Sharma B, Singh P, et al. Water quality assessment in terms of water quality index. Am J Water Resour. 2013;1(3):34–38.
[9] Shrestha S, Kazama F. Assessment of surface water quality using multivariate statistical techniques: a case study of the Fuji river basin, Japan. Environ Model Softw. 2007;22(4):464–475.
[10] Simeonov V, Stratis J, Samara C, et al. Assessment of the surface water quality in Northern Greece. Water Res. 2003;37(17):4119–4124.
[11] Zeinalzadeh K, Rezaei E. Determining spatial and temporal changes of surface water quality using principal component analysis. J Hydrol Region Stud. 2017;13:1–10.
[12] Jolliffe I. Principal component analysis. International encyclopedia of statistical science. New York: Springer; 2011.
[13] Mahapatra SS, Sahu M, Patel RK, et al. Prediction of water quality using principal component analysis. Water Qual Expos Health. 2012;4(2):93–104.
[14] Toledo LG, Nicolella G. Índice de qualidade de água em microbacia sob uso agrícola e urbano. Sci Agric. 2002;59(1):181–186.
[15] Fathi E, Zamani-Ahmadmahmoodi R, Zare-Bidaki R. Water quality evaluation using water quality index and multivariate methods, Beheshtabad River, Iran. Appl Water Sci. 2018;8(7):210.
[16] Amadi AN. Assessing the effects of Aladimma dumpsite on soil and groundwater using water quality index and factor analysis. Aust J Basic Appl Sci. 2011;5(11):763–770.
[17] Joung HM, Miller WW, Mahannah CN, et al. A generalized water quality index based on multivariate factor analysis. J Environ Qual. 1979;8(1):95–100.
[18] Climadate. Clima Dracena. (2018). [cited 26 Jul 2019]. https://pt.climate-data.org/location/34869/.
[19] Lelis LRM, Hespanhol RAM. Dinâmica agropecuária do município de Dracena – SP: da cafeicultura à cana-de-açúcar. Geogr Quest. 2013;6:2.
[20] Rice EW, Baird RB, Eaton AD, et al. Standard methods for the examination of water and wastewater. 22nd ed. Washington (DC): American Public Health Association (APHA), American Water Works Association (AWWA) and Water Environment Federation (WEF); 2012.
[21] Dobbie MJ, Dail D. Robustness and sensitivity of weighting and aggregation in constructing composite indices. Ecol Indic. 2013;29:270–277.

[22] Jolliffe IT, Cadima J. Principal component analysis: a review and recent developments. Philos Trans Roy Soc Math Phys Eng Sci. 2016;374:2065.
[23] Jackson DA. Stopping rules in principal components analysis: a comparison of heuristical and statistical approaches. Ecology. 1993;74(8):2204–2214.
[24] Jolliffe IT. Discarding variables in a principal component analysis. I: artificial data. J Roy Statis Soc Ser C. 1972;21 (2):160–173.
[25] Von Sperling M. Introdução à qualidade das águas e ao tratamento de esgotos. Belo Horizonte: Escola de Engenharia DESA-UFMG; 1996.
[26] Primavesi O, Freitas AR, Oliveira HT, et al. A qualidade de água na microbacia hidrográfica do Ribeirão Canchim, São Carlos, SP, ocupada por atividade pecuária. Acta Limnol Bras. 2000;12(1):95–111.
[27] Yin K, Lin Z, Ke Z. Temporal and spatial distribution of dissolved oxygen in the Pearl river Estuary and adjacent coastal waters. Cont Shelf Res. 2004;24(16):1935–1948.
[28] Américo JHP, Previato V, Carvalho SL. Qualidade da água de uma piscicultura em tanques-rede no rio São José dos Dourados, Ilha Solteira – São Paulo. Periód Eletr Fórum Ambient Alta Paul. 2013;9:2.
[29] Fundação Nacional de Saúde (FUNASA). Manual de controle da qualidade da água para técnicos que trabalham em ETAS / Ministério da Saúde, Fundação Nacional de Saúde. – Brasília, p. 112; 2014.
[30] Lima WP, Zakia MJB. Hidrologia de matas ciliares. Matas ciliares: conservação e recuperação. São Paulo: EDUSP/ FAPESP; 2000.
