Assessment of Spatial And Temporal Water Quality Distribution of Lake Ludas, Serbia

Zoltan Horvat

Univerzitet u Novom Sadu Gradevinski Fakultet Subotica

Mirjana Horvat

Univerzitet u Novom Sadu Gradevinski Fakultet Subotica

Kristian Pastor (✉ herr.pastor.kristian@gmail.com)

Univerzitet u Novom Sadu Teholoski Fakultet <https://orcid.org/0000-0003-0890-8171>

Research Article

Keywords: water quality assessment, Ludas lake, spatial and temporal analysis, data distribution, PCA, CCME WQI

Posted Date: August 9th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-637713/v1>

License: This work is licensed under a Creative Commons Attribution 4.0 International License. [Read Full License](#)

Abstract

This work presents the analysis of both spatial and temporal water quality distribution of Lake Ludas in the Republic of Serbia using water quality data from 2011 to 2018 at three different locations. By including a set of standard methods, we reduced the initial 15 water quality parameters to 7 parameters representative for the upcoming temporal and spatial considerations. The selected parameters were subject to a series of tests such as spatial and temporal analysis. Principal component analysis (PCA) was employed to present the variation of the measurements most efficiently and identify spatial and temporal tendencies. The PCA was expanded by the utilization of biplots providing a more comprehensive understanding of the measurements. Finally, the overall state of the lake's quality was evaluated using the Canadian Council of Ministers of the Environment Water Quality Index method for each sampling location, both annually and for the overall time interval, and as one representative value for the whole lake. The presented research lead to several conclusions, including the need for more detailed future measurements. It was shown that a reasonable monitoring approach leading to reliable conclusions should include much denser data in space and time. Furthermore, the necessity of three sampling locations remains relevant. In fact, it would be preferred to have a shorter list of monitored variables covering denser time and space data acquisition than having more diverse quality parameter evaluation at fewer locations or temporally sporadic measurements.

Introduction

Considering the overall declining water quality induced by anthropogenic activities, there is an increasing need for efficient water quality monitoring and management. Yadav et al. (2019) investigated the influence of various land uses on 16 water quality parameters in the Mun River, concluding that there is a strong link between the water quality parameters and land usage. The estimation of the overall quality of a water body is done using various approaches aiming to reduce a large number of data into a single number (Gradilla-Hernández et al., 2020). For example, a biological assessment of rivers was done using aquatic macro-invertebrates derived from the Tanzania River Scoring System (Dusabe et al., 2019). Computing the water quality index is another approach frequently utilized to evaluate water quality (Kükrer and Mutlu, 2019; Mutlu, 2019; Gradilla-Hernández et al., 2020; Radu et al., 2020). Others opt for hydrological simulations to provide the necessary current and future water quality information (Kumar, 2019). Models for flow and sediment transport simulations (Horvat et al., 2015), as well as heavy (toxic) metal transport simulations (Horvat and Horvat, 2016), can also be used to assess the water quality regarding sediment and/or other constituent content. However, one should keep in mind that the water quality is often described with a considerable number of inter-correlated parameters, making the task of evaluating the acquired data sets quite difficult. Consequently, the researchers opt for various approaches for establishing the most representative parameters to be utilized for a needed quality assessment. These methods include visual inspection of the data, evaluation of the Pearson's correlation between the measured parameters (Palaniet al., 2008; Radu et al., 2020), plotting of box and scatter plots, inspecting the most suitable distribution the given data will fit or implementing the principal component analysis (PCA) often used to provide a deeper insight into the water quality data (Olsen et al., 2012). A critical feature of the PCA is its potential to help identify possible clusters of the gathered measurements and examine the impact of water quality on other

environmental factors (Mandal et al., 2008). Researchers also used PCA to evaluate water quality monitoring stations (Ouyang, 2005), streamlining water quality management (Guo et al., 2018), and determining spatial and temporal changes in water quality (Zeinalzadeh and Rezaei, 2017).

The scope of this research was to implement exhaustive investigation and analysis on a series of temporally and spatially distributed data in order to establish a more reasonable monitoring approach for a shallow lake. Mutlu (2019) carried out a similar study, whose research included 11 points in space (sampling locations), which can provide excellent information considering the spatial distribution of water quality parameters. However, it had only one year of monthly data, making the temporal section of the research something that can still be enhanced. Since standard lake management approaches rely on comprehensive field and laboratory work, we aimed to implement exhaustive research that includes the consideration of spatial and temporal distributions of the water quality data while reducing the number of monitored parameters as much as possible, without significant loss of information. Lake Ludas in the Republic of Serbia was selected as an example of a shallow lake with systematic measurements both in space and time. Relying on standard approaches such as measurement visualization through the preparation of box plots and scatter plots, the use of Pearson's correlation coefficient, and inspection of numerous distributions and conducting distribution fit tests, to determine the most appropriate, we reduced the initial number of quality parameters from 15 to seven. Further data analysis was implemented relying on the PCA and preparation of biplots to identify potential grouping of the measurements. By establishing a reduced list of representative parameters for the evaluation of the water quality for Lake Ludas, we were able to simplify the quality water investigations, thus paving the road to measurements with increased frequency in space and time. Ultimately we employed a standard methodology for representing the water quality by computing the Canadian Council of Ministers of the Environment Water Quality Index (CCME WQI).

Materials And Methods

Study area

Lake Ludas is a Nature reserve that represents one of the Ramsar sites in the Republic of Serbia, located in northern Vojvodina, about 12 km from the city of Subotica. The lake is approximately 5 km long, stretching in the north-south direction, and from 1 km to 0.3 km wide, occupying an area of approximately 593 ha. There are two surface inflows into the lake. One is from the Palic-Ludas channel, entering the lake on its northwestern tip with an average annual discharge of roughly $0.3 \text{ m}^3/\text{s}$ and connecting it to Lakes Palic and Omladinsko (Fig. 1). In contrast, the other less significant inflow (considering the small discharge entering the lake this way) is from channel Kires on the lake's northern part. Seeing the inflow/outflow locations are positioned on the lake's northern region, it is justified to wonder how this influences the spatial distribution of the water quality. Furthermore, Lake Palic serves as a recipient of the treated wastewater from the Wastewater treatment plant Subotica (WWTPS), and is also mostly surrounded by agricultural land that drains into the lake. Finally, the Palic-Ludas channel is an unofficial recipient of wastewater from numerous homes surrounding Lake Palic and the channel itself. Considering all this, it is expected to see both spatial and temporal variations of the quality status in Lake Ludas.

Assessment Methodology

Another reason for selecting this location is the fact that Lake Ludas is a nature reserve site implying fairly regular measurements are conducted, covering a wide range of parameters for the time interval of the last eight years. Standard measurements conducted at Lake Ludas include a monthly water sampling at three different locations at the northern, middle, and southern parts of the lake, marked with diamonds on Fig. 1. In this research study, the 35 regularly assessed water quality parameters for over eight years, from 2011 until 2018, were firstly reduced to 15, which are presented in Table 1, along with the appropriate measurement methodologies employed.

