Cent. Eur. J. Chem. • 9(2) • 2011 • 265-274
DOI: 10.2478/s11532-011-0004-z

VERSITA

Central European Journal of **Chemistry**

Assessment of river water quality in the South Baltic coast by multivariate techniques

Research Article

Monika Cieszynska[1*], Marek Wesolowski[2], Maria Bartoszewicz[1], Malgorzata Michalska[1]

[1]*Department of Environmental Protection and Hygiene of Transport, Medical University of Gdansk, 81-519 Gdynia, Poland*

[2]*Department of Analytical Chemistry, Medical University of Gdansk, 80-416 Gdansk, Poland*

Received 29 August 2010; Accepted 22 December 2010

Abstract: The paper presents an example of using multivariate techniques to interpret a large data set obtained during a 4-year water quality monitoring program in the Gdansk Municipality region, on the southern coast of the Baltic Sea. From 2004 to 2007, 11 physicochemical water parameters were analyzed monthly at 15 sites within eight watercourses. Principal-components analysis and cluster analysis were used to explore the data. Spatio-temporal trends in water quality were evaluated, the variables that determined the data set's structure and the factors that affected the water's physicochemical composition identified, with the goal of helping to optimize future monitoring. To reduce the number of analyzed variables, relationships between the analyzed parameters were also identified. The results revealed that the differences in physicochemical water properties among stations were generally smaller than those between the warmer and cooler seasons. It was determined that seasonal intrusions of brackish water from the Gulf of Gdansk can modify the water properties of some watercourses in the study area, but that dissolved oxygen, chemical oxygen demand, and total phosphorus were the main parameters responsible for the overall variation in the observed data. These parameters are related to pollution of anthropogenic origin.

Keywords: *Baltic Sea • Cluster analysis • Physicochemical parameters • Principal-components analysis • River quality*

1. Introduction

Water is the most important and widespread natural resource, and it is an essential compound taken in and excreted by all living organisms [1,2]. It also serves as a solvent, substrate, or catalyst for industrial chemical reactions. The quality of water has a profound impact on human lives because it is so commonly consumed and used by households [3-6]. Drinking water is often obtained from freshwater sources such as rivers that are vulnerable to municipal or industrial wastewater discharges and runoff from agricultural or contaminated land. Therefore, to minimize health hazards, the quality of river water should be constantly monitored and analyzed with the aim of acquiring reliable information about the level and trends of water pollution [7,8].

Long-term surveys and monitoring programs include frequent water sampling at various sites, followed by determination of the values of many parameters that are usually characterized by high variability. Consequently, monitoring studies generate a large and complex database of multidimensional results that are difficult to interpret. To explore the information included in this matrix of environmental data, different chemometric methods can be applied. These allow presentation and visualization of the raw analytical data, while reducing the number of data dimensions without losing important information [8-10]. To make this theory more concrete, it was applied in a case study of water quality in watercourses of Poland's Gdansk Municipality, which is located on the southern shore of the Baltic Sea.

Such monitoring studies tend to be costly and time-consuming. In the present study, chemometric techniques were used to optimize the research effort and to plan future sampling campaigns on the assumption that the number of analyzed samples or measured parameters must be limited due to cost or time constraints [11,12]. The examined watercourses are small; however, local

* *E-mail: Cieszynskam@gumed.edu.pl*

authorities continually monitor their water quality because they enter the Gulf of Gdansk, in the Polish coastal zone. The Baltic is an extremely sensitive, semi-enclosed brackish sea, therefore the Gulf of Gdansk is vulnerable to pollutants carried in river water, which include wastewater, organic matter, and toxic substances originating from human activities [13].

Chemometric analysis was applied to detect similarities and differences among the physicochemical properties of water collected from different sites and in different sampling seasons [9,14]. This knowledge can reveal which samples could potentially be omitted from future surveys without a significant loss of information [15]. It was also attempted to define the pollution levels in the monitored watercourses and to distinguish watercourses and sampling stations with water quality that differed significantly from the other rivers or stations. Such surveys can reveal the potential causes that underlie specific properties and factors or processes that control water quality, and that might therefore be essential to protect the health of local inhabitants [12,16].

It was attempted to establish whether some of the measured variables could be excluded from regular future monitoring because they exhibited similar variation to those of other parameters (*i.e.*, whether some parameters could be used as proxies for other parameters). In such a case, one variable could be monitored instead of a larger group of physicochemical parameters. Some of the analyzed parameters could also be eliminated from future surveys, if they had relatively little influence on the overall data structure [14,17].

2. Experimental Procedure

2.1 Description of the study area

The Gdansk Municipality lies on the southern coast of the Baltic Sea, in northern Poland, at an average elevation of 13 m above sea level. It is characterized by a temperate climate with cold, cloudy, and moderately severe winters, and by mild summers with frequent rain and thunderstorms; four seasons can be clearly differentiated. Mean daily temperatures range from -3.4°C in January (winter) to 21.3°C in August (summer). Monthly rainfall varies from 17.9 to 66.7 mm, with the highest and lowest number of rainy days in November and December (16 days; autumn) and in April (11 days; spring), respectively.

