DOI: 10.2478/v10216-011-0016-9 ECOL CHEM ENG S. 2012;19(2):213-226

Petia PAPAZOVA[1] and Pavlina SIMEONOVA[1*]

LONG-TERM STATISTICAL ASSESSMENT OF THE WATER QUALITY OF TUNDJA RIVER

DŁUGOOKRESOWA OCENA JAKOŚCI WODY RZEKI TUNDJI

Abstract: Two major environmetric methods (*Cluster analysis* (CA) and *Principal components analysis* (PCA)) were applied for statistical assessment of the water quality of trans-border river Tundja. The study used long-term monitoring data from 26 sampling sites characterized by 12 physicochemical parameters. Clustering of chemical indicators results in 3 major clusters: the first one shows the impact of anthropogenic sources, the second - the impact of agriculture and farming activities and the last one describes the role of the physical parameters on the water quality and also the impact of urban wastes. For better assessment of the monitoring data, PCA was implemented, which identified four latent factors. Two of them - "urban wastes" factor and "agriculture" factor correspond almost entirely to clusters 3 and 2 from the previous statistical analysis. The third one, named "industrial wastes" factor, reveals a specific seasonal behavior of the river system. The last latent factor describes the active reaction of the water body and is determined as "acidity" factor. The linkage of the sampling sites along the river flow by CA formed two clusters with the spatial "upstream-downstream" separation. The apportionment model of the pollution determined the contribution of each one of identified pollution factors to the total concentration of each one of the water quality parameters.

Keywords: monitoring, river water, data treatment, cluster analysis, principal components analysis

Introduction

The assessment of the river water quality is usually based on the comparison of measured monitoring values of particular physicochemical parameters with the allowable threshold values defined in national or international directives. A much more sound and reliable approach seems to be the application of chemometric methods for classification and data interpretation of Bulgarian river water monitoring results since they consider the environmental system as a multivariate one and treat it respectively [1-15].

Usually, the studies performed try to assess the river water quality or to optimize the monitoring procedure by classifying the sampling locations, by revealing links between the water quality parameters, by identifying possible sources of pollution, by modeling the contribution of the identified sources to the formation of the total concentration of the

[1] Laboratory of Environmental Physics, Georgi Nadjakov Institute of Solid State Physics, Bulgarian Academy of Sciences, Tzarigradsko Chaussee Blv. 72, 1784 Sofia, Bulgaria, phone ++35929746265
[*] Corresponding author: poly-sim@issp.bas.bg

monitored chemical tracers. The goal of the present study is to classify, model and interpret monitoring data from Tundja River catchment in Bulgaria collected in a long-time period using multivariate statistics in order to assess in a reliable way the river water quality.

Experimental

Monitoring data collection

The Tundja River is part of the Maritza sub-basin, including Arda and Ergene tributaries, and one of the major river systems located in the eastern Balkans. It has a length of 350 km and the catchment area is 7884 km^2 in Bulgaria. Main cities along the catchment on Bulgarian territory are Kazanlak, Sliven and Yambol. The river then crosses into Turkey as Tunca (200 km^2) before flowing into Maritza river at the Greek-Turkish border near to the city of Edirne (Turkey).

The River Tundja springs out in the Kalofer part of the Stara Planina Mountain. In its upper reaches the river flows southwards and upon passing Kalofer turns towards the east to flow across the Kazanlak, Sliven, and Straldja lowlands. Near the village of Zavoi, the river turns southward to flow across the Yambol - Elhovo valley.

The River Tundja forms its runoff from the Central part of Eastern Stara Planina and from the Northern slopes of the Sredna Gora Mountain. The mid and downstream sections of the river cross the Kazanlushka valley, as well as several fields and low lands located in the eastern part of Southern Bulgaria.

In the Tundja river basin there are 252 settlements, and total population is 520,900 people. Population of principal cities or towns is: Sliven (111,301), Yambol (78,302), Kazanlak (60,764), Karnobat (19,315), Elhovo (10,846), Pavel Bania (3074).

