Pol. J. Environ. Stud. Vol. 27, No. 1 (2018), 19-30
DOI: *10.15244/pjoes/74367* ONLINE PUBLICATION DATE: 2017-10-30

Original Research

The Possibility of Applying the EM-PCA Procedure to Lake Water

Anna Bucior-Kwaczyńska*

Department of Chemistry and Natural Waters Management, Institute for Research on Biodiversity, Faculty of Biology, Szczecin University, Felczaka 3c, 71-412 Szczecin, Poland

Received: 8 May 2017
Accepted: 1 June 2017

Abstract

Missing elements in experimental data often occur in ecological and biological sciences. In this case, it is difficult to carry out any data analysis and their evaluation. This paper presents one of the chemometric techniques – principal component analysis (PCA) – used to classify water quality indices on data that contain missing elements. The surface water of Czajcze Lake in Wolin National Park (northwestern Poland) was investigated. Sixteen water-quality indices were appointed in a period from April to October during 1983-2013. Conducted analysis of experimental data by EM-PCA grouped the presented water quality indices in natural clusters, including several principal components (PCs) about similar features. EM-PCA applied in the present work shows that this method can be used to analyze experimental data with missing data on considerable seasonal changes.

Keywords: PCA, EM-PCA, water quality, Czajcze Lake, Wolin National Park

Introduction

Understanding the phenomena occurring in natural waters is based on the correct interpretation of analytical data obtained during the experiments. For this purpose, the individual variables are usually described by using the input values of minimum, average, maximum, standard deviation, and coefficient of the data variation, at least through the definition of the median or percentile. Nowadays, however, to understand the more subtle nature of the variables we more frequently use chemometric techniques of data analysis [1-7]. This allows for relatively simple and readable processing of the experimental data without any prior assumptions.

One of the most frequently used calculation methods in differenting scientific disciplines is the method of principal components (PCA), which is one of the main components of statistical analysis. PCA allows us to get readable information, often contained in a very large number of entered input data, that are often summarized to a few main components. Due to this fact, PCA is the first step in statistical data analysis. In mathematical terms: the entered data are grouped according to the criterion of maximization incompatibility. Accrued maximization is performed by using Lagrange's multipliers method. This statistical analysis of the data led principal components to the eigenvectors of the matrix of variance – covariance that is sensitive to the outstanding data. Missing data is increasingly frequent in biological and environmental sciences. Data that are actually temporary values can be missing for different reasons, so they should be not supplemented at random. According to [8], they should

*e-mail: aniabucior@wp.pl

be omitted or use specialized computational modules for their supplementation.

This paper attempts to use PCA to determine incomplete data [3, 6, 9-22] collected during investigations of Czajcze Lake in Wolin National Park (WNP) in 1983-2013.

Characteristic of Czajcze Lake

Czajcze Lake (Figs 1-2) is located on Wolin Island within WNP in the Warnowo Protection Zone in the northeastern region of WNP between the cities of Międzyzdroje, Warnowo, and Wisełka [23-24]. Czajcze is the third drainage lake of several flowing lakes related to Lewińska (Pojezierna) Stream on Wolin Island. Lewińska water flows down from Warnowo Lake through lakes Rabiąż, Czajcze, and Domysłowskie, and through other lakes of the Warnowsko-Kołczewskiego Lakes District not lying within WNP territory to Kamieński Lagoon [25].

Czajcze's basin has two distinct parts characterized by a flat and shallow bottom. The lake has the shape of a horseshoe. In the western part it is connected to Rabiąż by a narrow channel, while in the southeastern part it is connected to Domysłowskie by two overgrown watercourses lying next to each other. The channel connecting Czajcze with Rabiąż and Domysłowskie is Lewińska Stream. Czajcze morphometry and its partial cover from the wind suggest an average exposure to the wind and a limited mixing of waters to the bottom. Basic morphometrical and bathymetrical parameters [26] for Czajcze are collected in Table 1.

Material and Methods

Water samples were collected [after 27] at the water measuring sampling station (Fig. 2) from the surface layer (ca. 25 cm below the water surface) of Czajcze. Water samples were collected from the vegetation season from March to October with a frequency of once a month in arbitrarily selected dates in the years 1983-2013. Temperature [28] and pH of the water [29-30] were determined at the place of sample collection. Water samples were taken separately to determine dissolved oxygen concentration [29-30]. Water samples taken for determining the concentration of dissolved substances

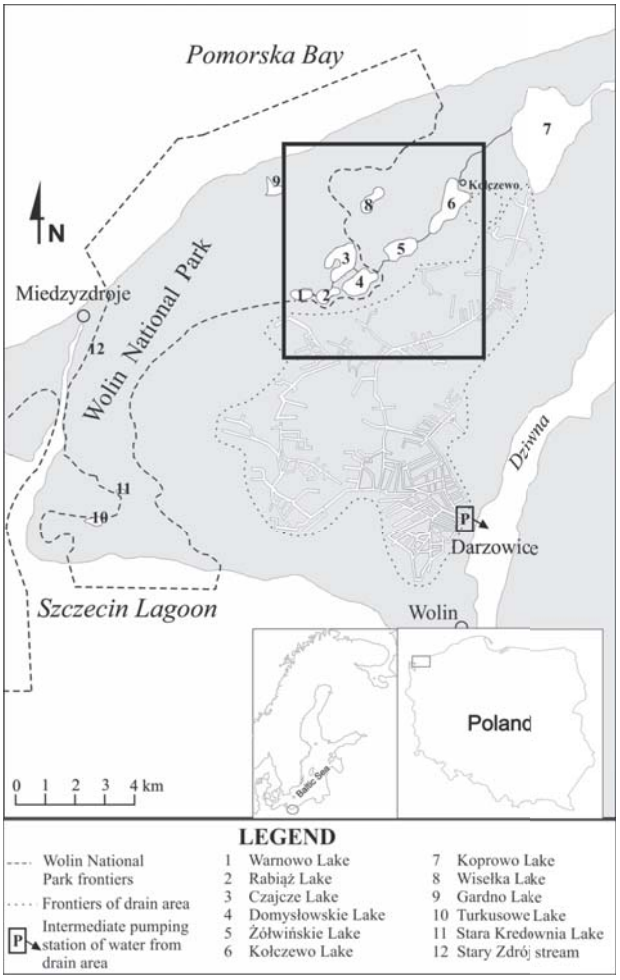

Fig. 1. Location of Czajcze Lake within Wolin National Park [23, with some changes].

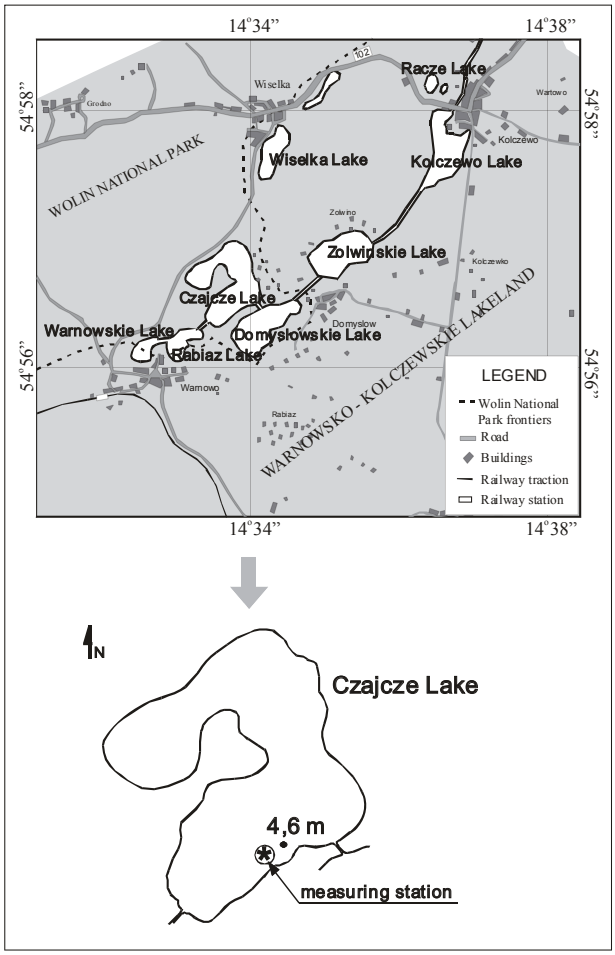

Fig. 2. Lakes Warnowsko-Kołczewskie Lakeland – measuring station location at Czajcze Lake [24, with some changes].