Fig. 1 shows the study area, including the monitoring stations used to provide study data. All watercourses in the study area (excluding the Vistula River, which is outside our study area) enter the Gulf of Gdansk *via* the Dead Vistula River, which is an artificial canal that intersects the Vistula River; only the Jelitkowski Stream flows directly into the Gulf of Gdansk. Table 1 lists the watercourses covered by the study, their general characteristics, as well as the geographical coordinates of the sampling stations.

The watercourses flow through areas that differ significantly not only with regard to their natural landscape, but also with respect to the type of land development in the surrounding area. Some of the sampling stations were situated in the woods, parks, and recreational areas that cover 24% of the city (the upper parts of the Strzyza Stream, both stations on Jelitkowski Stream, and the lower parts of the Siedlicki and Orunski streams), whereas others were located in areas covered by farming fields, privately-owned rural and suburban lands, and wastelands (the upper reaches of the Motlawa, Orunski, and Siedlicki streams). Gdansk is one of Poland's main industrial centers; therefore, some of the sampling stations were close to industrial facilities: Rozwojka is near an oil refinery, the lower Strzyza is near a shipyard and power station, and the Dead Vistula River is near waste dumps.

Moreover, the watercourses have different hydrological characteristics. The highest average flow was measured in the Dead Vistula River (600 to 1200 $m^3 s^{-1}$), which is the largest watercourse, with an average depth of 5 to 8 m, whereas the Rozwojka Canal and shallow Siedlicki and Orunski streams had the lowest average flows, ranging from 0.03 to 0.06 $m^3 s^{-1}$.

2.2 Water sampling

In this study, a total number of 675 surface-water samples were collected from 2004 to 2007 from eight watercourses situated in the Gdansk Municipality: the Strzyza Stream, Radunia Canal, Dead Vistula River, Rozwojka Canal, Motlawa River, Siedlicki Stream, Jelitkowski Stream, and Orunski Stream. Sampling sites were selected such that the one located farthest upstream was labelled by adding the letter "a" to its label. Each watercourse was sampled once per month. Two water samples from the main stream of each watercourse were collected at a depth of about 20 cm using a plastic scoop. Rozwojka Canal was an exception, because it was only sampled at one station. On rare occasions, it was not possible to collect samples from the main stream of certain watercourses (*i.e.*, the Dead Vistula and Motlawa rivers), and samples were instead collected at a distance equal to the maximum length of the scoop arm (about 3 m) from the bank.

Table 1. General characteristics of the watercourses studied in the Gdansk Municipality area, and details of the area surrounding the sampling stations and their locations.

Name of the watercourse	Sampling site	Geographical coordinates Longitude (E)	Latitude (N)	Site description	Average depth [m]	Length [km]	Average flow [$m^3 s^{-1}$]
Strzyza Stream	1	18.64°	54.38°	shipyard, power station (industrial area)	0.4	13.3	0.23
	1a	18.55°	54.35°	forest			
Radunia Canal	2	18.61°	54.32°	city center (urban area)	0.4	13.5	1.50
	2a	18.58°	54.23°	urban area with heavy vehicle traffic nearby			
Dead Vistula River	3	18.68°	54.36°	urban area	5.0 to 8.0	27.0	600.00 to 1200.00
	3a	18.76°	54.34°	mounds of phosphor-gypsum wastes and ashes from a power station (industrial area)			
Rozwojka Canal	4	18.72°	54.34°	oil refinery, heavy vehicle traffic nearby (industrial area)	1.0	4.6	controlled by inflow of wastewater
Motlawa River	5	18.66°	54.35°	city center (urban area)	2.0	65.0	6.80
	5a	18.65°	54.30°	fields (rural area)			
Siedlicki Stream	6	18.60°	54.30°	park in a city center	0.2	6.9	0.06
	6a	18.50°	54.25°	privately-owned rural, suburban lands, wastelands			
Jelitkowski Stream	7	18.60°	54.42°	park, beach (mouth lies on the Gulf of Gdansk)	0.3	9.7	0.25
	7a	18.51°	54.40°	forest, park			
Orunski Stream	8	18.58°	54.25°	park in a city center	0.2	7.5	0.03
	8a	18.50°	54.28°	fields, privately-owned rural, suburban lands and wastelands (rural area)			

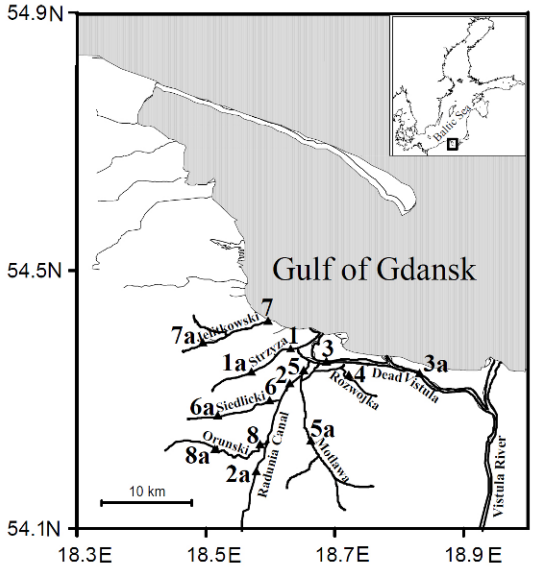

Figure 1. Map of the study area, and locations of the monitoring stations that provided data in the present study.