The Tundja River has 44 tributaries with a total length of 393.9 km. The main tributaries are Mochuritca River (catchment area ca 1278 km^2), Asenovska River (catchment area ca 89.7 km^2), Marash River (catchment area ca 74.5 km^2) and Eninska River (catchment area ca 45.2 km^2).

A very complex water resource system is created consisting of four big reservoirs - “Koprinka“, “Zrebchevo”, “Asenovets”, “Malko Sharkovo”, three smaller ones - “Ts. Tserkovski”, “Kirilovo” and “Dva Chouchoura”, five hydroelectric power plants, four large-scale irrigation systems, many irrigation fields, pumping stations, water supply groups, numerous small reservoirs, water intakes, many pits, river fisheries.

The main land uses include park land and protected fauna/flora areas, skiing, forestry, grazing, dry and irrigated agriculture, hydroelectricity, urban and scattered industry, fish farming and ponds and coastal tourism. Industrial emitters in Tundja River basin are distributed as follows: production and processing of metals - 20%; chemical industry - 60%; and intensive livestock production - 20%.

Monitoring of surface water is part of a *National Environmental Monitoring System* (NEMS) and includes programs for control and operational monitoring. The system is managed by the Minister of Environment and Water through the *Executive Environment Agency* (EEA). All measurements and observations are carried out by the structures of the EEA in common, unified methods for sampling and analysis in accordance with the procedures ensuring the quality of measurements and data. All EEA laboratories are accredited under the BS EN ISO/IEC 17025-(General requirements for competence in testing and calibration from EA BAS).

The data set used for the aims of the present study is part of NESM and involved 26 sampling sites characterized by 12 parameters - active reaction (pH), water temperature (T) [ºC], dissolved oxygen (O_2) [mg/dm^3], oxygen saturation [%], conductivity [mS/cm], non-dissolved matter [mg/dm^3], ammonia nitrogen (NH_4-N) [mg/dm^3], nitrate(V) nitrogen (NO_3-N) [mg/dm^3], orthophosphates (PO_4) [mg/dm^3], nitrate(III) nitrogen (NO_2-N) [mg/dm^3], *biological oxygen demand* (BOD) [mg/dm^3], *chemical oxygen demand* (COD) [mg/dm^3]. The analytical determination of the water indicators was performed according to the respective local and international standard methods.

All water samples were collected in the period between 2004 and 2009.

The catchment of Tundja River and the monitoring net of the river are presented in Figure 1.