Table 1. Basic morphometrical and bathymetrical parameters of Czajcze Lake in Wolin National Park on Wolin Island, NW Poland.

No.	Lake parameters	
1	Latitude	53º56,5' N
2	Longitude	14º34,0' E
3	Surface (ha)	71.5
4	Max. length (m)	1,630
5	Max. width (m)	640
6	Length of shoreline (m)	4,970
7	Development of shoreline	1.66
8	Altitude (m above sea level)	1.3
9	Average depth (m)	2.9
10	Max. depth (m)	4.6
11	Volume (thousand m^3)	2,073.5
12	Uncover factor	24.7

in laboratory analyses were fixed [31] at the place of sample collection. Water samples were investigated in the laboratory within 24 hours. Determined concentrations are: five-day biochemical oxygen demand (BOD_5), chemical oxygen demand (COD-Cr), nitrate (NO_3^-), nitrite (NO_2^-), ammonium ions (NH_4^+), total nitrogen (TN), soluble reactive orthophosphates (V) (SRP) and total phosphorus (TP), concentrations of calcium ions (Ca^{2+}), chloride ions (Cl^-), hydroxidodioxidocarbonate (HCO_3^-), and total concentrations of iron and manganese ($Fe_{tot.}$ and Mn_{tot}) [29-30]. In addition, the degree of water saturation by O_2 was calculated in the water samples taken for investigation.

In order to determine how the Czajcze water samples interact with each other, the collected results of the investigation were conducted with PCA using the computer software Statistica.

PCA is widely used to organize data by their respective grouping (packaging) and graphical presentation [32-33]. PCA is a nonparametric classification method whose main aim is to clarify the information contained in the data that were entered into the program by the so-called principal components (PCs). PCs are actually orthogonal and linear combinations of the data for their maximum diversity. Simultaneously, each of the PCs carries different information about data variability [3, 7, 34]. PCA allows us to group the experimental data and present them in the form of several PCs, and PCs contain nearly identical information to the experimental data. Such multidimensionality of the data through the explanation of correlations between variables is very useful, because during the PCA there is no loss of information contained in a set of experimental data [35]. Thus, PCA is the first step in most chemometric methods, such as cluster analysis or a neural network.

A good method for determining missing data is to predict an expansion to the maximum algorithm [36]. While the classic PCA taking multiple dimensions of data analysis is quite often used, the analysis of the main components for missing data (EM-PCA) is a little bit different [37-39]. EM-PCA begins with the initialization of the missing data by quoting the average values in the corresponding rows and columns, and then their iterations in such a way to substitute missing data with values predicted (estimated) by the PCA [40]. The algorithm for missing data [41-47] is repeated until the convergence criteria. The mean values obtained in the convergence criterion replace missing data in the corresponding rows and columns. Thus, the complete input data are obtained for PCA. The distribution of values in the full dataset and the prediction of X in the basic model of PCA are executed [3-5]:

$$X = T \cdot P^T$$

…where X is the matrix built from experimental data of the m characteristics and n variables, $T = m{\cdot}f$ is the dimension having a first recorded vector f, and $P\ (n{\cdot}f)$ is the dimension of matrix containing the shown experimental data by using the primarily defined variables, and then the expected missing data are compared with their previously obtained average values [9]. Thus determined, missing data are not taken into account in further calculations. Further calculations are executed only for observed experimental data.

In addition to the experimental data, regression equations were used wherever possible, and the diagrams were made depicting changes of the values investigating water quality parameters over time.

Results and Discussion

Sixteen variables of water quality indices are analyzed in this work. Table 2 shows the dependence between each index investigated in this work in the matrix correlation [6, 48-49] describing the degree of dependence on individual changeable interdependence. Numeric data of matrix correlation shows that the greater the absolute value between two variables (water quality indices), the more important are the correlations – positive or negative – that occur between these indices.

Then, eigenvalues of the matrix correlation were appointed (Table 3), which are a measure of the variability of the primary (original) data in the coordinates of the main components. On this basis, a graph was obtained (Figs 3-4) to illustrate (with appropriate to impinge data) which variables (water quality indices) show a similar pattern of changes and which are clearly distinguished. The direction and the length of the eigenvector (i.e., charge) in Fig. 3 assigns the degree to which each of the investigated water quality indices affect the main components. The analysis of the eigenvalues main components > 1 PCA shows that the plane of the first

Table 2. Significant level of Pearson correlation between investigated water quality indices.

Water quality indices (units)	pH (pH units)	DO (mg O_2/dm^3)	Water saturation (%)	BOD_5 (mg O_2/dm^3)	COD-Cr (mg O_2/dm^3)	NO_3^- (mg N/dm^3)	NO_2^- (mg N/dm^3)	NH_4^+ (mg N/dm^3)
pH (pH units)	1.000							
DO (mg O_2/dm^3)	0.439	1.000						
water saturation (%)	0.383	0.984***	1.000					
BOD_5 (mg O_2/dm^3)	0.284*	0.702***	0.770***	1.000				
COD-Cr (mg O_2/dm^3)	-0.494**	-0.611***	-0.499**	-0.374*	1.000			
NO_3^- (mg N/dm^3)	-0.730***	-0.394*	-0.354*	-0.414*	0.447*	1.000		
NO_2^- (mg N/dm^3)	-0.381*	-0.328*	-0.291*	-0.283*	0.944***	0.456**	1.000	
NH_4^+ (mg N/dm^3)	0.540**	0.486**	0.445*	0.316*	-0.303*	-0.719***	-0.357*	1.000
TN (mg N/dm^3)	0.393*	0.549**	0.527**	0.301*	-0.357*	-0.586**	-0.342*	0.977***
SRP (mg PO_4/dm^3)	0.436*	0.361*	0.309*	0.699***	-0.281*	-0.882***	-0.336*	0.504**
TP (mg PO_4/dm^3)	0.532**	0.307*	0.333*	0.296*	-0.721***	-0.880***	-0.785***	0.392*
Ca^{2+} (mg Ca/dm^3)	0.464**	0.479**	0.598**	0.368*	-0.344*	-0.284*	-0.318*	0.241
Cl^- (mg Cl/dm^3)	0.841***	0.394*	0.363*	0.324*	-0.801***	-0.820***	-0.795***	0.329*
HCO_3^- (mg HCO_3/dm^3)	0.291*	0.790***	0.831***	0.978**	-0.342*	-0.328*	-0.313*	0.289*
$Fe_{tot.}$ (mg Fe/dm^3)	-0.334*	-0.933***	-0.867***	-0.638***	0.793***	0.351*	0.590**	-0.333*
$Mn_{tot.}$ (mg Mn/dm^3)	-0.322*	-0.928***	-0.861***	-0.632***	0.781***	0.346*	0.587**	-0.326*
pH (pH units)								
DO (mg O_2/dm^3)								
water saturation (%)								
BOD_5 (mg O_2/dm^3)								
COD-Cr (mg O_2/dm^3)								
NO_3^- (mg N/dm^3)								
NO_2^- (mg N/dm^3)								
NH_4^+ (mg N/dm^3)								
TN (mg N/dm^3)	1.000							

Table 2. Continued.