In order to investigate which of the remaining parameters are representative for the upcoming quality analysis, the Pearson correlation coefficient (Hinkle et al., 2003) was computed for each of the quality parameter combinations. In this way, a relation between the considered parameters by evaluating both the strength and direction of the potential correlations, was obtained. Thus, the most appropriate parameters to represent each water quality characteristic were selected and reduced to seven.

These were further used in extensive data analysis, including spatial and temporal water quality distributions, the principal component analysis (PCA), and determining the water quality index relying on the Canadian Council of Ministers of the Environment Water Quality Index (CCME WQI).

Results

Evaluation of the water quality data

As previously mentioned, standard monitoring of Lake Ludas includes reasonably regular analysis of 35 water quality parameters. The main issue with these data is that over the years, amid various restructuring, some of the monitored parameters were replaced with others. Another inconvenience is that while some of these quality parameters are analyzed monthly, others are examined only four times a year. This is the consequence of the complexity of the overall location. Namely, as seen in Fig. 1, the complete area is made up of nine separate segments. Lake Palic incorporates seven of them. Lake Omladinsko is the eighth, and Lake Ludas the ninth. Hence, just one monthly sample within each segment would result in 315 analyzed data monthly, or 3780 annually. This inevitably leads to different attempts at reducing the number of tests being conducted. Therefore, if we were to conduct thorough data analysis, we would have to significantly reduce the number of data included in the research, either by reducing the variety of examined quality data or by decreasing the temporal frequency at least three times (from 12 data sets per year to 4). Considering this research aims to investigate temporal and spatial variations of the quality data, the Authors opted to reduce the variety of water quality parameters included in the research. This way, the initial number of 35 parameters was reduced to 15 parameters (Table 1). These are the measurements that were generally implemented regularly at a monthly pace.

Table 1
Methods and measuring uncertainties of the selected water quality parameters

No.	Parameter	Notation	Units	Method	Measuring uncertainty
1	Temperature	Temp.	°C	2250B*	1.15%
2	Power of hydrogen	pH	S/cm	500-H + B*	1.23%
3	Conductivity	Cond.	mg/L	SRPS EN 27888/2009	17%
4	Dissolved oxygen	DO	mg/L	SRPS EN 25813:2009/1/2011	19%
5	Dichromate chemical oxygen demand	COD _{Cr}	mg/L	EPA 410.4 US 5220D, ISO 15705	17%
6	Potassium permanganate chemical oxygen demand	COD _{Mn}	mg/L	4500-KMnO4*	11%
7	Five-day biochemical oxygen demand	BOD ₅	mg/L	SRPS EN 1899-1:2009	22%
8	Chlorophyll-a	Chl-a	mg/m ³	10200-H*	25%
9	Suspended sediments	SS	mg/L	2540D*	22%
10	Total phosphorus	TP	mg/L	SRPS EN ISO 6878:2008	13%
11	Orthophosphates	PO ₄ ³⁻	mg/L	SRPS EN ISO 6878:2008	13%
12	Total Kjeldahl nitrogen	TN	mg/L	4500-N.C*	3.5%
13	Nitrite nitrogen	NO ₂ ⁻ -N	mg/L	SRPS EN 26777:2009	21%
14	Nitrate nitrogen	NO ₃ ⁻ -N	mg/L	SRPS EN ISO 10304-1:2009	23%
15	Ammonia nitrogen	NH ₄ ⁺ -N	mg/L	**	2.2%
* Clesceri et al. (1999) ** Čoha (1990)					

Table 2 gives the descriptives, including the mean value, standard deviation (SD), minimum, median, and maximum value of the 15 quality parameters considered.

Table 2
Quality parameter characteristics

No.	Variable	Mean	SD	Minimum	Median	Maximum
1.	Temp	16.371	7.701	-4.30	16.20	29.50
2.	pH	9.034	0.660	7.06	9.05	10.63
3.	Cond.	1354.303	446.474	805.00	1253.00	4162.00
4.	DO	10.843	4.582	1.78	10.64	24.40
5.	COD _{Cr}	185.765	111.819	31.00	155.00	860.00
6.	COD _{Mn}	26.350	12.892	6.36	24.59	92.43
7.	BOD ₅	58.076	45.731	3.00	46.00	206.00
8.	Chl-a	382.078	364.549	0.10	273.00	2422.00
9.	SS	73.890	112.143	0.13	38.00	800.00
10.	TP	0.315	0.298	0.01	0.23	1.58
11.	PO ₄ ³⁻	5.215	8.806	0.00	0.06	49.37
12.	TN	16.938	8.413	3.29	15.54	49.46
13.	NO ₂ ⁻ -N	0.011	0.018	0.00	0.00	0.12
14.	NO ₃ ⁻ -N	0.266	0.166	0.05	0.24	0.90
15.	NH ₄ ⁺ -N	1.367	0.584	0.15	1.27	3.53

The next step in the study was to investigate which of the remaining parameters are representative for the upcoming quality analysis, by computing the Pearson correlation coefficient. The Pearson correlation coefficient r will have the value ranging from -1 , indicating a complete negative correlation of the considered parameters, to $+1$, suggesting their complete positive correlation. The value of r approaching zero illustrates the absence of correlation for the examined parameters.

Figure 2 presents examples of different quality correlations of the Pearson correlation, e.g., Fig. 2(b) displays the value of r for SS and Cond. where it is quite apparent these two quality parameters aren't correlated. On the other hand, Fig. 2(a) shows the correlation of SS and COD, and Fig. 2(c) presents the correlation for SS and Chl-a, where both show distinct positive correlations of the considered quality parameters.

Table 3 presents the correlation matrix for the 15 remaining quality parameters, where a correlation is considered significant if r is greater or equal $r = 0.5$. Significant correlations were noticed between pH and

Temp ($r = 0.526$), COD_{Cr} with BOD_5 ($r = 0.595$), Chl-a (0.648) and SS ($r = 0.546$), Chl-a with SS ($r = 0.549$) and TN ($r = 0.502$).