2.3 Parameters studied and analytical methods

For all surface water samples, the values of 11 parameters that characterized the water quality were determined: the total suspended solids content (TSS), dissolved oxygen concentration (DO), water temperature (T), oxygen saturation (OS), 5-day biochemical oxygen demand (BOD), chemical oxygen demand (COD), total phosphorus concentration (TP), total nitrogen concentration (TN), chloride concentration (Cl^-), pH of the water, and electrical conductivity (EC). Determinations were performed using the procedures recommended in [18]. Table 2 summarizes the units and abbreviations for the parameters, the analytical techniques, and the equipment used to measure these parameters. Samples were collected in polypropylene bottles, and all physicochemical water parameters were determined immediately upon arrival at the laboratory. Sample storage was avoided because that might have changed their chemical composition.

Table 2. List of physicochemical parameters measured for the river water.

Parameter	Units	Abbreviation used in the text	Applied analytical techniques	Equipment/reagents used
Water temperature	°C	**T**	thermometer	Model HI 145-20, HANNA instruments Inc., Woonsocket, RI, U.S.A.
5-day biochemical oxygen demand	mg dm^{-3} O_2	**BOD**	incubation, Winkler titration	All the analytical reagents were supplied by the POCH Joint-Stock Company, Poland, and by Merck Poland.
Chemical oxygen demand	mg dm^{-3} O_2	**COD**	permanganate titration	All the analytical reagents were supplied by the POCH Joint-Stock Company, Poland, and by Merck Poland.
Total nitrogen concentration	mg dm^{-3} N	**TN**	peroxydisulfate oxidation, cadmium-copper reduction, spectrophotometry	Spectronic Genesys 5 Spectrophotometer, Milton Roy Company, Rochester, N.Y., U.S.A. All the analytical reagents were supplied by the POCH Joint-Stock Company, Poland, and by Merck Poland.
Total phosphorus concentration	mg dm^{-3} P	**TP**	peroxydisulfate oxidation, spectrophotometry	Spectronic Genesys 5 Spectrophotometer, Milton Roy Company. All the analytical reagents were supplied by the POCH Joint-Stock Company, Poland, and by Merck Poland.
Oxygen saturation	%	**OS**	calculated based on the DO values after accounting for water temperature (T), atmospheric pressure, and water salinity	
Dissolved oxygen concentration	mg dm^{-3} O_2	**DO**	Winkler titration	All the analytical reagents were supplied by the POCH Joint-Stock Company, Poland, and by Merck Poland.
Total suspended solids	mg dm^{-3}	**TSS**	filtration and drying	filtration system, Millipore Sp. z o. o., Warsaw, Poland
pH	pH units	**pH**	pH meter	Model HI 9025, HANNA instruments Inc.
Electrical conductivity	μS cm^{-1}	**EC**	conductometer	Model HI 9835, HANNA instruments Inc.
Chloride concentration	mg dm^{-3} Cl^-	**Cl^-**	Mohr's titration	All the analytical reagents were supplied by the POCH Joint-Stock Company, Poland, and by Merck Poland.

2.4 Data analysis by means of multivariate statistical methods

To analyze the large database of more than 7000 measurements that were obtained during the monitoring program, two chemometric techniques were applied: principal-components analysis (PCA) and cluster analysis (CA).

PCA attempts to explain the relationships within a large dataset using a smaller set of orthogonal variables, called principal components (PCs), with the minimum loss of original information. PCs are weighted linear combinations of the standardized original variables. Identification of PCs relies on the fact that some of the measured variables are correlated and overlap. In such cases, some information included in the original variables is redundant (*i.e.*, some variables can be explained by the other variables, as they overlap), and hence, the overall dataset can be explained by fewer variables; the variables that do the best job of explaining the dataset are the PCs [8,16]. The strengths of the correlations between PCs and the original variables are given by their loadings; individual transformed observations are called scores. Each PC is described by a specific eigenvalue and the percentage of the variance it explains. In this paper, the number of significant PCs was established using an eigenvalue-based criterion of the standardized data set, and only PCs that accounted for variance greater than that of any single variable were retained (*i.e.*, with eigenvalues > 1). This criterion has been used by many authors, including [8,19].

The evaluation of water quality was also supported by an unsupervised pattern-recognition technique, namely agglomerative hierarchical cluster analysis (CA). In contrast with PCA, CA accounts for the total variation

in the dataset, and no simplification of the information is necessary [16,20,21]. CA classifies objects into clusters based on their similarity or difference. Grouping of the data is performed in such a way that objects with similar properties fall into the same cluster, which differs significantly from other clusters in the dataset [22]. The number of significant clusters was determined using the more restrictive criterion of Sneath's index, at 1/3 of the maximum distance, D_{max} [11]. All dendrograms that depicted object clustering in the CA were plotted using Ward's method, and the Euclidean distance was applied as the measure of similarity [23].