Water quality indices (units)	TN (mg N/dm^3)	SRP (mg PO_4/dm^3)	TP (mg PO_4/dm^3)	Ca^{2+} (mg Ca/dm^3)	Cl^- (mg Cl/dm^3)	HCO_3^- (mg HCO_3/dm^3)	$Fe_{tot.}$ (mg Fe/dm^3)	$Mn_{tot.}$ (mg Mn/dm^3)
SRP (mg PO_4/dm^3)	0.357*	1.000						
TP (mg PO_4/dm^3)	0.280*	0.798***	1.000					
Ca^{2+} (mg Ca/dm^3)	0.261	-0.515**	-0.429*	1.000				
Cl^- (mg Cl/dm^3)	0.286*	0.613***	0.862***	-0.005	1.000			
HCO_3^- (mg HCO_3/dm^3)	0.363*	0.539**	0.332*	0.316*	0.307*	1.000		
$Fe_{tot.}$ (mg Fe/dm^3)	-0.387*	-0.333*	-0.516**	-0.881***	-0.495**	-0.738**	1.000	
$Fe_{tot.}$ (mg Fe/dm^3)	-0.387*	-0.333*	-0.516**	-0.881***	-0.495**	-0.738**	1.000	
$Mn_{tot.}$ (mg Mn/dm^3)	-0.380*	-0.326*	-0.511**	-0.877***	-0.489**	-0.725**	0.995***	1.000
Significance level: *** $\alpha \leq 0.001$, ** $0.001 < \alpha \leq 0.01$, * $0.01 < \alpha \leq 0.05$								

and second principal components describes 71.00% of the variance of the primary (original) data (Fig. 3). According to the above criteria, only principal components (PCs) with values higher than the values of the principal components were considered [35, 50]. The percentage explained by the first two dimensions is 71%, and according to [19] it is very high. In literature it is assumed that if the percentage explained by the first two dimensions is 75%, then it is statistically significant [34]. In [51] the authors prefer 70% and 64% in [52] the percentage explained by the first two dimensions, as it is very significant statistically.

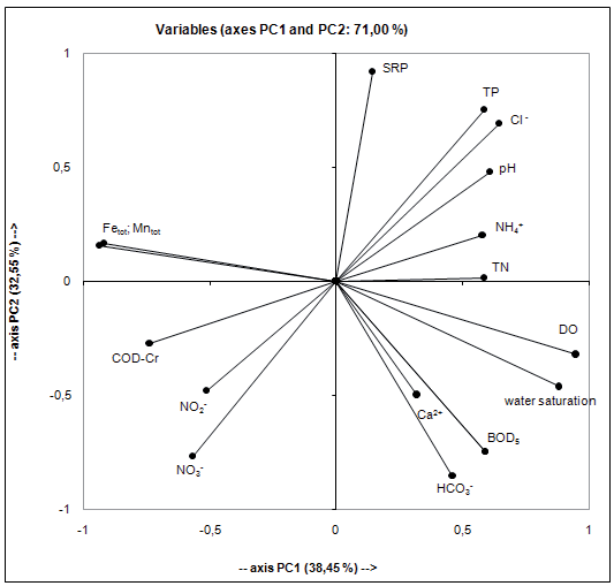

Fig. 3. PCA ordination diagram of investigated water quality indices.

Fig. 3 clearly shows which of the investigated water quality indices have a similar pattern of change, and which are completely different from each other. Those of the investigated parameters (water quality indices that had a higher eigenvalues) had the eigenvector corresponding to that value less correlated with other parameters [53]. On this basis, the results of 16 water quality indices that are presented in this work were reduced in the space diagram (Fig. 3) to the several sets of main PC components. Fig. 3 shows that exposure to PCA data is packed into eight groups.

Individual groups include the following indices of the investigated water quality:

– I – one water quality indicator: SRP (mg PO_4/dm^3).
– II – three water quality indices: TP (mg PO_4/dm^3), Cl^- (mg Cl/dm^3), pH (pH units).
– III – two water quality indices: NH_4^+ (mg N/dm^3), TN (mg N/dm^3).
– IV – two water quality indices: DO (mg O_2/dm^3), water saturation by O_2 (%).
– V – three water quality indices: BOD_5 (mg O_2/dm^3), Ca^{2+} (mg Ca/dm^3), HCO_3^- (mg HCO_3/dm^3).
– VI – two water quality indices: NO_3^- (mg N/dm^3), NO_2^- (mg N/dm^3).
– VII – one water quality indicator: COD-Cr (mg O_2/dm^3).
– VIII – two water quality indices: $Fe_{tot.}$ (mg Fe/dm^3), $Mn_{tot.}$ (mg Mn/dm^3).

The above-mentioned groups of water quality indices that correlate with each other were deployed in four quarters of the diagram in the following way:

– Quarter I – indices of groups I, II and III.
– Quarter II – indices of groups IV and V.
– Quarter III – indices of groups VI and VII.

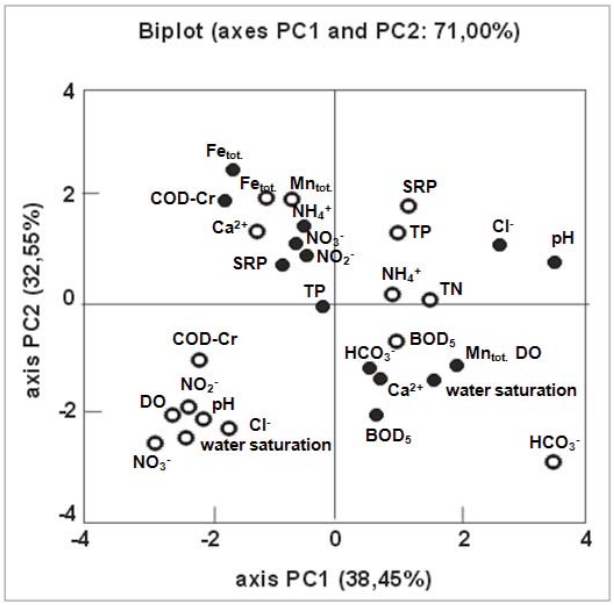

Fig. 4. Map of PCA main factors showing projection of cases before (●) and after (○) 2006.

– Quarter IV – indices of group VIII.

The values of water quality indices belonging to each of the quarters I-IV, located in the area of Fig. 3 close to each other, are positively correlated with each other. When the separate indicators are closer to each other, their correlation is more important. The values of water quality indices belonging to groups I-III that are located in quarter I are positively correlated with each other, e.g., the value of the correlation coefficients between concentrations TP and Cl^- is 0.862, between concentrations SRP and TP is 0.798, and between concentrations SRP and NH_4^+ is 0.504. The same correlations are between groups of parameters that belong to quarters II (the value of the correlation coefficient between the concentrations DO and Ca^{2+} is 0.479, and between the concentrations Ca^{2+} and BOD_5 is 0.368), III (the value of the correlation coefficient between concentrations NO_3^- and NO_2^- is 0.456, and between concentrations of COD-Cr and NO_3^- is 0.447), and IV (the value between concentrations $Fe_{tot.}$ and $Mn_{tot.}$ is 0.995) of the diagram presented in Fig. 3.