Table 3 The Pearson correlation table

	Temp	pH	Cond.	DO	COD_{Cr}	COD_{Mn}	BOD_5	Chl-a	SS	TP	PO_4^{3-}	TN	NO_3^- -N	NO_2^- -N	NH_4^+ -N
Temp	1														
pH	0.526	1													
Cond.	-0.202	-0.141	1												
DO	-0.096	0.011	-0.114	1											
COD_{Cr}	0.227	0.463	0.244	-0.085	1										
COD_{Mn}	0.045	-0.158	0.261	-0.052	0.138	1									
BOD_5	0.226	0.318	0.204	-0.022	0.595	-0.01	1								
Chl-a	0.142	0.442	-0.075	0.064	0.648	-0.013	0.432	1							
SS	0.164	0.357	-0.006	-0.015	0.546	-0.041	0.208	0.549	1						
TP	0.11	0.173	-0.047	0.272	0.39	0.019	0.368	0.605	0.432	1					
PO_4^{3-}	-0.071	-0.241	0.132	0.024	0.085	0.152	0.137	0.143	-0.035	0.325	1				
TN	-0.129	0.056	0.125	-0.053	0.288	0.38	0.121	0.502	0.175	0.06	0.008	1			
NO_3^- -N	-0.005	-0.054	-0.038	0.005	-0.059	-0.032	-0.111	-0.119	-0.044	-0.077	-0.018	-0.129	1		
NO_2^- -N	-0.009	-0.03	-0.049	-0.015	-0.045	-0.019	0.041	-0.063	-0.101	0.065	0.012	-0.1	-0.023	1	
NH_4^+ -N	-0.124	-0.176	0.254	-0.102	0.043	0.186	0.212	0.007	-0.041	0.119	0.171	0.15	-0.007	0.022	1

Further analysis helped us additionally reduce the number of parameters by including only one parameter from a group of available data for a considered characteristic. For example, the presence of organic matter in the water is described with COD_{Cr} , COD_{Mn} , and BOD_5 , yet including all of these data in further research would provide a biased appearance of the water quality. Thus, further studies that helped to select the most appropriate parameter to represent each water quality characteristic, were conducted. To pinpoint the most suitable parameters, the authors examined which distribution is best suited for the measured data for each of the considered groups separately, the phosphorus content, nitrogen, and organic matter, arguing that

such numerous water quality data should follow some type of distribution. This helped us differentiate the most reliable data in the group, and decide which should be discarded as less trustworthy. It should be pointed out that the representative distributions in this part of the study were not the best fitting ones for each of the compounds within a group; e.g., for TP, the best distribution was the Frechet distribution. At the same time, for PO_4^{3-} it was a mixture distribution. The selection was made to compare the results, meaning the same distribution had to be assigned for all compounds within a group. Consequently, the five best-fitting distributions for each compound in a group were considered and compared. After identifying a distribution that is common for all the data within a group, and is the best among the available, the distributions presented in Table 4 were selected: the Extreme value distribution for the nitrogen group; the Student T distribution for the phosphorus group; and the LogNormal distribution for the evaluation of the organic matter content.

Table 4
Distribution fit test results

Distribution fit test results								
Parameter	Distribution*	PCS +	CVM +	AIC -	BIC -	HQIC -	LogLik. +	Com. -
TN	Extreme Value	0.219030	0.726715	-6.730	-6.745	-6.746	-3.352	2
NO ₂ ⁻ N		0.000125	0.048696	2.557	2.542	2.541	1.292	2
NH ₄ ⁺ -N		0.273204	0.085680	0.332	0.318	0.318	0.178	2
NO ₃ ⁻ N		0.171210	0.515795	-1.948	-1.963	-1.963	-0.961	2
PO ₄ ³⁻	StudentT	0.000000	0.000000	-3.917	-3.939	-3.940	-1.938	3
TP		0.000021	0.067025	-0.264	-0.286	-0.286	-0.113	3
BOD5	LogNormal	0.413103	0.731957	-9.809	-9.824	-9.823	-4.893	2
COD _{Cr}		0.531894	0.735407	-11.580	-11.595	-11.594	-5.779	2
COD _{Mn}		0.004310	0.238261	-7.612	-7.627	-7.627	-3.793	2
* the + and – signs indicate whether higher test results are better (+) or lower (-)								

Additionally to the selected distributions, Table 4 presents the test results utilized for the decision process that includes the Pearson χ^2 (Chi-square) test (PCS), Cramer von Mises (CVM), Bayesian information criterion (BIC), Akaike information criterion (AIC), Hannan-Quinn information criterion (HQC), Log Likelihood (LogLik) and the complexity (Com). The PCS tests the goodness of fit, where the null hypothesis was that the data is drawn from a population with the considered distribution, while the alternative hypothesis was that the data was not drawn from the given population. The results are presented using the p value, with the significance level of 0.05, where small p values indicate a small chance that the data came from the considered distribution. The CVM test is also presented in the form of p values, with the same meaning as in the previous (PCS) case, where smaller p values mean there is a smaller probability the data is from the

considered distribution. The AIC is a test that includes the penalty for model complexity to protect against over-fitted models. Since it relies on the assumption of an infinite-sized sample, it is reliable only as a comparison test of various models instead as a general decision-making approach. Although for sample sizes less than 40, one can use the adjusted AIC formula, the considered dataset has a much larger size (over 100) and was analyzed using the standard AIC test, with lower results indicating more appropriate models. The BIC is also called the Schwarz information criterion, and much like AIC, it includes a penalty criterion against over-fitting models. A better model fit is indicated with smaller values of the results. The HQC is often utilized as an alternative to the AIC or BIC and measures the goodness of fit, where lower values of HQC imply either fewer explanatory variables, better fit of data, or both. The LogLik test also measures the model's goodness of fit, where higher values indicate a better fit. The Log Likelihood value ranges from negative infinity to positive infinity, making it inappropriate for drawing straightforward conclusions. Instead, it can be used to compare these values between various models and aid decision-making. Finally, the analysis includes the Com. indicating the complexity of the data distribution, where higher complexity is considered a negative characteristic, since it can lead to over-fitted models.

Based on the results, we selected the TN, TP, and COD_{Cr} as the most suitable parameters to represent the considered characteristics. After choosing the representative parameters, we further evaluated the measured data against the most appropriate distribution for the given method to analyze how well the data and the best distribution fit. The best-fitting distributions are not the same as the ones used for the selection of the representative parameters, since at this point, there was no need to accommodate various distributions anymore. The best distribution for COD_{Cr} and TN was the LogNormal distribution with distribution parameters $\mu_{COD}=4.9987$ and $\sigma_{COD}=0.5266$ for COD_{Cr} and $\mu_{TN}=2.6767$ and $\sigma_{TN}=0.4746$ for TN. The most suitable distribution for TP was the Frechtel Distribution with the distribution parameters $\alpha_{TP}= 1.8384$, $\beta_{TP}=0.2297$, and $\mu_{TP}=-0.0752$. Figure 3 displays the probability scale plots for these distributions, confirming a good fit of the data.

Regarding the selected distributions, we implemented the distribution fit test with the results given in Table 5, where it is clearly displayed the data match the selected distributions. Based on the given test results, it was safe to conclude the chosen parameters include reliable data and can provide a reasonable representation of the compound group, making them a sound choice for the upcoming evaluation.