All data were initially standardized through z-score transformation [24]. Then statistical and mathematical calculations were performed using version 8.0 of the Statistica software (StatSoft Inc., Cracow, Poland) and Microsoft Office Excel 2003 spreadsheet (Microsoft Corporation, Warsaw, Poland).

3. Results and Discussion

A concise statistical description of 11 parameters analysed and values of guidelines are presented in Table 3. The results obtained in this study for most of the parameters measured seem to be rather satisfactory. Even though the values obtained are characterized by high variability, what is specific for environmental studies, the mean and median values of most parameters fall into the first class of water quality according to Polish standards for surface waters. The exception was noted for COD, TP, EC and Cl^-. Increased COD values indicate elevated levels of organic contamination in water or a high rate of organic matter production in the warm season. In the case of a high TP mean value, this may be attributed to site 3a on the Dead Vistula River and release of phosphates from the waste dump in Wislinka district, which could also increase observed high mean EC. For Cl^-, and to some extent EC, it is most likely that the elevated mean values were caused by seasonal intrusions of brackish water from the Gulf of Gdansk.

3.1 Principal-components analysis

Table 4 presents the PCA results for the entire study period and for individual years. In each case, more than 75% of the total variance was explained by the first four PCs. The first four significant PCs generally had eigenvalues >1, and were therefore retained for interpreting the information included in the original data set; PC4 in 2006 was only 0.92, but it was retained for consistency with the other analyses. For PCA carried out on the entire dataset, PC1 explained 31.0% of the overall variance and was highly influenced by seven variables. The highest contributions were observed for DO, COD, and TP (all magnitudes >0.7), whereas OS, BOD, Cl^-, and EC showed less influence. Therefore, variation in the data was attributed primarily to values related to the oxygen concentration in the water (DO and OS, with a positive loading, and COD and BOD, with a negative loading) and to dissolved ions (TP, Cl^-, and EC, with a negative loading). All these physicochemical water parameters are considered to be related to anthropogenic pollutants [9]. The authors of reference [9] also observed a strong influence of BOD, COD, and TP on the data structure of the Nakdong River watershed in South Korea.

PC2 explained 17.9% of the total variance, and was determined primarily by DO and OS (negative loadings, both with magnitudes > 0.5) as well as by Cl^- and EC (also negative loadings, both with magnitudes ≥ 0.68). The influence of DO and OS on PC2 was weaker than that on PC1, whereas the opposite was true for Cl^- and EC; that is, PC2 depended more strongly than PC1 on the latter variables. PC3 explained 14.6% of the total variance, and depended most strongly on TSS, BOD, and pH (negative loadings, all magnitudes >0.5). Based on this analysis, it can be concluded that the TSS values were mostly influenced by the intensity of organic matter production, which increases both BOD and pH [15,25,26]. The last significant PC, PC4, explained 12.5% of the total variance, and was most strongly influenced by water temperature (T, with a loading of 0.75) and total nitrogen (TN, with a loading of –0.68). T and TN seemed to influence the overall data structure less than the other parameters.

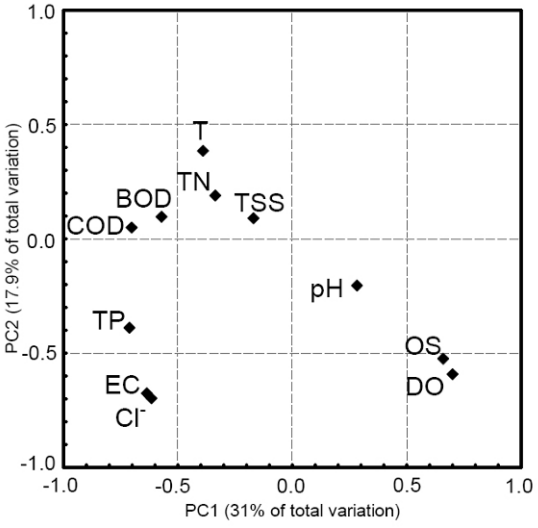

Figure 2. Loadings of the 11 physicochemical water parameters in the PC1–PC2 plane, calculated using the entire dataset.

Table 3. Statistical description of physicochemical parameters measured for the river water and recommended water quality criteria.

Parameter	Statistical description of measured parameters		Polish standards, which implement EU Council Directive 2000/60/WE: limit values for surface water quality		EU drinking water standards:(Council Directive 98/83/EC)	WHO guidelines for drinking water
			I class	II class		
T [°C]	Range Mean Median SD	0.0-25.7 11.2 11.6 6.9	≤22	24		
BOD [mg dm^{-3} O_2]	Range Mean Median SD	0.3-7.8 2.8 2.5 1.4	≤3	6		
COD [mg dm^{-3} O_2]	Range Mean Median SD	1.9-18.6 6.9 6.8 2.3	≤6	12	5.0	
TN [mg dm^{-3} N]	Range Mean Median SD	0.2-17.1 4.8 3.9 3.2	≤5	10		
TP [mg dm^{-3} P]	Range Mean Median SD	0.03-2.40 0.23 0.15 0.27	≤0.2	0.4		
OS [%]	Range Mean Median SD	4.3-121.0 80.0 84.3 19.6	Not defined	Not defined		
DO [mg dm^3 O_2]	Range Mean Median SD	2.3-14.5 9.1 9.2 2.8	≥7	5		
TSS [mg dm^{-3}]	Range Mean Median SD	0.5-198.0 10.1 6.1 14.7	≤25	50		Not defined
pH	Range Mean Median SD	6.36-8.65 7.84 7.84 0.32	6.0-8.5	6.0-9.0	6.5-9.5	6.5-9.5
EC [μS cm^{-1}]	Range Mean Median SD	67-18822 1643 666 2393	≤1000	1500	2500	2500
Cl^- [mg dm^{-3} Cl^-]	Range Mean Median SD	6.3-5934 361 55.4 779	≤200	300	250	250