Also, statistically significant positive correlations show indicators that are located in quarter I and are compared to parameters of quarter II, as well as indicators that are located in quarter III and are compared to parameters of quarter IV. However, these correlations are weaker than those correlations that occur between indicators within each group, e.g., the value of the correlation coefficient between the pH values and concentration BOD_5 is 0.284, while the value of the correlation coefficient between

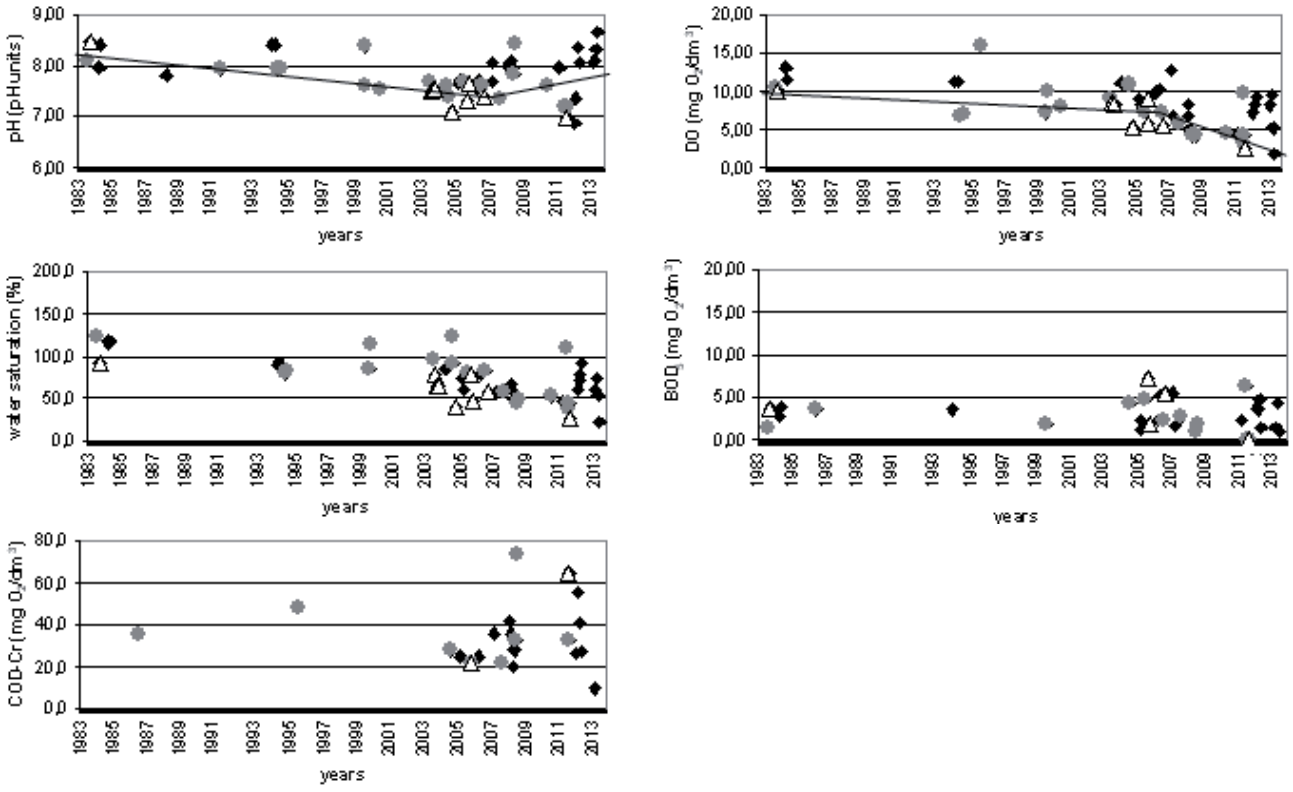

Fig. 5. Changes of concentrations of general chemical water-quality indices of surface waters in spring (♦), summer (●), and autumn (△), 1983-2013.

Annotation: Some of the figures also show graphically the equation that is statistically significant and presents the relationship between parameter value and the date of measurement (Table 4).

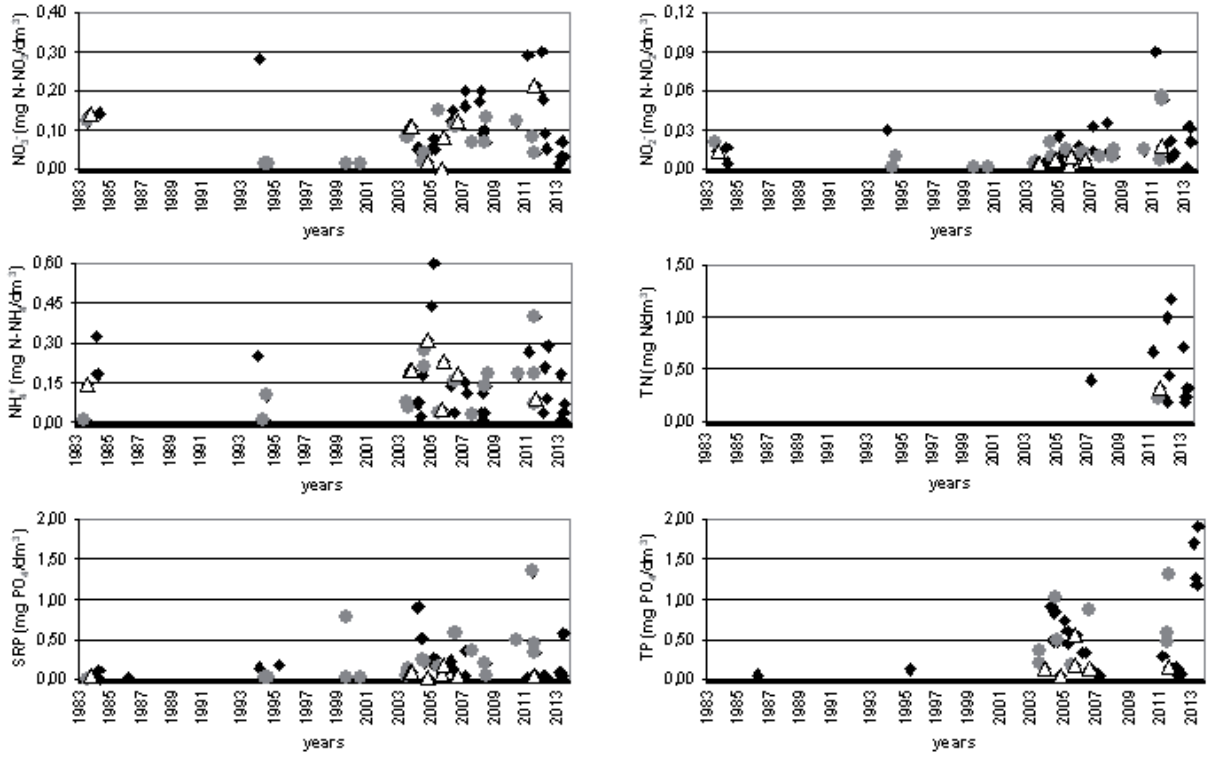

Fig. 6. Changes of selected indices characterizing concentrations in waters of nitrogen and phosphorus substances in spring (♦), summer (●), and autumn (△), 1983-2013.