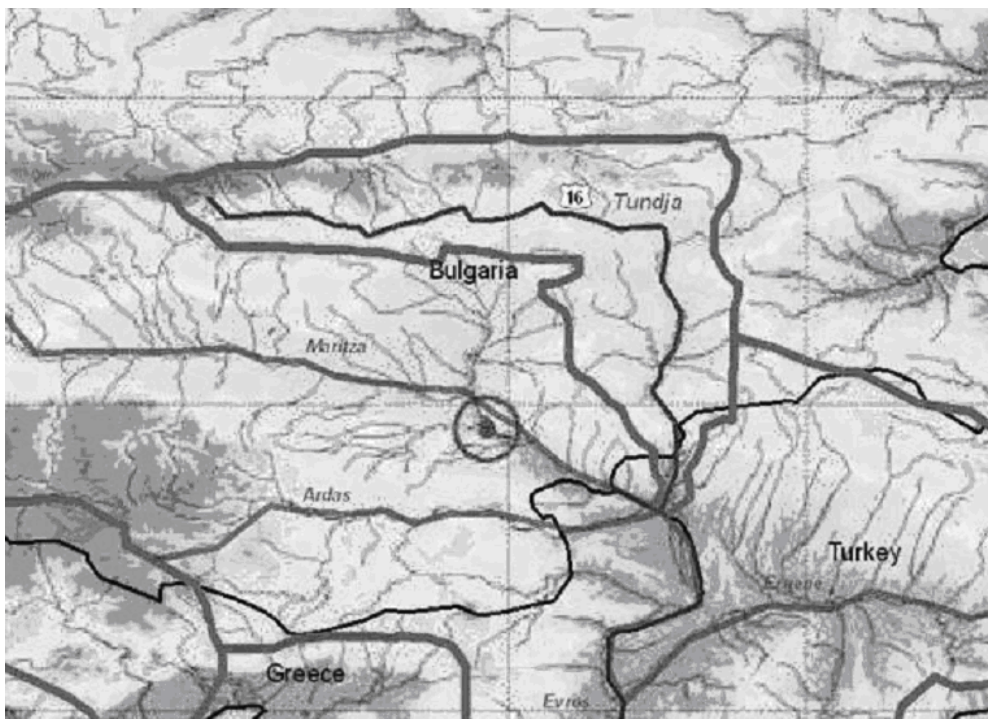

Fig. 1. The Tundja River catchment and monitoring net

Environmetric methods

In the data treatment two major environmetrics approaches were used: cluster analysis and principal components analysis.

Cluster analysis [16] is a well-known and widely used classification approach. In order to cluster objects characterized by a set of variables one has to determine their similarity. A preliminary step of data scaling is necessary, where normalized dimensionless numbers replaces the real data values in order to eliminate dimension differences. Then, the similarity (or the distance) between the objects in the variable space can be determined.

Very often the Euclidean distance is used for clustering purposes. Thus, from the input matrix (raw data) a similarity matrix is calculated. There is a wide variability of hierarchical algorithms but the typical ones include the single linkage, the complete linkage and the average linkage methods. The representation of the results of the cluster analysis is usually performed by a tree-like scheme called dendrogram comprising a hierarchical structure (large groups are divided into small ones).

Principal components analysis (PCA) [17] is a typical display method, which allows estimating the internal relations in the data set. There are different variants of PCA but basically, their common feature is that they produce linear combination of the original columns in the data matrix (data set) responsible for the description of the variables characterizing the objects of observation. These linear combinations represent a type of abstract measurements (factors, principal components) being better descriptors of the data structure (data pattern) than the original (chemical or physical) measurements. Usually, the new abstract variables are called latent factors and they differ from the original ones named manifest variables. It is a common finding that just a few of the latent variables account for a large part of the data set variation. Thus, the data structure in a reduced space can be observed and studied. The new coordinates are called factor scores and the regression coefficients from the linear combination of the old variables - factor loadings.

In case of many studies related to natural ecosystems PCA and other multivariate statistical techniques are used to determine possible natural or anthropogenic influences in the formation of the determinants total mass. However, PCA does not provide a direct balancing and apportionment. After the pollution sources identification by the application of PCA, the next calculation step in modeling and balancing of pollution impacts is the apportioning itself. It is performed mostly by *absolute principal components analysis* (APCA). The procedure introduced by Thurston and Spengler [18] is well developed and often applied for apportionment purposes, mainly in apportionment of airborne particulate matter. However, recent applications of the approach proved its effectiveness in apportionment monitoring studies for other environmental compartments like surface water, soils, sediments, and biota.

The first step in the source apportionment methodology of Thurston and Spengler is performing of principal components analysis. The PCA assumes that the total concentration of each element is made up of the sum of elemental contributions from each of *f* pollution source components. Hence

$$Z_{ik} = \sum W_{ij} P_{jk} \quad (\text{for } j = 1\ldots p)$$

where: P_{jk} is the *j*th component's value for observation *k*; $j = 1\ldots p$ is the number of pollution sources influencing the data and W_{ij} is the coefficient matrix of the components.

The first step in the derivation of source impacts is to calculate component scores for each sample (object). Rotated PC coefficients, *B**, are calculated by applying the rotation transformation matrix [T] to [B]

$$[B]^{*}_{pxn} = [B]_{pxn}[T]_{nxn}$$

Rotated PC scores are computed using the transformed [B] matrix

$$[P]^{*}_{pxm} = [B]^{*}_{pxn}[Z]_{nxm}$$

These PC scores are correlated with their respective pollution source impacting the site (ie a higher component score P_{jk}^* implies a higher pollution impact by the pollution source *j* during observation *k*). However, because they are computed from the normalized elemental concentrations Z_{ik}, they too are normalized. Each component indicates deviations from the mean source impacts; they are not proportional to these pollution impacts.