Table 4. PCA results for the whole study period (TOTAL) and for the individual years (2004 to 2007). The table presents variable loadings, eigenvalues, and the proportion of the variance explained for the first four PCs. Boldfaced values represent parameters with significant loadings.

Studied variables	TOTAL (675×11)				2004 (135×11)				2005 (180×11)				2006 (180×11)				2007 (180×11)			
	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4	PC1	PC2	PC3	PC4
TSS	-0.17	0.09	-0.70	0.00	-0.01	-0.11	0.64	0.01	-0.13	-0.38	-0.52	-0.40	-0.13	0.17	-0.75	-0.45	-0.31	-0.03	-0.27	0.69
DO	0.71	-0.59	-0.27	-0.23	0.71	-0.61	-0.26	-0.18	0.53	-0.78	0.01	0.30	0.84	-0.37	-0.19	-0.26	0.57	0.72	0.10	0.35
T	-0.39	0.38	0.06	0.75	-0.36	0.28	0.31	0.72	-0.23	0.41	0.25	-0.76	-0.61	0.25	-0.24	0.64	-0.19	-0.8	0.47	-0.02
OS	0.66	-0.53	-0.31	0.22	0.66	-0.63	-0.14	0.22	0.54	-0.73	0.20	-0.15	0.73	-0.32	-0.44	0.11	0.60	0.38	0.46	0.43
BOD	-0.57	0.10	-0.56	0.16	0.03	-0.47	0.52	-0.19	-0.58	-0.27	-0.46	-0.33	-0.70	0.17	-0.54	0.03	-0.63	-0.34	0.12	0.54
COD	-0.70	0.05	-0.42	-0.16	-0.45	-0.22	0.66	-0.35	-0.73	-0.12	-0.42	0.07	-0.82	0.02	-0.29	-0.08	-0.64	0.14	-0.24	0.37
TP	-0.71	-0.39	-0.01	-0.04	-0.63	-0.31	-0.21	-0.03	-0.78	-0.30	0.17	0.11	-0.63	-0.49	-0.11	-0.06	-0.82	0.23	0.06	0.07
TN	-0.34	0.19	-0.26	-0.68	0.03	0.16	0.12	-0.80	-0.17	0.15	-0.64	0.52	-0.65	0.34	0.14	-0.39	-0.28	0.44	-0.64	-0.07
Cl^-	-0.62	-0.70	0.19	0.10	-0.73	-0.58	-0.25	0.03	-0.75	-0.36	0.45	0.13	-0.50	-0.81	0.06	0.01	-0.58	0.47	0.57	-0.21
EC	-0.64	-0.68	0.21	0.07	-0.74	-0.56	-0.23	0.04	-0.78	-0.25	0.42	0.16	-0.51	-0.82	0.05	0.01	-0.62	0.50	0.50	-0.25
pH	0.28	-0.21	-0.55	0.42	0.39	-0.46	0.51	0.38	0.10	-0.56	-0.17	-0.44	0.52	-0.11	-0.56	0.27	0.23	-0.11	0.46	0.40
Eigenvalues	3.41	1.97	1.61	1.38	2.91	2.10	1.72	1.55	3.33	2.16	1.59	1.48	4.38	2.07	1.56	0.92	3.14	2.17	1.81	1.49
% of variance explained	31.0	17.9	14.6	12.5	26.5	19.1	15.7	14.1	30.2	19.6	14.5	13.5	39.8	18.8	14.2	8.39	28.5	19.7	16.4	13.4
% of cumulative variance (sum of the values for the individual PCs)	31.0	48.9	63.5	76.1	26.5	45.6	61.3	75.3	30.2	49.8	64.3	77.8	39.8	58.6	72.8	81.2	28.5	48.3	64.7	78.2

Fig. 2 shows the loading plot for all 11 variables using the overall dataset from 2004 to 2007. As shown in Table 4, the first two PCs accounted for 48.9% of the overall variance. Based on this plot, EC and Cl^-, which describe the mineral content and salinity of water, had the most similar variation structures [16,27]. These two variables were also strongly and significantly correlated; Pearson's correlation coefficient (p=0.05) ranged from 0.87 for the 2005 samples to 0.99 for the 2004 samples. Based on these results, Cl^- could be eliminated from subsequent analysis and not considered in future monitoring studies, as was previously suggested by [14] for a river in Argentina.