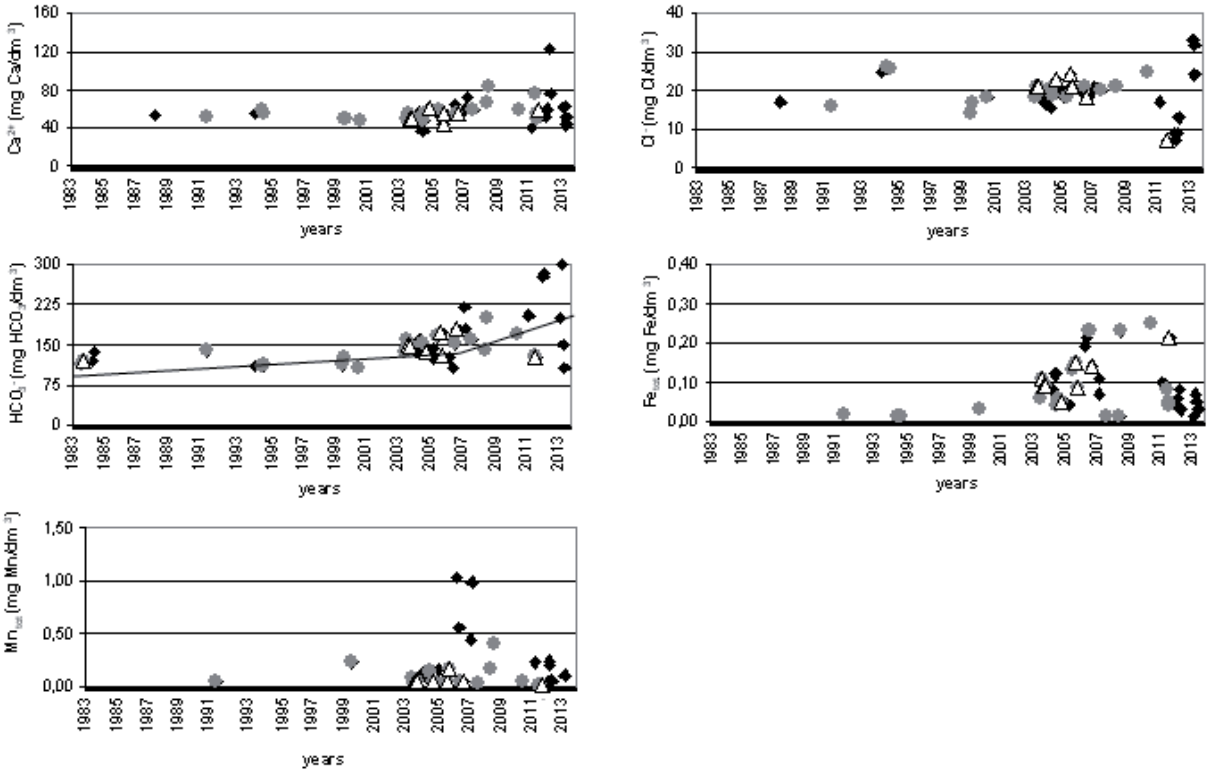

Fig. 7. Changes concentrations of selected mineral macro and micro component indices of surface waters in spring (♦), summer (●), and autumn (△), 1983-2013.
Annotation: Some of the figures also show graphically the equation that is statistically significant and presents the relationship between parameter value and the date of measurement (Table 4).

Table 3. PCA Results – eigenvalue principal component of matrix correlation.

No.	Water quality indices (units)	Eigenvectors			
		PC1	PC2	PC3	PC4
1	pH (pH units)	0.231	-0.317	-0.107	-0.168
2	DO (mg O_2/dm^3)	0.231	-0.317	-0.107	-0.168
3	water saturation (%)	0.238	0.204	0.274	0.296
4	BOD_5 (mg O_2/dm^3)	0.371	-0.136	0.010	0.011
5	COD-Cr (mg O_2/dm^3)	0.344	-0.196	0.055	0.043
6	NO_3^- (mg N/dm^3)	-0.289	-0.116	0.326	-0.210
7	NO_2^- (mg N/dm^3)	0.056	0.392	0.081	-0.220
8	NH_4^+ (mg N/dm^3)	0.228	0.320	-0.160	-0.104
9	TN (mg N/dm^3)	-0.367	0.067	0.190	0.037
10	SRP (mg PO_4/dm^3)	0.058	-0.096	0.266	0.565
11	TP (mg PO_4/dm^3)	0.180	-0.362	-0.130	-0.091
12	Ca^{2+} (mg Ca/dm^3)	0.124	-0.211	0.374	0.352
13	Cl^- (mg Cl/dm^3)	0.252	0.294	-0.040	0.211
14	HCO_3^- (mg HCO_3/dm^3)	-0.222	-0.326	-0.143	0.125
15	$Fe_{tot.}$ (mg Fe/dm^3)	0.226	0.087	0.407	-0.288
16	$Mn_{tot.}$ (mg Mn/dm^3)	0.228	0.007	0.380	-0.348
	Eigenvalue	6.537	5.533	2.598	2.332
	Variability (%)	38.452	32.549	15.281	13.718
	Cumulative %	38.452	71.001	86.282	100.000

the concentrations NO_3^- and $Fe_{tot.}$ is 0.351. Exception are values of Ca^{2+} concentrations (quarter II of the diagram), which are not correlated with some of water quality indices such as TN, NH_4^+, and Cl^- (located in quarter I of the diagram).

In turn, quarter I parameters show quite a significant negative correlation compared to quarter III parameters (e.g., the value of the correlation coefficient between concentrations Cl^- and NO_3^- is -0.820), as well as quarter II parameters compared to quarter IV parameters (e.g., the value of the correlation coefficient between the concentrations BOD_5 and $Fe_{tot.}$ is -0.638). Quarter I Parameters also show negative correlations compared to parameters of quarter IV as well as parameters of quarter II compared to parameters of quarter III (e.g., the value of the correlation coefficient between the concentrations Cl^- and $Fe_{tot.}$ is -0.495, and the correlation coefficient between the concentrations Ca^{2+} and NO_3^- has a value of -0.284). However, these correlations are not very significant statistically.

The correlations presented graphically in Fig. 3 are very similar to the correlation between investigated water quality indices shown numerically in Table 2. Other similar satisfactory optimization results were determined [54] by his research on hydrological parameter classification. [54] has modeled in his studies on study [55-58], showing that the values of parameters on PCA diagrams can be presented as tables correlating between investigated parameters. At the same time, in order to identify the missing data and their uncertainties in complex computational models, a reliable approach to analysis is needed.

Multidimensional analysis of incomplete data sets in environmental studies is rarely used [21, 59-60], e.g., [22] this fact translates human uncertainty into the method used. All the more so because of often-used individual and variable removal or medium imputation to handle incomplete data. Therefore, in this work additionally, for verification results obtained by EM-PCA, graphical dependences were plotted (Figs 5-7) showing the changes of selected parameters – indices of water quality as a function of time. It has been observed that since 2006 the surface waters of Czajcze Lake followed significant sustainable changes of concentrations of water quality indices that are determined in this work. For this purpose, the regression equation $y_i = a_0+a_1\cdot\tau+a_2\cdot(\tau-\tau_1)\cdot(\tau>\tau_1)$ was appointed to describe the changes in the long-term period of each of the water quality indices. The values of the coefficients of the regression equations and their

Table 4. Parameters (a_0, a_1, a_2) in equation $y_i = a_0+a_1\cdot\tau+a_2\cdot(\tau-\tau_1)\cdot(\tau>\tau_1)$, significance level of parameters, time (τ') as the moment pronouncement of visible changes in lake reservoir, SEE and R^2.