Table 5
Distribution fit test results

Pair	COD _{Cr}		TN		TP	
Null hypothesis	The data is distributed according to the LogNormal distribution with parameters (4.9987,0.5266)		The data is distributed according to the LogNormal distribution with parameters (2.6767, 0.4746)		The data is distributed according to the Frechet distribution with parameters (1.8384, 0.2297, -0.0752)	
Test	Statistic	p value	Statistic	p value	Statistic	p value
Anderson-Darling	0.2481	0.9714	0.6153	0.6333	1.8801	0.1071
Cramer von Mises test	0.0433	0.9154	0.0861	0.6571	0.3376	0.1063
Pearson χ^2	12.0989	0.7371	22.8684	0.0624	32.9398	0.0048
Conclusion	The null hypothesis is not rejected at the 5 percent level based on the Cramer von Mises test		The null hypothesis is not rejected at the 5 percent level based on the Cramer von Mises test		The null hypothesis is not rejected at the 5 percent level based on the Cramer von Mises test	

The statistical significance for Pearson's correlation was assessed through the p values for the given correlations. The null hypothesis was that the correlation coefficient of the bivariate population is equal to zero, while the alternative hypothesis was that the correlation coefficient is not equal to zero. Keeping in mind higher p values (with the 5% limit) represent statistically insignificant correlations, suggesting strong evidence for the null hypothesis or the absence of correlation. These values helped us distinguish quality parameter pairs with reasonable correlations. After evaluating the results, the following parameters were selected to be included in the further analysis: Cond. and pH indicating the ions content of the sample; COD_{Cr} measuring all organic contaminants, including those that are not biodegradable; Chl-a as an indicator of algae; SS indicating the presence of minerals and organic substances.

The p values matching the Pearson coefficients for the selected parameters are given in Table 6, where upper index ** suggests the null hypothesis should not be rejected, meaning the pair's correlation may be zero.

Table 6
Table of p values for the Pearson coefficients

Pair	Pearson r	95% CI	p-value	Pair	Pearson r	95% CI	p-value
Cond., COD _{Cr}	0.263	0.097 to 0.415	0.0022	pH, Cond.	-0.109	-0.274 to 0.063	0.2126**
Cond., Chl-a	-0.072	-0.239 to 0.100	0.4122**	pH, COD-Cr	0.468	0.324 to 0.591	< 0.0001
Cond., SS	0.014	-0.157 to 0.184	0.8729**	pH, Chl-a	0.45	0.302 to 0.576	< 0.0001
Cond., TP	-0.018	-0.187 to 0.153	0.8388**	pH, SS	0.376	0.219 to 0.513	< 0.0001
Cond., TN	0.099	-0.072 to 0.265	0.2562**	pH, TP	0.2	0.031 to 0.358	0.0208
COD _{Cr} , Chl-a	0.64	0.528 to 0.731	< 0.0001	pH, TN	0.056	-0.115 to 0.224	0.5193**
COD _{Cr} , SS	0.549	0.418 to 0.658	< 0.0001	SS, TP	0.444	0.296 to 0.571	< 0.0001
COD _{Cr} , TP	0.395	0.241 to 0.530	< 0.0001	SS, TN	0.166	-0.004 to 0.327	0.0560**
COD _{Cr} , TN	0.273	0.108 to 0.423	0.0015	TP, TN	0.038	-0.133 to 0.207	0.6617**
Chl-a, SS	0.57	0.443 to 0.675	< 0.0001	Chl-a, TN	0.331	0.170 to 0.474	0.0001
Chl-a, TP	0.47	0.326 to 0.593	< 0.0001				

Spatial and temporal distribution of the water quality data

Hoping to understand better the causes and nature behind the spatial and temporal alterations, the selected water quality data was further assessed by constructing box plots, Figures 4, 5, and 6 with marked median, interquartile range (IRQ) as the range between the 25th (Q₁) and 75th (Q₃) percentile, maximum (as Q₃ + 1.5 IRQ), minimum (as Q₁ - 1.5 IRQ), and outliers. Figures 4 and 5 present the temporal distribution of the box plots for the seven selected water quality parameters for the three sampling locations separately.

A general overview of the data shows an improvement of the water quality during the time interval from 2013 to 2016 (possibly 2017). Although, it is not a distinct improvement apparent through all of the quality indicators, there is a noticeable tendency supporting this statement. As an example, we can single out a decrease of the concentration of nutrient elements (TN and TP) on both north and middle part of Lake Ludas in 2013, and a straightforward increase in 2017, Figures 4 (a), (b), (e) and (d). Matching trends are

noticeable with SS and Chl-a, given in Figures 4(j), (k), (m) and (n), showing decreased concentrations from 2013 until 2016.

On the other hand, the same parameters show conflicting information on the south part of the lake, Figures 4 (c), (f), (l), and (o). These different spatial tendencies can be explained as consequences of the inflow location of the Palic-Ludas channel, containing untreated wastewater as well as runoff from agricultural lands, and vicinity of the two sampling locations (north and middle), opposed to the remoteness of the south sampling location. Further examination supports this assumption, since all of the considered quality parameters have milder variations on the south, compared to the north and middle.

A more detailed examination of the data shows that in 2015 there was a significant rise in TP at the southern sampling location that can be detected on Figure 4(f), while according to Figures 4(d) and 4(e) the two other sites didn't endure such sudden changes in phosphorus concentrations. Considering the measurements are conducted only once a month, it is impossible to determine what caused this event with sufficient certainty. This type of conclusion could be made by a significant increase in the data collection frequency in space or time, potentially both, since information like that would provide significantly reliable information through a spatially and/or temporally denser data distribution. By examining Figures 5(a), (b), and (c) we can recognize higher conductivity in the southern part of the lake in all the examined years, also implying there is a significant spatial variation of the water quality between the considered locations. These spatial alterations of the water quality parameters also support the idea that additional water quality measurements are required.

Figure 6 displays the spatial distribution box plots jointly for the overall time interval from 2011 to 2018. This representation is used to eliminate the temporal influence of the data and help focus the attention onto the spatial variations. The results once again exhibit clear spatial distinctions of the water quality, affirming the previously noted poorer quality on the north part (Figures 6 (a), (b), (d), and (e)) that seems to be improving as one approaches the south part.