DO and OS also had similar loadings for the first two PCs. This can be explained by the fact that OS was calculated based on the DO value after accounting for water temperature, atmospheric pressure, and salinity [28,29]. COD and BOD also had similar loadings, since BOD accounts for biotic decomposition of organic matter, whereas COD represents organic matter that is oxidized by strong oxidants [15,30].

Fig. 3 represents a score plot for all the analyzed water samples in the plane defined by the first two PCs. The large number of objects (*i.e.*, 675) projected on the plane defined by PC1 and PC2 made it difficult to classify these objects and detect groups of similar objects. Similar observations have been made by [11]. However, some patterns could still be detected. For example, the distribution of the water samples was influenced by the season of sample collection. Samples collected approximately from May to September, in the warm (spring and summer) season, are located in the upper part of the plot for positive values of PC2, whereas samples collected approximately from October to April, in the cool (autumn and winter) season, appear mostly in the lower part of the plot for negative values of PC2 and positive values of PC1. These seasonal patterns were mainly detected for DO and OS; the maximum values of both parameters were observed during the cool season, and were accompanied by the lowest values of COD and BOD. This phenomenon was confirmed by the high positive loadings of DO and OS for PC1 and the moderate negative loadings for PC2, with the opposite pattern for COD and BOD (Fig. 2). The projection of the samples in the PC1–PC2 plane suggested that the differences in physicochemical water properties among sites were smaller than the differences between seasons. An exception for two watercourses was noted: data for the Dead Vistula River (site 3a) were clustered primarily in the lower left quadrant, with negative scores for both PCs, whereas data for Rozwojka Canal (site 4) were located primarily in the upper left quadrant, with

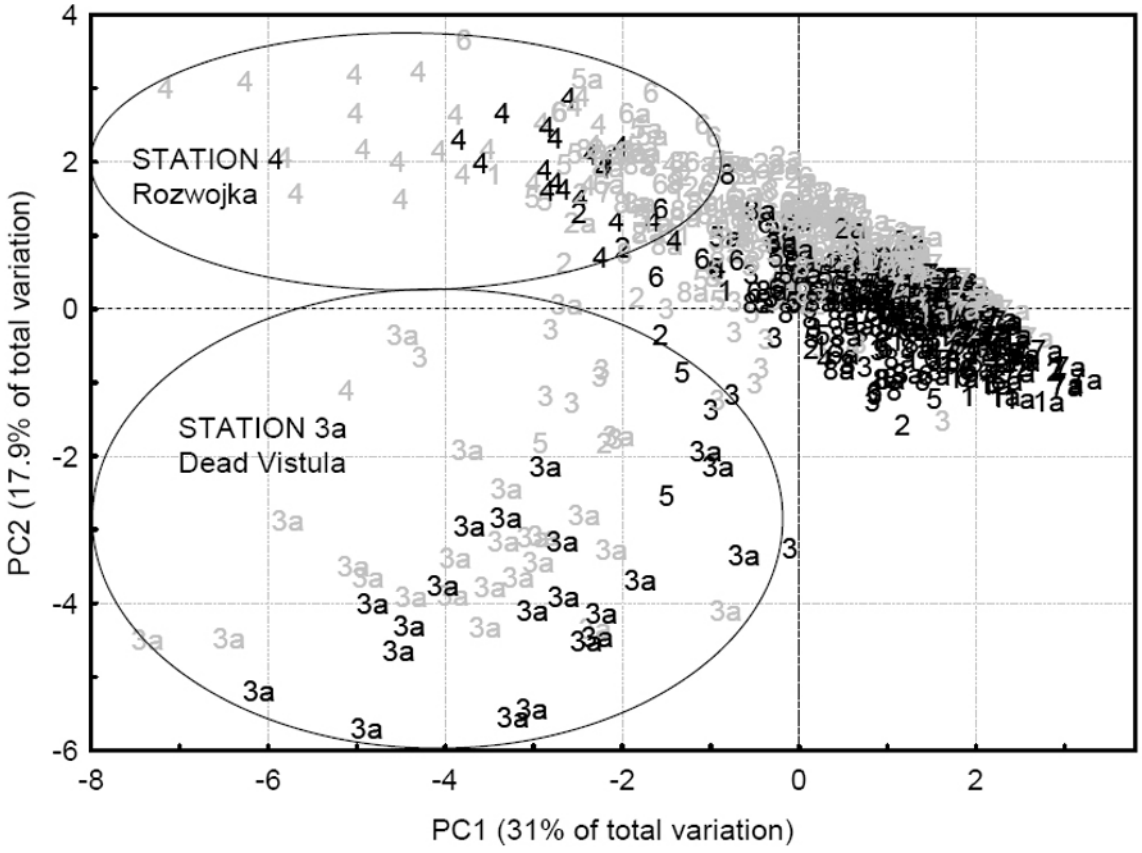

Figure 3. Score plot representing the projection of the entire data set (n = 675 samples, 11 parameters) in the PC1–PC2 plane. Samples collected in the warm (spring and summer) period are shown in grey, whereas samples from the cool (autumn and winter) period are shown in black. Numbers represent the sampling stations in Table 1.