Water quality indices (units)	Parameters in equation $y_i=a_0+a_1\cdot\tau+a_2\cdot(\tau-\tau_1)\cdot(\tau>\tau_1)$	τ‘	Significance level of parameters*	SEE	R^2
DO ($mg\ O_2/dm^3$) $n = 55$	a_0=23.73±6.00 a_1=-0.61±0.23 a_2=0.54±0.29	January 2006	0.00023 0.01139 0.06628	0.92	0.41
pH (pH units) $n = 56$	a_0=8.32±0.16 a_1=-0.03±0.01 a_2=0.07±0.03	January 2007	0.00000 0.00062 0.01621	0.73	0.45
HCO_3^- ($mg\ HCO_3/dm^3$) $n = 47$	a_0=-58.4±12.3 a_1=6.9±2.3 a_2=-5.4±02.8	January 2006	0.82829 0.00415 0.05793	0.6	0.54

Explanations: * Statistical significant at p = 90% for DO and HCO_3^- concentrations, and p = 95% for pH. n - data number
For other investigated water quality indices the equation of regression was not included because they were not statistically significant.

Table 5. General statistical characteristics of collected results of investigated water quality indices on Czajcze Lake before and after 2006.

No.	Water quality indices (units)	General statistical characteristics									
		n	before 2006					after 2006				
			min		max	SD	CV	min		max	SD	CV
1	pH (pH units)	56	7.10	7.78	8.47	0.37	0.04	6.86	7.76	8.64	0.46	0.06
2	DO ($mg\ O_2/dm^3$)	55	5.40	10.67	26.70	5.05	0.47	1.80	6.58	12.60	2.67	0.40
3	water saturation (%)	51	40.0	100.0	273.5	55.2	0.55	21.0	59.24	110.7	19.9	0.34
4	BOD_5 ($mg\ O_2/dm^3$)	47	1.30	3.30	7.40	1.74	0.53	0.10	2.74	6.40	1.96	0.72
5	COD-Cr ($mg\ O_2/dm^3$)	26	22.0	30.7	48.0	11.77	0.38	9.3	35.5	73.4	16.86	0.47
6	NO_3^- ($mg\ N/dm^3$)	52	0.01	0.10	0.85	0.16	1.72	0.01	0.12	0.03	0.07	0.61
7	NO_2^- ($mg\ N/dm^3$)	54	0.001	0.015	0.170	0.03	2.13	0.001	0.021	0.090	0.02	0.91
8	NH_4^+ ($mg\ N/dm^3$)	52	0.01	0.25	1.47	0.32	1.26	0.01	0.13	0.40	0.09	0.73
9	TN ($mg\ N/dm^3$)	21	-	-	-	-	-	0.19	0.49	1.17	0.33	0.67
10	SRP ($mg\ PO_4/dm^3$)	52	0.01	0.21	0.90	0.26	0.24	0.01	1.24	1.34	0.31	1.27
11	TP ($mg\ PO_4/dm^3$)	40	0.03	0.45	1.02	0.32	0.71	0.05	0.60	1.91	0.61	1.01
12	Ca^{2+} ($mg\ Ca/dm^3$)	46	38	52	62	6.34	0.12	41	63	122	16.42	0.26
13	Cl^- ($mg\ Cl/dm^3$)	43	14	20	26	3.33	0.17	7	19	33	7.83	0.40
14	HCO_3^- ($mg\ HCO_3/dm^3$)	47	131.1	165.5	210.4	23.30	0.14	128.1	215.3	360.0	70.47	0.33
15	$Fe_{tot.}$ ($mg\ Fe/dm^3$)	43	0.01	0.07	0.15	0.04	0.57	0.01	0.10	0.25	0.08	0.80
16	$Mn_{tot.}$ ($mg\ Mn/dm^3$)	37	0.03	0.09	0.23	0.06	0.60	0.01	0.25	1.03	0.31	1.24

Explanations: n - data number, min - minimum value, $\bar{x}$ - average value, maximum value, SD - standard error of estimation, CV - standard deviation

significance level, as well as the month of the year, from which appeared clear changes in the investigated indices of water quality in the lake are presented in Table 4 only for selected indicators. Additionally, Table 5 presents general statistical characteristics before and after 2006 for selected investigated index-parameters of water quality. On the basis of the obtained regression equations and the general statistical characteristics it has been established that the changes that occurred have initiated a new qualitative composition of Czajcze waters.

The cause of a completely different qualitative composition of Czajcze waters since 2006 was detected in the first place in climatic conditions. There was an attempt to link them with an increased amount of precipitation. It has been observed that since the second half of 2006, the amount of atmospheric precipitation on the territory of Warnowskie Lakes has increased slightly. Unfortunately, it was not a large enough growth of atmospheric precipitation (with a few exceptions) to give a clear explanation for this state of affairs. Consider, in turn, a tributary of the waters from Rabiąż Lake (eastern Warnowo) which is connected to Czajcze by a clearly narrow channel watercourse did not make sense, because at present the channel-watercourse connecting the two lakes is completely overgrown [61] and has bottom sediments. So any water flow from Rabiąż to Czajcze is impossible. The gradual increase in water level in the lake and the establishment of a new qualitative composition of waters were explained only during the local vision conducted. We observed characteristic traces of beaver functioning in the environment, suggesting that probably the only explanation for the increase in water level in Czajcze was due to a beaver dam. So beaver activity in the lake water body initiated the appearance of a new qualitative composition of the selected water quality indices in Czajcze. The determination achieved earlier confirms the analytical results of experimental data presented in this work and analyzed using PCA (EM-PCA).

As [59] and [22] PCA method for missing data, because of its simplicity in considering the missing data, can be considered as a relative success. It provides results and loads minimizing the least squares criterion with respect to observed values [19], attribution of experimental results, and missing values in a dataset [62].

Openness remains the credibility of missing data. Several authors [15, 18-20] undertook the task of estimating the maximum part of missing data values. They showed that this is impossible because there is no mathematical upper limit for the missing data estimation. According to the author of this paper, when monitoring the natural environment with more variables, estimating missing data values is more accurate, i.e., the error of estimation is smaller. This is confirmed by [14] and [18], who state that with fewer variables and more missing data, the estimation error increases. In contrast, [15] suggests that in the absence of a data rate of 20%, they cannot be reliably estimated because the estimation error exceeds 10%, and [12] says that 50% of missing data can be estimated with good accuracy. These simulation studies suggest that this strategy is promising [63] for environmental studies. Also, the results got [35] applying chemometric analysis for river water classification, and for rapid assessment of water qualities.

In conclusion, it can be stated that PCA is one of the most important and powerful methods in chemometrics [64] and can successfully serve as a basis for handling missing data to classify water quality indices for other more accurate methods in multivariate analysis.

Conclusions

1. Applying chemometric techniques for data analysis enabled presentation of the experimental data in a simple manner without any prior assumptions. The EM-PCA method can be applied because the variables configuration will be saved (as for full data), and the omission of missing data does not affect the right described general dependencies between the individual variables.
2. EM-PCA analysis shows that the method that was used can be applied for incomplete data with a large number of experimental parameters-indices of water quality for their suitable grouping ("reduction") to the preliminary interpretation.
3. Classical statistical methods of processing experimental data confirm the results obtained by the EM-PCA. However, the results obtained through the use of classical statistical methods allow for a deeper (more accurate) analysis of collected experimental data.

Acknowledgements

I am thankful to the management of Wolin National Park for giving permission to conduct field research.