Considering that the current data are available only within the displayed locations, the only equitable explanation one can extract is that the systematic spatial quality alterations result from the inflow location of the Palic-Ludas channel on the north part of the lake. The influence of the Kires channel, or the effect the agricultural lands surrounding the lake have on its quality, would need to be monitored more thoroughly. This type of examination could be made by eliminating the temporal considerations and conducting extensive data sampling all around the lake, since it could provide an insight into the micro-distribution of the quality parameters around it. On the other hand, a more comprehensive understanding of temporal changes in the lake's water quality would require more frequent water sampling. For this purpose, the authors initially propose daily sampling of the lake for a shorter time interval (e.g., a couple of weeks). This type of data would provide understanding the frequency and intensity of temporal changes in the water quality data. Furthermore, these values could be associated with precipitation measurements to assess the influence of the weather on the quality variations as well. Subsequently, an even more frequent sampling should be conducted, e.g., hourly values, to identify short-term alterations of the quality data and the causes behind them. Seeing the restrained amount of information that can be drawn out of such long data series,

regardless of the spatially distributed data availability, the authors would suggest implementing the proposed much denser data measurements, at least once for any lake, to help better adapt the monitoring approach for the encountered circumstances.

Finally, Fig. 7 presents the scatter plot matrix providing a graphical representation of the correlations between pairs of the analyzed water quality parameters. The red lines mark the 95% density ellipses indicating the correlation between the two parameters, where narrower ellipses suggest a stronger correlation, such as the correlation between COD_{Cr} and Chl-a, and COD_{Cr} and SS, also supported by the Pearson's correlation coefficients given in Table 3. The different symbols employed on the scatter plot mark the locations (north, middle, and south) in Lake Ludas. Another interesting observation that can be made based on the scatter plot is that better the correlation between the considered parameters coincides with more distinct clustering of the data. For example, a higher correlation of previously mentioned COD_{Cr} and Chl-a, and COD_{Cr} and SS, also shows more pronounced grouping compared to TP and pH, or TP and TN, where both have low correlation coefficients, circle-shaped ellipses, and quite dispersed measured values.

Principal component analysis

The principal component analysis (PCA) boils down to finding new variables, called the principal components (PC), that capture as much of the data variation as possible (Pastor et al., 2016). The principal components are the linear combination of the analyzed variables (in this case, water quality parameters), computed so, that they are not correlated amongst each other, and so, that the first principal component captures the most of the original data variation, the second PC includes the next most variation, etc. When the analyzed data sets are presented in a graph where the principal components are the axes, similar data points will cluster together, which can be used to attain a deeper insight into the water quality parameter's nature (e.g., temporal, spatial, or other tendencies).

Table 7 gives the standardized variance where the sum of all variances equals the number of analyzed parameters (in this case, 7), the proportion of the variance contained in each PC, and the cumulative proportion of the standardized variance. We can see the first two components include 59% of the total variance, while the first three include over 72%.

Table 7
Variance of the principal components

Component	Variance	Proportion	Cumulative proportion
1	2.948	0.421	0.421
2	1.180	0.169	0.590
3	0.931	0.133	0.723
4	0.804	0.115	0.838
5	0.491	0.070	0.908
6	0.381	0.054	0.962
7	0.266	0.038	1.000

The proportion of the standardized variance in each PC is also presented by a scree plot on Fig. 8, where it is visually clearly displayed that the first PC encloses most of the data variance, as indicated with the apparent drop of variance between components 1 and 2. In contrast, the variances between all the subsequent components are minor in comparison.

The coefficients that define the principal components, as a linear combination of the analyzed water quality parameters, are listed in Table 8. The coefficients' higher absolute values imply a more significant impact of the water quality parameter on the deliberated PC.

Table 8
Coefficients of the principal components

Parameter	Principal component						
	PC1	PC2	PC3	PC4	PC5	PC6	PC7
pH	0.361	0.280	-0.037	0.709	-0.297	-0.426	-0.128
Cond.	0.038	-0.791	0.464	0.153	-0.058	-0.133	-0.337
COD _{Cr}	0.487	-0.228	0.122	0.194	-0.063	0.410	0.697
Chl-a	0.498	0.074	-0.175	-0.085	-0.064	0.574	-0.612
SS	0.454	0.076	0.122	-0.109	0.813	-0.317	-0.023
TP	0.363	0.175	0.360	-0.615	-0.477	-0.316	0.054
TN	0.211	-0.450	-0.770	-0.196	-0.110	-0.322	0.075

For instance, PC1 is relatively evenly influenced by three quality parameters, COD_{Cr}, Chl-a, and SS, with over 45%, following the impact of pH and TP with 36%. The next component, PC2, primarily consists of Con. and TN, where we can notice the significance of Cond. is more pronounced at 79% when compared to TN with

45%. A similar distribution is observed at PC3, with TN carrying 77% and Cond. 46%. In the case of PC4, pH encompasses almost 71% of it, while TP includes 61% of the effect.

Since the first two principal components contain nearly 60% of the data variance, a two-dimensional representation of the measurements is reasonable. After the computation of the principal components, the original data can be presented on biplots, presented on Figs. 9, 10, and 11, where the horizontal and vertical axes mark the first two principal components. The biplots also contain (angled) axes representing those water quality parameters whose variation is included in the first two principal components, as well as the percentage of their impact.

The biplot on Fig. 9 lets us recognize a specific clustering of the data regarding the sampling location (north, middle, and south), regardless of the sampling time. Nevertheless, some influence of the timing of the sampling is anticipated. Consequently, we provided two additional biplots, given in Figs. 10 and 11, showing potential grouping in terms of the timing of the data acquisition.

Figure 10 displays the accumulation of data depending on the sampling months for all three locations. In light of the large amount of evaluated data, we utilized the same symbols and colors by seasons, dividing each year into four groups of three months (January to March – group 1, April to June – group 2, July to September – group 3 and October to December – group 4). To help distinguish data within one group, we varied the sizes of the symbols in decreasing order, labeling the symbol of the first month in each group as the largest and the symbol of the last month as the smallest.

The analysis of the result given in Fig. 10 does indicate a certain grouping of data both through months and seasons, regardless of the location. Yet, the large number of data makes it hard to identify more definite characteristics. Ergo, we included the biplot presented in Fig. 11, that gives the monthly clusterings using analogous denotation, as on Fig. 10, but depicting only data for the north part of the lake. Based on these results, we were able to make some clear-cut conclusions. Namely, there is a distinct grouping of the data sampled in the first half of each year (January to June), that includes groups 1 and 2, marked with squares and circles and located on the left side of the biplot, in contrast to the values sampled in the second half of the year (July to December), that combines groups 3 and 4, located on the right side of the biplot, marked with triangles and diamonds.

Further analysis suggests a noticeable accumulation of the data during the three-month intervals, previously established as groups 1, 2, 3, and 4. Although this grouping is less pronounced, it is clearly identifiable since the first data group is mainly located in the lower-left quadrant of the biplot. The second is predominantly grouped within the upper left segment. In contrast, groups 3 and 4 seem to be almost evenly distributed on the upper and lower right quadrants.

A more exhaustive investigation implied considering monthly clusterings within each group (i.e., include the symbol sizes into our considerations). Although there are some vague tendencies of monthly data cumulation, they are not nearly enough to make educated, reliable conclusions regarding their behavior. For that purpose, the authors would once again suggest a more frequent data gathering campaign (e.g., daily sampling for a shorter time interval).