Table 5. Composition of the clusters detected by the cluster analysis of water samples collected from 2004 to 2007; 11 water parameters were determined at each sampling site. Site locations and characteristics are described in Table 1. Data in the table represent the number of samples that were grouped into a given cluster.

Watercourse	Years									
	2004			2005			2006		2007	
	Clusters			Clusters			Clusters		Clusters	
	I	II	III	I	II	III	I	II	I	II
Strzyza Stream (1)	9			2	10		12		12	
Strzyza Stream (1a)	9			1	11		12		12	
Radunia Canal (2)	9				12		10	2	12	
Radunia Canal (2a)	9				12		12		12	
Dead Vistula River (3)	1	8		12			7	5	9	3
Dead Vistula River (3a)			9	1		11		12	1	11
Rozwojka Canal (4)	2	7		11		1	12		12	
Motlawa River (5)	5	4		9	3		9	3	12	
Motlawa River (5a)	9				12		12		12	
Siedlicki Stream (6)	9			2	10		12		12	
Siedlicki Stream (6a)	9			1	11		12		12	
Jelitkowski Stream (7)	9				12		12		12	
Jelitkowski Stream (7a)	9				12		12		12	
Orunski Stream (8)	9			1	11		12		12	
Orunski Stream (8a)	9			1	11		12		12	
TOTAL	107 (79%)	19 (14%)	9 (7%)	41 (23%)	127 (70%)	12 (7%)	158 (88%)	22 (12%)	166 (92%)	14 (8%)

negative scores for PC1 and positive scores for PC2. Samples from these sites were less variable with regard to the season of sample collection. The high negative loading of EC and Cl^- for the first two PCs (Fig. 2) suggested that salinity and mineral content of the water could have been influenced by the sampling season.

The Dead Vistula River (site 3a) and Rozwojka Canal (site 4) samples appeared to differ noticeably from those collected from the other watercourses. Data points for the Dead Vistula and Rozwojka were characterized by high EC and Cl^- values. Moreover, samples collected from Rozwojka tended to have lower DO and OS, and high BOD and COD values (which is typical for the warm season). Samples from site 3a on the Dead Vistula River were extremely polluted: for example, they contained the highest TP concentrations of all the samples. Some samples collected from site 5 on Motlawa and site 3 on the Dead Vistula also displayed distinctive characteristics because these watercourses are influenced by seasonal inflows of water from the Gulf of Gdansk, which increase the EC and Cl^- values.

PCA allowed identifying important pollution sources in the Gdansk Municipality area. In the Dead Vistula River (site 3a), the elevated concentrations of TP and EC (higher than in the waters of the Gulf of Gdansk) may be caused by the release of phosphates from a phosphogypsum waste dump in the Wislinka district near station 3a. Considering the rapid water flow in the Dead Vistula River (600 to 1200 $m^3 s^{-1}$) and its depth of more than 5 m, it can be concluded that the amount of phosphates and other ions delivered from the waste dump into the Dead Vistula should not be neglected. The elevated Cl^- concentrations, on the other hand, can be attributed to intrusions of brackish water from the Gulf of Gdansk.

The characteristics of water quality in the Rozwojka Canal samples (site 4) arise from the fact that sampling station 4 was located near the Gdansk oil refinery. Rozwojka Canal receives an inflow of drainage waters and treated wastewater from the refinery's wastewater treatment plant. Effluents from the refinery cannot be quickly flushed due to the relatively low water flow, which includes periods with no flow, in Rozwojka Canal.

3.2 Cluster analysis

Table 5 presents the results of the CA, which confirmed the PCA results. Based on CA, the analyzed water samples also had low variability among sampling sites. In 2004 and 2005, the samples clustered into three groups, whereas the samples collected in 2006 and 2007 formed only two clusters. Similarly to the PCA outcome, the physicochemical properties of water samples collected from the Dead Vistula River (stations 3 and 3a), Rozwojka Canal (station 4), and Motlawa River (station 5) differed enough from those at other stations to form separate clusters each year. The remaining samples were grouped into one large cluster that contained from 70% (2005) to 92% (2007) of the total number of samples. Within this cluster, some trends in the distribution related to the sampling season in each year were noted, but an overall pattern could not be determined.

3.3 Synthesis of the PCA and CA results

The application of PCA and CA provided useful information about the data matrix structure that can support optimization of the future monitoring strategy for these watercourses. The physicochemical properties of most of the samples showed relatively little variance among sampling sites, but noticeable variability between sampling seasons (warm *vs.* cold). Therefore, if time and money constraints limit the number of samples that can be examined in a monitoring program, samples could be collected less frequently (*e.g.* once every 2 months). This is however not applicable to the Dead Vistula River (station 3a) and Rozwojka Canal (station 4) as samples from those stations varied less between the seasons than other sites. The number of sampling sites within each watercourse should be also reduced to only one, but in case of The Dead Vistula River (sites 3 and 3a) and Motlawa River (sites 5 and 5a) the two stations (*i.e.*, upstream and lower) have such dissimilar properties that samples from both sites must always be analyzed and the number of sampling should not be reduced at either site.