It has been shown that the regression of a dependent variable Y_k on the daily scores of components P_{jk} could be presented by the formula

$$Y_k = Y_a + \sum \zeta_j P_{jk} \quad (\text{for } j = 1\ldots p)$$

where Y_a equals the mean of Y_k. If the dependent variable Y_k is the total mass (for air particulate matter, in [μg m^{-3}]), then ζ_j are the conversion coefficients of the non-dimensional PC score deviations into mass deviations from the mean source impact. Since the components are not scored as deviations from zero, but instead as deviations from the mean, this results in the presence of Y_a in the equation.

As the factor scores obtained from PCA are normalized, with mean zero and standard deviation equal to unity, the true zero for each factor score is calculated by introducing an artificial sample with concentration equal to zero for all variables.

$$(Z_0)_i = \frac{(0 - C_i)}{s_i} = -\frac{C_i}{s_i}$$

where: C_i - arithmetic mean concentration of analyte *i* (understood as feature), s_i - standard deviation of variable *i*.

Then the rotated absolute zero PC scores, P_0^* for each of *p* components are calculated

$$P_{0p}^* = \sum B_{pi}^* (Z_0)_I \quad (\text{for } i = 1\ldots n)$$

These estimates of the PC scores for each component at absolute zero are then used to estimate *Absolute PC Scores* [APCS] for each component on each sampling day as follows:

$$[APCS]_{pxj}^* = [P]_{pxj}^* - [P_0]_{pxj}^*$$

where the *j* columns of *[P0]** are all identically equal to the values calculated for P_{0p}^*. It can be proved in a straight forward manner that the calculation for *[APCS]** gives the exact score which would be achieved had the original scoring been executed using unnormalized data.

Regressing (multiple linear regression) the monitoring results on these APCS give estimates of the coefficients which convert the APCS into pollutant source mass contributions (in [μg m^{-3}]) from each source for each sample.

The source contributions to C_i can be calculated by mentioned above linear regression procedure according to the following:

$$C_i = (b_0)_i + \sum APCS_p \cdot b_{pi}, \quad p = 1,2,\ldots,n$$

where: $(b_0)_i$ - constant term of multiple regression for variable *i*, b_{pi} - the coefficient of multiple regression of the source *p* for variable *i*, $APCS_p$ - scaled value of the rotated factor *p* for the considered sample, $APCS_p \cdot b_{pi}$ represents the contribution of source *p* to C_i.

The mean of the product $APCS_p \cdot b_{pi}$ on all samples represents the average contribution of the sources. The method estimates source profiles and contributions but its serious disadvantage is error propagation in centering and uncentering of data. This balancing approach accepts that all sources have been identified by the principal components analysis and all of them participate in the source contribution procedure.

All statistical calculations were performed by the use of the software package STATISTICA 7.0.

Results and discussion

The monitoring data set involved 26 sampling sites characterized by 12 parameters (water temperature **T**; *active reaction* or pH marked as **AR**; *dissolved oxygen* **DO**; *oxygen saturation* **OSat**; *conductivity* **COND**; *non-dissolved matter* **NDMat**; *ammoniac nitrogen* **NH_4-N**; *nitrate(V) nitrogen* **NO_3-N**; orthophosphates **P**; nitrate(III) nitrogen **NO_2-N**; *biological oxygen demand* **BOD**; *chemical oxygen demand* **COD**). The sampling period was between 2004 and 2009 but due to lack of data for some sampling sites and sampling periods the whole data matrix size finally was [555x12]. The data quality was proved by checking of the all aspects of the analytical procedures used - uncertainty of sampling and measurement, detection limit determination for each parameter, using of standard materials for method calibration. There were no missing data in the final data set.

The basic statistics of the data is presented in Table 1.