CCME water quality index

Although any substantial research requires a fair amount of data, standard approaches in water quality assessment rely on finally representing the overall water quality via a simplified methodology. Consequently, to provide a clearer understanding of the water quality, the evaluation of Lake Ludas was carried out by computing the water quality index (WQI), relying on the Canadian Council of Ministers of the Environment Water Quality Index (CCME WQI), as the most suitable for the lake at hand (Davies, 2006; Bilgin, 2018; Horvat and Horvat, 2020). This method was selected after analyzing several approaches presented in numerous research papers (Lumb et al., 2011; Voudouris and Voutsas, 2012; Bhateria and Jain, 2016; Ji et al., 2016; Liu et al., 2017; de Almeida and de Oliveira, 2018). The paper gives only a short overview of the CCME WQI method, while a detailed representation can be found (CCME, 2001, 2012; Lumb et al., 2006, 2011; Horvat and Horvat, 2020).

The water quality index CCME WQI is computed as a combination of scope, marked as F_1 , frequency denoted with F_2 and amplitude F_3 :

$$CCME\ WQI = 100 - \left[\frac{\sqrt{F_1^2 + F_2^2 + F_3^2}}{1.732} \right] \quad (1)$$

whereas the factors are estimated using the following equations:

$$F_1 = 100 \cdot \left[\frac{n_{nmo}}{n_{par}} \right], \quad F_2 = 100 \cdot \left[\frac{n_{fail}}{n_{total}} \right] \quad (2)$$

$$F_3 = \left[\frac{n_{nse}}{0.01 \cdot n_{nse} + 0.01} \right] \quad (3)$$

where n_{nmo} denotes the number of parameters not meeting the objectives, n_{par} is the total number of considered parameters (in this case $n_{par} = 7$), n_{fail} stands for the number of failed tests, and n_{total} marks the total number of tests. The total number of tests is the product of the total number of considered parameters and the number of times the parameters were tested. The value n_{nse} is the normalized sum of excursion showing the collective amount of individual tests that don't meet the objective and is computed using equation:

$$n_{nse} = \left[\frac{\sum_{i=1}^n excursion_i}{n_{test}} \right] \quad (4)$$

while the individual excursions i represents the number of times individual measurements fail the objective and are determined according to:

$$excursion_i = \left[\frac{x_{fail,i}}{y_{obj}} \right] - 1, \quad excursion_i = \left[\frac{y_{obj}}{x_{fail,i}} \right] - 1 \quad (5)$$

where $x_{fail,i}$ is the value of the failed test, and y_{obj} is the objective of the test.

Table 9
Water quality objectives for the computation of CCME WQI for Lake Ludas

No.	Parameter	Objectives
1.	SS (mg/L)	35
2.	TP (mg/L)	1
3.	Cond. ($\mu\text{S}/\text{cm}$)	1500
4.	COD _{Cr} (mg/L)	50
5.	Chl-a (mg/m^3)	100
6.	pH (-)	8.5
7.	TN (mg/L)	10 ^a , 20 ^b
^a Objective from December until end of April, ^b Objective from May until end of November		

The CCME WQI was determined in two ways: using the 7 selected water quality parameters to establish one representative WQI for the complete time interval of eight years, and computing one WQI for each year separately to detect alterations of the lake's quality through the years. Considering the significant spatial variations of the quality parameters identified during the previous examinations, both approaches were performed for the three sampling locations separately. The desired objectives for the considered water quality parameters are given in Table 9.

The results for the CCME WQI are presented graphically in Figure 12, where Figure 12(a) depicts the water quality index for every year and sampling location, while Figure 12(b) presents the water quality index for the entire analyzed time period. A higher value of the CCME WQI describes a better water quality and *vice versa*.

Discussion

After implementing a series of statistical analyses, including the determination of the distributions most suitable for the considered parameters, data visualization, and evaluation of the correlation so efficient, the initially considered 15 water quality parameters were reduced to 7 parameters sufficient to provide a trustworthy display of Lake Ludas quality state. Although the implemented analysis did not yield a clear tendency, it seems to imply the inflow on the north of the lake as a possible cause in the water quality alterations, possibly identifying the Palic-Ludas channel as a contaminant source. By studying the temporal variations of the quality parameters, we noted an increase in the water quality from 2013 to 2016 (2017) in both the north and middle part of the lake, while the south part seems to have its own trends (Fig. 4). On the other hand, the southern part of the lake has a higher conductivity virtually in all the examined years (Fig. 5), indicating possible pollution from fertilizers from adjacent fields on this narrow part of the lake.

Employing the PCA allowed us a more comprehensive insight into the overall quality alterations within the lake, both in space and time. Table 7 and Fig. 8 present the variations included by the computed principal components, where we can see that PC1 and PC2 contain nearly 60% of the data variation. Figure 9

presents a spatially oriented biplot that includes the first two principal components intended for the identification of possible clusters regarding the locations. The displayed results support this idea, as the measurements obviously group according to the sampling location, which is in accordance with the analyzed box plots. Although there is some overlapping, the clustering according to sampling location confirms the spatial variation in the measured data, as well as the source of pollution coming from the northern part of the lake.

Figures 10 and 11 depict the same measurements in the form of biplots, now labeled with regards to the month of sampling. The clustering is less evident for the results given in Fig. 10, since it encloses all of the measurements both in space and time, making it harder to make out potential accumulation based on various criteria. Regardless, we can notice some clustering inclination suggesting that some months (e.g., January, February, March, etc.) seem to be grouped more around one side of the plot, while others (e.g., October, November, and December) on the other. To point out this kind of behavior, we employed the same notation and coloring for groups of three months, as explained in Section *Principal component analysis*, differentiating the months within a group only by varied symbol sizes. Using this notation principle accompanied with the reduction of the displayed data to the north sample location, the biplot given in Fig. 11 makes it relatively easy to detect a couple of characteristics. There is a clear distinction between the first two data groups, including sampling from January to June, and the third and fourth groups, including data from July to December. Furthermore, the data from the first group (January to March) have generally different quality compared to the remaining groups. The same can be stated for the second group, while the third and fourth groups have a more pronounced overlapping of the quality characteristics. This suggests that the quality variations tend to decrease towards the second half of the year as the colder months approach. Although the data gives an impression of additional grouping between every month, making such conclusions based on the presented data would be bold but reckless. Such considerations would require a more thorough insight into the data that can be provided by introducing temporally much denser data sampling. For instance, daily measurements of properly selected water quality parameters would allow us to identify additional time-dependent behavior. Nonetheless, it can be safely concluded that the data clustering is more pronounced regarding the sampling locations than the data acquisition timing. This judgment also collaborates with the results identified on the box plots and makes it safe to suggest that the number of monitored sampling locations shouldn't be reduced and could potentially be increased.