[31] Barros RVG, De Souza CA. Qualidade do recurso hídrico do córrego André Mirassol d'oeste, MT. Revis Bras Ciênc Ambient. 2012;24:1–16.
[32] Siqueira L, Barbieri R, Rojas M, et al. Bioensaio e estudo da decomposição de Ruppia Marítima L. da Laguna da Jansen, São Luís – MA (Brasil). Acta Tecnol. 2011;6(1):62–72.
[33] Varela ZCR, Troncone F, Sánchez J, et al. Nitrógeno y fósforo totales de los ríos tributarios al sistema lago de Maracaibo, Venezuela. Revis Ciencia Tecnol Am. 2009;34 (5):308–314.
[34] Hannouche A, Chebbo G, Ruban G, et al. Relationship between turbidity and total suspended solids concentration within a combined sewer system. Water Sci Technol. 2011;64(12):2445–2452.
[35] Skraber S, Gassilloud B, Gantzer C. Comparison of coliforms and coliphages as tools for assessment of viral contamination in river water. Appl Environ Microbiol. 2004;70 (6):3644–3649.
[36] Mitch AA, Gasner KC, Mitch WA. Fecal coliform accumulation within a river subject to seasonally-disinfected wastewater discharges. Water Res Shift Paradig Assess Recreat Water Qual. 2010;44(16):4776–4782.
[37] Vis C, Hudon C, Cattaneo A, et al. Periphyton as an indicator of water quality in the St Lawrence river (Québec, Canada). Environ Pollut. 1998;101(1):13–24.
[38] Capodaglio AG. Improving sewage treatment plant performance in wet weather. (J. Marsalek et al., Eds.) Enhancing urban environment by environmental upgrading and restoration. Nato Sci Ser IV Earth Environ Sci. 2004;43:175–185.
[39] He L, Tan T, Gao Z, et al. The shock effect of inorganic suspended solids in surface runoff on wastewater treatment plant performance. Int J Environ Res Public Health. 2019;16:3.
[40] Lowrance R, Leonard R, Sheridan J. Managing riparian ecosystems to control nonpoint pollution. J Soil Water Conserv. 1985;40(1):87–91.
[41] Bernhardt ES, Band LE, Walsh CJ, et al. Understanding, managing, and minimizing urban impacts on surface water nitrogen loading. Ann N Y Acad Sci. 2008;1134 (1):61–96.
[42] Kalmykova Y, Harder R, Borgestedt H, et al. Pathways and management of phosphorus in urban areas. J Ind Ecol. 2012;16(6):928–939.
[43] Mitsch WJ, Day JW, Zhang L, et al. Nitrate-nitrogen retention in wetlands in the Mississippi River basin. Ecol Eng. 2005;24(4):267–278.
[44] Petrone KC. Catchment export of carbon, nitrogen, and phosphorus across an agro-urban land use gradient, Swan-Canning river system, southwestern Australia. J Geophys Res Biogeosci. 2010;115:G1.
[45] Li H-M, Tang H-J, Shi X-Y, et al. Increased nutrient loads from the changjiang (Yangtze) river have led to increased harmful algal blooms. Harmful Algae. 2014;39:92–101.
[46] Smith VH, Tilman GD, Nekola JC. Eutrophication: impacts of excess nutrient inputs on freshwater, marine, and terrestrial ecosystems. Environ Pollut. 1999;100(1):179–196.
[47] Haycock NE, Pinay G, Walker C. Nitrogen retention in river corridors : European perspective. Nitrogen retention in river corridors. Eur Persp. 1993;22(6):340–346.
[48] Conselho Nacional do Meio Ambiente - CONAMA. Resolução n. 357, de 17 de março de. Dispõe sobre a classificação daságuas doces, salobras e salinas do território nacional. Brasília: Diário Oficial da União; 2005.
[49] Kaiser HF. The application of electronic computers to factor analysis. Educ Psychol Meas. 1960;20(1):141–151.
[50] Semesi IS, Kangwe J, Björk M. Alterations in seawater pH and CO_2 affect calcification and photosynthesis in the tropical coralline alga, Hydrolithon sp. (Rhodophyta). Estuarine. 2009;84(3):337–341.