Table 1

Basic statistics of the data set (N = 555)

Parameter	Mean	Median	Minimum	Maximum	SD
T	14.28	14.5	1.0	28.3	6.46
AR	7.89	7.9	6.66	9.52	0.42
DO	7.14	7.07	0.7	14.64	2.41
OSat	70.85	72.0	8.0	131.0	21.35
COND	533.91	508.0	34.0	2700.0	286.8
NMat	20.95	16.0	2.0	190.0	18.8
NH_4-N	0.63	0.105	0.001	11.2	1.57
NO_3-N	2.74	1.34	0.01	113.0	8.65
P	3.31	0.238	0.006	549.0	28.09
NO_2-N	0.07	0.040	0.001	0.83	0.094
COD	23.24	23.0	0.073	125.0	17.21
BOD	3.53	2.73	0.0001	47.8	3.42

If cluster analysis (standardized data set, squared Euclidean distance as similarity measure, Ward's method of linkage) of the variables is performed using all data three major clusters are formed (Fig. 2).

Cluster 1 includes the indicators OSat, DO and AR and forms a pattern showing the impact of anthropogenic sources (eg industrial wastes) causing the oxidation properties of the water body. Cluster 2 contains another three parameters (P, NO_3 and NH_4) which probably get into one group of similarity due to their common origin, eg agricultural and farming activities along the river catchment being the reason for enrichment with phosphate and nitrogen - containing substances. The last identified cluster 3 unites the rest of the water quality parameters. It could be conditionally divided into two sub-clusters: (BOD, COD,

NDMat) and (NO_2, COND, T). This clustering resembles the role of the physical parameters on the water quality (temperature, conductivity) and, thus, the formation of possible seasonal patterns. Additionally, the biological impact of urban wastes (characterized by the correlation between BOD, COD and non-dissolved matter) contributes to the complete assessment of the river water quality and the creation of the respective water quality pattern.

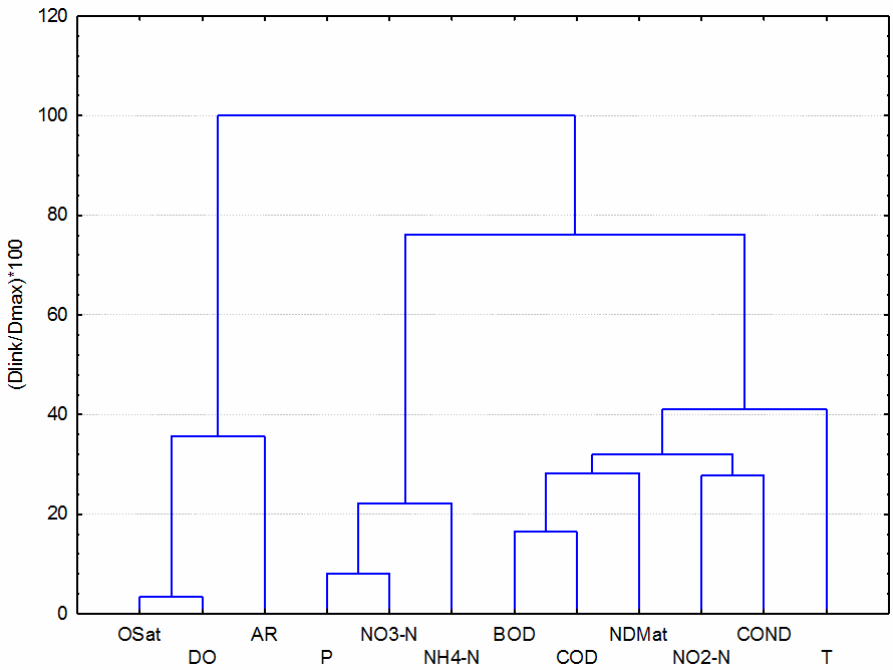

Fig. 2. Hierarchical dendrogram for linkage of 12 water parameters

In order to better explain the formation of different water quality pattern formation and, respectively, to identify the sources responsible for the data set structure in this case, principle components analysis was performed using the standardized data set and Varimax mode of presenting the results. In Table 2 the factor loadings are presented.