The computed CCME WQI (Fig. 12) demonstrated that the water quality of Lake Ludas got progressively better from 2013 to 2016 (2017), which is following the conclusions attained from Figs. 4 and 5. Furthermore, these results support the idea that the water quality seems to improve farther from the inflow location on the north, aside from the timing of the data sampling suggesting that the significant source of pollution is located in the northern part of the lake, most probably the inflow water from the Palic-Ludas channel. Another conclusion that can be drawn is that the presently used three sampling locations are indeed necessary since there is a calculable difference in CCME WQI, and thus in the water quality as well between the northern part and the rest of Lake Ludas. Considering the lake's hydraulic peculiarity, described in Sect. 2.1, the primary mechanism of transferring pollutants to the southern part of the lake is wind. Therefore, while the number of monitored parameters can be reduced and still be enough to assess the lake's water quality, the three sampling locations should remain since they can give valuable information

about the faith of pollution coming from the lake's northern part. Moreover, including additional sites for data analysis should be contemplated, even if as just a one-time undertaking, since it would provide valuable insight into the spatial distribution of the water quality.

Conclusion

Lake Ludas is a Nature Reserve site. Thus there are fairly regular monitoring activities of its water quality that include both spatially and temporally distributed data. These measurements include a monthly water sampling at three different locations at the north, middle and south parts of the lake. Although 35 physical and chemical parameters are being monitored, preliminary analysis revealed only 15 of them are observed monthly during the entire time interval from 2011 to 2018. Aiming to reduce the number of parameters that would be included in the upcoming investigations, we engaged in the statistical analysis, including computing the correlation matrix, establishing the data distributions, and implementing distribution fit test, helping us select which water quality parameters should be chosen as the most representative for further analysis. The chosen parameters included Cond., pH, COD_{Cr}, Chl-a, SS, TP, and TN.

Further temporal and spatial analysis of the water quality included box plots, scatter plots, and principal component analysis. Box plots made it apparent that there is both a spatial and a temporal change of the water quality that can be detected. The temporal alteration suggests an improvement of the quality status during the time interval from around 2013 until 2016 and 2017. This improvement is noticeable in the north and middle parts of the lake, while it was less clear on the south part. Another distinct tendency is the spatial variation of the water quality. It was identified that the north part of the lake has the worst quality, which seems to improve as one departs from the north, implying that the primary source of pollution comes from the Palic-Ludas channel that connects to the lake at its north part. Another observation suggests that the south part of the lake has a higher conductivity. Considering how narrow the lake is at this location and the fact it is surrounded by agricultural land would suggest fertilizing material as a source of pollution.

The principal component analysis provided additional insight into the data characteristics by preparing biplots focusing on both time and space distribution. We were able to confirm the existence of data clustering in both aspects. Additionally, we showed a general seasonal character of the measurements that can indeed be grouped through time. By compiling groups made of three subsequent months, we showed that the measurements show altering water quality during the first and second half of a year. Furthermore, there is a noticeable change in the quality in the first and second trimester, while the third and fourth trimester alterations seem less obvious. Finally, one can also recognize variations of the water quality within a trimester by comparing values throughout months, but in this case, the changes are less explicit. As an overall conclusion, we found that a more thorough evaluation of the spatial and temporal characteristics of the water quality changes requires much denser data gathering that would include daily measurements for the temporal analysis and a seriously increased number of sampling locations for the spatial evaluation.

The computation of the CCME WQI supported the previously noted conclusions regarding the spatial and temporal tendencies, once again proposing that a significant source of pollution is the inflow of water from the Palic-Ludas channel. Furthermore, one can offer a suggestion that although the monitored water quality

parameters can be reduced significantly and still be able to compute a representative CCME WQI, the necessity of three sampling locations remains relevant. The data gathered from these three locations can give valuable information about the spread of pollution coming from the northern part of the lake, as well as information regarding possible pollution from agricultural land located on the south section of Lake Ludas. Ultimately, we would propose implementing an enhanced data sampling campaign regarding enriched data acquisition both in space in time to allow a better understanding of the lake quality alterations. Namely, the presented data includes eight years of monthly data on three locations, yet proved insufficient for an overall unambiguous quality assessment that would result in clear-cut conclusions regarding the contaminant sources or aftereffects.

Declarations

Acknowledgments

This work was funded by the Ministry of Education, Science and Technical Development of the Republic of Serbia [Programs no. 451-03-9/2021-14/200134 and 451-03-9/2021-14/200093].

Author contribution

Z. Horvat – methodology, data treatment and visualization, writing; M. Horvat – conceptualization and supervision, methodology; K. Pastor – writing, reviewing and editing.

Availability of data and materials

All data generated or analyzed during this study are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

The authors declare no competing interests.

References

1. Bhateria R, Jain D (2016) Water quality assessment of lake water: a review. *Sustain Water Resour Manag* 2:161–173. <https://doi.org/10.1007/s40899-015-0014-7>
2. Bilgin A (2018) Evaluation of surface water quality by using Canadian Council of Ministers of the Environment Water Quality Index (CCME WQI) method and discriminant analysis method: a case study