4. Conclusions

Both multivariate techniques (principal-components analysis and cluster analysis) revealed that for large environmental data sets such as the one in the present study (n > 7000), characterized by high variation, detection of relationships and similarities among data objects may be challenging, but patterns can be detected that might otherwise not be visible. In particular, the analysis revealed sampling stations with distinctive water parameters (*e.g.* Dead Vistula River, Rozwojka Canal). Poor water quality at these sites was mainly attributed to the mode of land use in the surrounding area (industrial) and the hydrological properties of the watercourse (flow and depth). In the Motlawa River (site 5) and Dead Vistula River (site 3) samples, the water composition was substantially modified by intrusions of brackish water from the Gulf of Gdansk.

The main variables responsible for the overall water quality were DO, COD, and TP, followed by OS, BOD, Cl^-, and EC to a lesser extent. All these water parameters are related to anthropogenic pollution.

PCA and CA showed that the differences among sites in the physicochemical properties of the water samples were smaller than those between sampling seasons. This finding can be used to optimize the future monitoring strategy so as to reduce the number of samples collected and analyzed with minimal loss of information.

References

[1] A.J. Bates, Food Chem. Toxicol. 38, S29 (2000)
[2] S. Dolnicar, A.I. Schäfer, J. Environ. Manage. 90, 888 (2009)
[3] V.V Goncharuk, J. Water Chem. Technol. 30, 129 (2008)
[4] M.S. Holt, Food Chem. Toxicol. 38, S21 (2000)
[5] L. Petraccia, G. Liberati, S.G Masciullo, M. Grassi, A. Fraioli, Clin. Nutr. 25, 377 (2006)
[6] F.X. R. Van Leeuwen, Food Chem. Toxicol. 38, S51 (2000)
[7] S.W. Liao, H.S. Gau, W.L Lai, J.J Chen, C.G. Lee, J. Environ. Manage. 88, 286 (2008)
[8] K.P. Singh, A. Malik, V.K. Singh, Water Air Soil Pollut. 170, 383 (2005)
[9] S. Han, E. Kim, S. Kim, KSCE J. Civ. Eng. 13(2), 97 (2009)
[10] V. Simeonov, J.W. Einax, I. Stanimirova, J. Kraft, Anal. Bioanal. Chem. 374, 898 (2002)
[11] A. Astel, S. Tsakovski, P. Barbieri, V. Simeonov, Water Res. 41, 4566 (2007)
[12] V. Simeonov, J.A. Stratis, C. Samara, G. Zachariadis, D. Voutsa, A. Anthemidis, M. Sofoniou, T. Kouimtzis, Water Res. 37, 4119 (2003)
[13] G.P. Glasby, P. Szefer, Sci. Total Environ. 212, 49 (1998)
[14] P.M Castañé, M.G Rovedatti, M.L Topalián, A. Salibián, Environ. Monit. Assess. 117, 135 (2006)
[15] P.R. Kannel, S. Lee, Y.S Lee, J. Environ. Manage. 86, 595 (2008)
[16] M. Vega, R. Pardo, E. Barrado, L. Debán, Water Res. 32, 3581 (1998)
[17] J.R. King, D.A. Jackson, Environmetrics 10, 67 (1999)
[18] APHA, Standard Methods for the Examination of Water & Wastewater (American Public Health Association, Washington D.C., 2005)
[19] E.M. Andrade, H.A.Q Palácio, I.H. Souza, R.A.O. Leão, M.J. Guerreiro, Environ. Res. 106, 170 (2008)
[20] T. Hill, P. Lewicki, Statistics, methods and applications: A comprehensive reference for science, industry, and data mining (StatSoft, Inc., Tulsa, 2006)
[21] M. Otto, Chemometrics: Statistics and computer application in analytical chemistry (Wiley-VCH, New York, 1999)
[22] S. Shrestha, F. Kazama, Environ. Modell. Softw. 22, 464 (2007)
[23] U.C. Panda, S.K. Sundaray, P. Rath, B.B. Nayak, D. Bhatta, J. Hydrol. 331, 434 (2006)
[24] E. Marengo, M.C. Gennaro, D. Giacosa, C. Abrigo, G. Saini, M.T. Avignone, Anal. Chim. Acta. 317, 53 (1995)
[25] D. Bellos, T. Sawidis, J. Environ. Manage. 76, 282 (2005)
[26] E. Perona, I. Bonilla, P. Mateo, Sci. Total Environ. 241, 75 (1999)
[27] T.G. Kazi, M.B. Arain, M.K. Jamali, N. Jalbani, H.I. Afridi, R.A. Sarfraz, J.A. Baig, A.Q. Shah, Ecotoxicol. Environ. Safety 72, 301 (2009)
[28] H. Chang, Water Air Soil Poll. 161, 267 (2005)
[29] A.J. M. Yunus, N. Nakagoshi, Chin. Geogr. Sci. 14, 153 (2004)
[30] C.A. Almeida, S. Quintar, P. González, M.A. Mallea, Environ. Monitor. Assess. 133, 459 (2007)