In principle, factor loadings higher than 0.7 are considered as statistically significant (they are marked by bold in the Table). Additionally, a second level of significance (bold + italics) is marked for better interpretation. Four latent factors explain over 65% of the total variance of the system and confirm the results obtained by cluster analysis. The first latent factor PC1 explains almost 20% of the total variance and indicates the strong impact of biological pollution parameters. It could be conditionally named "*urban wastes*" factor. The high factor loadings for BOD, COD, ND Mat, NO_2-N and COND correspond to the grouping of the water quality parameters in cluster 3 from the previous statistical analysis.

A second specific source in the river catchment is strongly related to those parameters which are linked to the concentration of the nutritional components in the water body - nitrogen containing species and phosphates. As conditional name "*agricultural*" factor

seems suitable since it resembles the role of this type of parameter correlation in describing the river catchment water quality. PC2 fits quite well to cluster 2.

Table 2

Factor loadings table

Variables	PC1	PC2	PC3	PC4
T	–0.10	–0.01	**0.70**	0.17
AR	0.026	–0.01	–0.10	**0.92**
DO	–0.21	–0.03	**–0.92**	0.24
OSat	–0.30	–0.05	**–0.77**	0.34
COND	***0.52***	0.02	0.37	0.26
NDMat	***0.62***	–0.06	–0.04	–0.10
NH_4-N	0.27	**0.75**	0.11	–0.03
NO_3-N	–0.06	**0.92**	–0.05	0.01
P	–0.11	**0.90**	–0.01	–0.01
NO_2-N	***0.46***	0.02	0.21	–0.29
COD	**0.80**	0.03	0.03	0.09
BOD	**0.78**	0.09	0.12	0.01
Expl. Var. [%]	19.6	18.5	17.9	10.1

A slight difference between the results of clustering and principal components analysis is in the explanation of the next two latent factors and cluster 1. The third principal component PC3 involves high factor loadings for dissolved oxygen (DO) and saturation with oxygen (OSat) along with the temperature parameter. However, T is reversely correlated to the other two indicators. Thus, the conditionally named "*industrial wastes*" factor (explanation of nearly 18% of the total variance) reveals a specific seasonal behavior of the river system - the oxidation ability of the water body is definitively different in the winter and in the summer season. Using only cluster analysis one cannot find out this property in assessing the water quality. The active reaction (AR) of the water body is separated in PC4 (explanation of 10.1% of the total variance) and does not correlated with any other water quality indicator. This specific latent factor is probably related to the natural water acidity and since no pH values are available in the data set we could determine it as "*acidity*" factor.

The linkage of the sampling locations (26, for each one various number of observation is registered; in total 552 separate cases are clustered) is performed using the same methods as in variables linkage (standardized data set, squared Euclidean distance as similarity measure and Ward's method of linkage). The hierarchical diagram is presented in Figure 3.

Two major clusters are formed but the detailed interpretation is difficult since the number of observation for each one of the monitoring sampling sites (in total 26) is very different. For some locations over 100 results are given (long-time complete monitoring) since other sites are included only by two to thirteen cases (limited monitoring results). Nevertheless, the formation of the two clusters is obvious and the careful checking indicated that the separation is due to geographical reasons - the right cluster consists dominantly of observations made from sites upstream and the left one includes dominantly downstream sites. No specific seasonal patterns were detected, probably due to the fact that the data set was not evenly populated for all sampling locations.

Next step in the statistical analysis was the clustering of the averages for the sampling sites. Thus, in total 26 objects were grouped. The results are shown in Figure 4.