- Coruh River Basin. *Environ Monit Assess* 190:554. <https://doi.org/10.1007/s10661-018-6927-5>
3. Canadian Council of Ministers of the Environment (2001) Canadian Water Quality Guidelines for the Protection of Aquatic Life, CCME Water Quality index 1.0, User's Manual. CCME, Winnipeg. <https://ccme.ca/en>
 4. Canadian Council of Ministers of the Environment (2012) Synthesis of Research and Application of the CCME Water Quality Index, CCME Project No. Malahat, British Columbia, pp 520–2012
 5. Clesceri LS, Greenberg AE, Eaton AD (eds) (1999) Standard methods for the examination of water and wastewater, 20th edition. American Public Health Association (APHA), the American Water Works Association (AWWA), and the Water Environment Federation (WEF), Washington DC, USA. http://srjcstaff.santarosa.edu/~oraola/Assets/APHA_SM_20.pdf
 6. Čoha F (1990) *Voda za piće - Standardne metode za ispitivanje higijenske ispravnosti*. Savezni zavod za zdravstvenu zaštitu. NIP Privredni pregled, Belgrade, Serbia (in Serbian). <https://www.tehnologijahrane.com/knjiga/voda-za-pie>
 7. Davies J-M (2006) Application and Tests of the Canadian Water Quality Index for Assessing Changes in Water Quality in Lakes and Rivers of Central North America. *Lake Reserv Manag* 22(4):308–320. <https://doi.org/10.1080/07438140609354365>
 8. de Almeida GS, de Oliveira IB (2018) Application of the index WQI CCME with data aggregation per monitoring campaign and per section of the river: case study - Joanes River, Brazil. *Environ Monit Assess* 190:195. <https://doi.org/10.1007/s10661-018-6542-5>
 9. Dusabe M-C, Wronski T, Gomes-Silva G, Plath M, Albrecht C, Apio A (2019) Biological water quality assessment in the degraded Mutara rangelands, northeastern Rwanda. *Environ Monit Assess* 191:139. <https://doi.org/10.1007/s10661-019-7226-5>
 10. Gradilla-Hernández MS, de Anda J, Garcia-Gonzalez A, Montes CY, Barrios-Pina H, Ruiz-Palomino P, Díaz-Vázquez D (2020) Assessment of the water quality of a subtropical lake using the NSF-WQI and newly proposed ecosystem specific water quality index. *Environ Monit Assess* 192:296. <https://doi.org/10.1007/s10661-020-08265-7>
 11. Guo X, Zhang X, Yue H (2018) Evaluation of hierarchically weighted principal component analysis for water quality management at Jiaozuo mine. *Int Biodeterior Biodegradation* 128:182–185. <https://doi.org/10.1016/j.ibiod.2017.11.012>
 12. Hinkle DE, Wiersma W, Jurs SG (2003) *Applied Statistics for the Behavioral Sciences* – 5th edition. Houghton Mifflin, Boston
 13. Horvat M, Horvat Z (2020) Implementation of a monitoring approach: the Palic-Ludas lake system in the Republic of Serbia. *Environ Monit Assess* 192:150. <https://doi.org/10.1007/s10661-019-7927-9>
 14. Horvat Z, Horvat M (2016) Two Dimensional Heavy Metal Transport Model for Natural Watercourses. *River Res Appl* 32(6):1327–1341. <https://doi.org/10.1002/rra.2943>
 15. Horvat Z, Isic M, Spasojevic M (2015) Two dimensional river flow and sediment transport model. *Environ Fluid Mech* 15(3):595–625. <https://doi.org/10.1007/s10652-014-9375-y>
 16. Ji X, Dahlgren RA, Zhang M (2016) Comparison of seven water quality assessment methods for the characterization and management of highly impaired river systems. *Environ Monit Assess* 188:15.

<https://doi.org/10.1007/s10661-015-5016-2>

17. Kumar P (2019) Numerical quantification of current status quo and future prediction of water quality in eight Asian megacities: challenges and opportunities for sustainable water management. *Environ Monit Assess* 191:319. <https://doi.org/10.1007/s10661-019-7497-x>
18. Kükre S, Mutlu E (2019) Assessment of surface water quality using water quality index and multivariate statistical analyses in Saraydüzü Dam Lake, Turkey. *Environ Monit Assess* 191:71. <https://doi.org/10.1007/s10661-019-7197-6>
19. Liu WJ, Li ZB, Zou DS, Ren CJ, Pei QB (2017) The characteristics and evaluation of water pollution in Ganjiang Tail River. *IOP Conference Series: Earth and Environmental Science*, IOP Publishing, 82. <https://iopscience.iop.org/article/10.1088/1755-1315/82/1/012022/pdf>
20. Lumb A, Halliwell D, Sharma T (2006) Application of CCME water quality index to monitor water quality: a case study of the Mackenzie River Basin, Canada. *Environ Monit Assess* 113:411–429. <https://doi.org/10.1007/s10661-005-9092-6>
21. Lumb A, Sharma TC, Bibeault JF, Klawunn P (2011) A Comparative Study of USA and Canadian Water Quality Index Models. *Water Qual Expo Health* 3(3–4):203–216. <https://doi.org/10.1007/s12403-011-0056-5>
22. Mandal UK, Warrington DN, Bhardwaj AK, Bar-Tal A, Kautsky L, Minz D, Levy GJ (2008) Evaluating impact of irrigation water quality on a calcareous clay soil using principal component analysis. *Geoderma* 144:189–197. <https://doi.org/10.1016/j.geoderma.2007.11.014>
23. Mutlu E (2019) Evaluation of spatio-temporal variations in water quality of Zerveli stream (northern Turkey) based on water quality index and multivariate statistical analyses. *Environ Monit Assess* 191:335. <https://doi.org/10.1007/s10661-019-7473-5>
24. Olsen RL, Chappell RW, Loftis JC (2012) Water quality sample collection, data treatment and results presentation for principal components analysis - literature review and Illinois River watershed case study. *Water Res* 46:3110–3122. <https://doi.org/10.1016/j.watres.2012.03.028>
25. Ouyang Y (2005) Evaluation of river water quality monitoring stations by principal component analysis. *Water Res* 39:2621–2635. <https://doi.org/10.1016/j.watres.2005.04.024>
26. Palani S, Liong S-Y, Tkachik P (2008) An ANN application for water quality forecasting. *Mar Pollut Bull* 56:1586–1597. <https://doi.org/10.1016/j.marpolbul.2008.05.021>
27. Pastor KA, Acanski MM, Vujic DjN, Dj J, Wienkoop S (2016) Authentication of Cereal Flours by Multivariate Analysis of GC-MS Data. *Chromatographia* 79(19–20):1387–1393. <https://doi.org/10.1007/s10337-016-3142-9>
28. Radu V-M, Ionescu P, Deak Gy, Diacu E, Ivanov AA, Zamfir S, Marcus M-I (2020) Overall assessment of surface water quality in the Lower Danube River. *Environ Monit Assess* 192:135. <https://doi.org/10.1007/s10661-020-8086-8>
29. Voudouris K, Voutsas D (2012) Water Quality Monitoring and Assessment. InTech Open, Rijeka, Croatia. <https://www.intechopen.com/books/water-quality-monitoring-and-assessment>
30. Yadav S, Babel M, Shrestha S, Deb P (2019) Land use impact on the water quality of large tropical river: Mun River Basin, Thailand. *Environ Monit Assess* 191:614. <https://doi.org/10.1007/s10661-019-7779-3>

31. Zeinalzadeh K, Rezaei E (2017) Determining spatial and temporal changes of surface water quality using principal component analysis. J Hydrol Reg Stud 13:1–10.
<https://doi.org/10.1016/j.ejrh.2017.07.002>

Figures

Figure 1
Lakes Palic and Lake Ludas with the sampling locations at Lake Ludas

Figure 2
Examples of the Pearson's correlation coefficients

Figure 3

Probability scale plots for CODCr, TN and TP successively

Figure 4

Temporal distribution of water quality parameters

Figure 5

Temporal distribution of water quality parameters

Figure 6

Spatial distribution of water quality parameters

Figure 7

Scatter plot matrix

Figure 8

Scree plot

Figure 9

Biplot with labeled sampling locations: north (N), middle (M) and south (S)

Figure 10

Biplot with labeled months for all three locations: north, middle and south

Figure 11

Biplot with labeled months for the north part of Lake Ludas

Figure 12

Changes of water quality index throughout space and time