References

1. DASZYKOWSKI M., KACZMAREK K., VANDER HEYDEN Y, WALCZAK B. Robust statistics in data analysis - a review basic concepts. Chemometrics and Intelligent Laboratory Systems, **85** (2), 203, **2007**. DOI: 10.1016/j.chemolab.2006.06.016
2. STANIMIROVA I., ZEHL K., MASSART D.L., VANDER HEYDEN Y., EINAX J.W. Chemometric analysis of soil pollution data applying Tucker N-way method. Analytical and Bioanalytical Chemistry, **385** (4), 771, **2006**. DOI: 10.1007/s00216-006-0445-y
3. STANIMIROVA I., DASZYKOWSKI M., WALCZAK B. Dealing with missing values and outliers in principal component analysis. Talanta, **72** (1), 172, **2007**. DOI: 10.1016/j.talanta.2006.10.011
4. WU T., ZHAO W., GUO H., LIM H., YANG Z. A streaming PCA based VLSI chip for neural data compression. In: IEEE Biomedical Circuits and Systems Conference, Shanghai, China 17-19 October 2016, 192, **2017**. DOI: 10.1109/BioCAS.2016.7833764

5. KAYA I.E., PEHLIVANLI A.Ç., SEKIZKARDEŞ E.G., IBRIKCI T. PCA based clustering for brain tumor segmentation of T1w MRI images. Computer Methods and Programs in Biomedicine, **140**, 19-28, **2017**. DOI: https://doi.org/10.1016/j.cmpb.2016.11.011
6. SMOLIŃSKI A. ,FALKOWSKA L., PRYPUTNIEWICZ D. Chemometric exploration of sea water chemical component data sets with missing elements. Oceanological and Hydrobiological Studies, **3**, 49, **2008**. DOI: 10.2478/v10009-008-0005-1
7. BAILEY S. Principal Component Analysis with Noisy and/or Missing Data. Publications of the Astronomical Society of the Pacific, **124** (919), 1015, **2012**. DOI: 10.1086/668105
8. SERNEELS S., VERDONCK T. Principal component analysis for data containing outliers and missing elements. Copmutational Statistics and Data Analysis, **52** (3), 1712, **2008**. DOI: 10.1016/j.csda.2007.05.024
9. WALCZAK B., MASSART D.L. Dealing with missing data: Part I. Chemometrics and Intelligent Laboratory Systems, **58**, 15–27, **2001a**. http://doi.org/10.1016/S0169-7439(01)00131-9
10. WALCZAK B., MASSART D.L. Dealing with missing data: Part II. Chemometrics and Intelligent Laboratory Systems, **58**, 29, **2001b**. http://doi.org/10.1016/S0169-7439(01)00132-0
11. SMOLIŃSKI A., WALCZAK B. Exploratory analysis of chromatographic data-sets with missing elements. Initialization of the expectation-maximization algorithm. Acta Chromatographica, **12**, 30, **2002**.
12. STRAUSS R.E., ATANASSOV M.N., ALVES DE OLIVEIRA J. Evaluation of the principle-component and expectation-maximization methods for estimation of missing data in morphometric studies. Journal of Vertebrate Paleontology , **23** (2), 284, **2003**.
13. NAKAGAWA S., FRECKLETON R.P. Missing inaction: the dangers of ignoring missing data. Trends Ecology and Evolution, **23** (11), 592, **2008**. DOI: 10.1016/j.tree.2008.06.014
14. NEESER E., ACKERMANN R.R., GAIN J. Comparing the accuracy and precision of three techniques used for estimating missing landmarks when reconstructing fossil hominin crania. American Journal of Physical Anthropology, **140** (1), 1, **2009**. DOI: 10.1002/ajpa.21023.
15. COUETTE S., WHITE J. 3D geometric morphometrics and missing data. Can extant taxa give clues for the analysis of fossil primates? C. R. Palevol., **9**, 423, **2010**. DOI:10.1016/j.crpv.2010.07.002
16. BENTLER P.M., YUAN K.H. Positive Definiteness via Off-diagonal Scaling of a Symmetric Indefinite Matrix. Psychometrika, **76** (1), 119, **2011**. DOI: 10.1007/s11336-010-9191-3
17. JOSSE J., PAGÈS J., HUSSON F. Multiple imputation in principal component analysis. Adv Data Anal Classif, **5**, 231-246, **2011**. DOI 10.1007/s11634-011-0086-7
18. BROWN C. M., ARBOUR J. H., JACKSON D. A. Testing of the Effect of Missing Data Estimation and Distribution in Morphometric Multivariate Data Analyses. Syst. Biol., **61** (6), 941, **2012**. DOI:10.1093/sysbio/sys047
19. JOSSE J., HUSSON F. Handling missing values in exploratory multivariate data analysis methods. Journal de la Société Française de Statistique, **153** (2), 79, **2012**.
20. CLAVEL J., MERCERON G., ESCARGUEL G. Missing Data Estimation in Morphometrics: How Much is Too Much? Syst. Biol., **63** (2), 203, **2014**. DOI:10.1093/sysbio/syt100
21. HOU D., LIU S.,ZHANG J., CHEN F., HUANG P., ZHANG G. Online Monitoring of Water-Quality Anomaly in Water Distribution Systems Based on Probabilistic Principal Component Analysis by UV-Vis Absorption Spectroscopy. Journal of Spectroscopy, Article ID 150636, 9 pages, **2014**. DOI: http://dx.doi.org/10.1155/2014/150636
22. DRAY S., JOSSE J. Principal component analysis with missing values: a comparative survey of methods. Plant Ecology, **216** (5), 657, **2015**. DOI: 10.1007/s11258-014-0406-z
23. POLESZCZUK G., SADOWSKA B., KARPOWICZ K., GRZEGORCZYK K. Open water ecosystems of Wolin National Park – natural characterization. Baltic Coastal Zone, **7**, 37, **2002/2003**.
24. BUCIOR A., POLESZCZUK G. What happens in the waters of the Warnowo, Rabiąż, Czajcze and Domysłowskie Lakes in the Wolin National Park during summer stagnation?. Ecological Chemistry and Engineering, Series A, **20** (1), 7, **2013**. DOI:10.2428/ecea.2013.20(01)001
25. WAWRZYNIAK W., POLESZCZUK G., BUCIOR A., PIERWIENIECKI J., LASKOWSKI F., TYMANOWSKI Ł., GRYNFELDER K., RUTKOWSKA J. Wody powierzchniowe jezior Pojezierza Warnowsko-Kołczewskiego w Wolińskim Parku Narodowym – status troficzny wiosną 2012 roku. W: Zaborowski T (red.), Satori w publicznym bezpieczeństwie. Wyd. Inst. Badań i Ekspertyz Nauk. w Gorzowie Wlkp., Gorzów Wlkp.-Poznań, 350, **2012**.
26. JAŃCZAK J. (red.) Atlas jezior Polski – tom II. Jeziora zlewni rzek Przymorza i dorzecza dolnej Wisły. IMGW, Bogucki Wydawnictwo Naukowe, Poznań, 256, **1997**. ISBN: 83-86001-43-7
27. ISO 5667-4:2016. Water quality -- Sampling -- Part 4: Guidance on sampling from lakes, natural and man-made, **2016**.
28. HERMANOWICZ W., DOJLIDO J., DOŻAŃSKA W., KOZIOROWSKI B., ZERBE J. Fizyczno-chemiczne badanie wody i ścieków. Wyd. Arkady, 555, **1999**.
29. APHA, AWWA, WEF. Standard methods for the examination of water and wastewater. 20th ed. Washington, D.C.: APHA-AWWA-WEF, 1268, **1998**.
30. APHA. Standard methods for the examination of water and wastewater. 21st ed. Washington, D.C.: APHA, 1368, **2005**.
31. ISO 5667-3:2012. Water quality – Sampling – Part 3: Preservation and handling of water samples, **2012**.
32. WOLD S., ESBENSEN K., GELADI P. Principal Component Analysis. Chemometrics and Intelligent Laboratory Systems, **2**, 37, **1987**.
33. DASZYKOWSKI M., WALCZAK B., MASSART D.L. Projection methods in chemistry. Chemometrics and Intelligent Laboratory Systems, **65**, 97-112, **2003**. http://doi.org/10.1016/S0169-7439(02)00107-7
34. CRUZ A.G., CADENA R.S., ALVARO M.B.V.B., SANT'ANA A.S., OLIVEIRA C.A.F., FARIA J.A.F., BOLINI H.M.A., FERREIRA M.M.C. Assessing the use of different chemometric techniques to discriminate low-fat and full-fat yogurts. LWT – Food Science and Technology, **50**, 210, **2013**. http://dx.doi.org/10.1016/j.lwt.2012.05.023
35. KANNEL P.R., LEE S., KANEL S.R., KHAN S.P. Chemometric application in classification and assesment of monitoring locations of an urban river system. Analytica Chimica Acta, **582** (2), 390, **2007**. DOI: 10.1016/j.aca.2006.09.006
36. DEMPSTER A.P., LAIRD N.M., RUBIN D.B. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society, Series B (Methodological), **39**, 1, **1977**.

37. BOUKOUVALA F., MUZZIO F.J., IERAPETRITOU M.G. Predictive Modeling of Pharmaceutical Processes with Missing and Noisy Data. AIChE Journal, **56** (11), 2860, **2010**. DOI: 10.1002/aic.12203
38. SANTOS A., SANTOS R., SILVA M., FIGUEIREDO E., SALES C., COSTA J. C. W. A. A global Expectation-Maximization approach based on memetic algorithm for vibration-based structural damage detection. IEEE Transactions on Instrumentation and Measurement, **66** (4), 661, **2017**. DOI: 10.1109/TIM.2017.2663478
39. WOYANN L.G., BENIN G., STORCK L., TREVIZAN D. M., MENEGUZZI C., MARCHIORO V.S., TONATTO M., MADUREIRA A. Estimation of missing values affects important aspects of GGE biplot analysis. Crop Science, **57** (1), 40, **2016**. DOI: 10.2135/cropsci2016.02.0100
40. LI G., SHEN H. HUANG J.Z. Supervised sparse and functional principal component analysis. Journal of Computational and Graphical Statistics, **25** (3), 859-878, **2016**. DOI: http://dx.doi.org/10.1080/10618600.2015.1064434
41. STANIMIROVA I., WALCZAK B. Classification of data with missing elements and outliers. Talanta, **76** (3), 602, **2008**. DOI: 10.1016/j.talanta.2008.03.049
42. LIEW A.W.C., LAW N.F., YAN H. Missing value imputation for gene expression data: computational techniques to recover missing data from available information. Briefings in bioinformatics, **12** (5), 498, **2011**. DOI: https://doi.org/10.1093/bib/bbq080
43. ANDERSEN T., CARSTENSEN J., HERNANDEZ-GARCIA E., DUARTE C. M. Ecological thresholds and regime shifts: approaches to identification. Trends in Ecology and Evolution, **24** (1), 49, **2009**. DOI: 10.1016/j.tree.2008.07.014
44. HRON K., TEMPL M., FILZMOSER P. Imputation of missing values for compositional data using classical and robust methods. Computational Statistics and Data Analysis, **54** (12), 3095, **2010**. DOI:10.1016/j.csda.2009.11.023
45. SCHLOMER G. L., BAUMAN S., CARD N. A. Best practices for missing data management in counseling psychology. Journal of Counseling psychology, **57** (1), 1, **2010**. DOI: 10.1037/a0018082
46. BAŃBURA M., MODUGNO M. Maximum likelihood estimation of factor models on datasets with arbitrary pattern of missing data. Journal of Applied Econometrics, **29** (1), 133, **2014**.
47. ILIN A., RAIKO T. Practical approaches to principal component analysis in the presence of missing values. Journal of Machine Learning Research, **11**, 1957, **2010**.
48. JOLLIFFE I.T. Principal Component Analysis. 2rd Ed. Springer Verlag, New York, 488, **2002**.
49. LITTLE R.J., RUBIN D.B. Statistical analysis with missing data. John Wiley and Sons, **2014**.
50. KOWALKOWSKI T., ZBYTNIEWSKI R., SZPEJNA J., BUSZEWSKI B. Application of chemometrics in river water classification. Water Research, **40** (4), 744, **2006**. DOI: 10.1016/j.watres.2005.11.042
51. UKALSKI K., ŚMIAŁOWSKI T. Multivariate analysis of data from preliminary trials with winter rye, Biul. IHAR, **260/261**, 251, **2011**.
52. HOWANIEC N., SMOLIŃSKI A. Influence of fuel blend ash components on stream co-gasification of coal and biomass – Chemometric study. Energy, **78**, 814, **2014**. DOI: https://doi.org/10.1016/j.energy.2014.10.076
53. SMOLIŃSKI A., DROBEK L., DOMBEK V., BĄK A. Modeling of experimental data on trace elements and organic compounds content in industrial waste dumps. Chemosphere, **162**, 189-198, **2016**. DOI: https://doi.org/10.1016/j.chemosphere.2016.07.086
54. REN H., HOU Z., HUANG M., BAO J., SUN Y., TESFA T., LEUNG L. R. Classification of hydrological parameter sensitivity and evaluation of parameter transferability across 431 US MOPEX basins. Journal of Hydrology, **536**, 92, **2016**. DOI: https://doi.org/10.1016/j.jhydrol.2016.02.042
55. HOU Z., HUANG M., LEUNG L.R., LIN G., RICCIUTO D.M. Sensitivity of surface flux simulations to hydrologic parameters based on an uncertainty quantification framework applied to the Community Land Model. Journal of Geophysical Research, **117** (15), D15108, 2012. DOI: 10.1029/2012JD017521
56. RAY J., HOU Z., HUANG M., SARGSYAN K., SWILER L. Bayesian calibration of the Community Land Model using surrogates, SIAM/ASA J. Uncertainty Quantification, **3** (1), 199, **2015**. DOI: http://dx.doi.org/10.1137/140957998
57. GONG W., DUAN Q., LI J., WANG C., DI Z., DAI Y., YE A., MIAO C. Multi-objective parameter optimization of common land model using adaptive surrogate modeling, Hydrology and Earth System Sciences, **19** (5), 2409, **2015**. DOI: 10.5194/hess-19-2409-2015
58. BAO J., HOU Z., HUANG M., LIU Y. On approaches to analyze the sensitivity of simulated hydrologic fluxes to model parameters in the community land model. Water., **7**, 6810, **2015**. DOI: http://dx.doi.org/10.3390/w7126662
59. DRAY S., PETTORELLI N., CHESSEL D. Multivariate analysis of incomplete mapped data. Trans GIS, **7**, 411-422, **2003**. DOI: 10.1111/1467-9671.00153
60. LOURENÇO N.D., PAIXÃO F., PINHEIRO H. M., SOUSA A. Use of Spectra in the Visible and Near-Mid-Ultraviolet Range with Principal Component Analysis and Partial Least Squares Processing for Monitoring of Suspended Solids in Municipal Wastewater Treatment Plants. Applied Spectroscopy, **64** (9), 1061, **2010**. DOI: 10.1366/000370210792434332
61. DĄBROWSKI K. DĄBROWSKI K. Program ochrony środowiska dla gminy Wolin. W: BIP UM w Wolinie. Wyd. Związek Gmin Wyspy Wolin, Międzyzdroje, 169, **2005**.
62. JOSSE J., HUSSON F., PAGÈS J. Handling missing values in Principal Component Analysis, Journal de la Société Française de Statistique **150** (2), 28, **2009**.
63. GRAHAM J.W. Missing data analysis: Making it work in the real world. Annual review of psychology, **60**, 549, **2009**. DOI: 10.1146/annurev.psych.58.110405.085530
64. BRO R.A., SMILDE A.K. Principal component analysis, Anal. Methods, **6**, 2812, **2014**. DOI: 10.1039/c3ay41907j