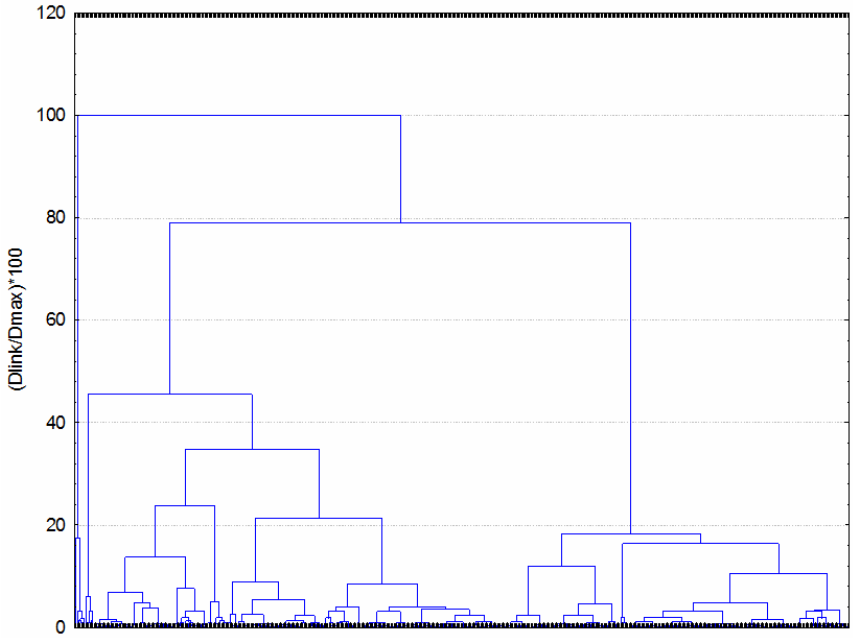

Fig. 3. Hierarchical dendrogram for linkage of 552 observations (cases)

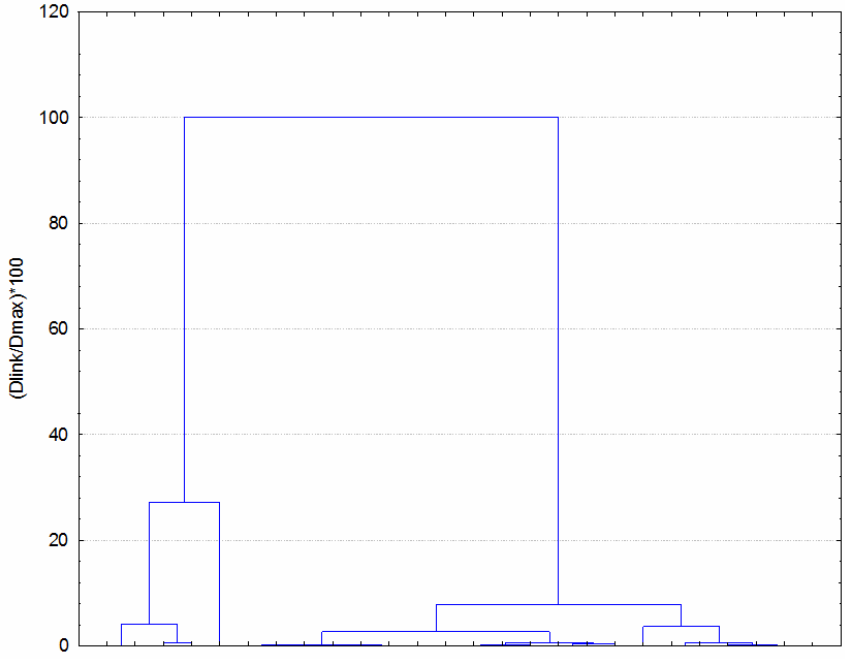

Fig. 4. Hierarchical dendrogram for linkage of 26 locations (cases)

In this situation the division between the sampling sites into two major clusters is quite clear and it is easier to interpret the cluster population. Cluster 1 (left side) includes sites with conditional numbers 2, 3, 4, 5, 6, all of them upstream sites and cluster 2 (right side of the dendrogram) - the rest of the sites (most of them located downstream). Again, the spatial separation "upstream - downstream" is proved. It is interesting to note that similar pattern of spatial separation is observed for other river systems [19, 20].

In Table 3 the factors responsible for the water quality around each sampling site are thoroughly discussed.

Table 3

Factors affecting the river water quality

Number of sampling sites	Conditional site number	Name and situation of sampling site	Municipality	Impact factors or sources of pollution
1.	**2**	River Tundja before town Kalofer	**Kalofer**	- lack of sewerage system - livestock production - illegal dumps - no WWTP
2.	**3**	River Tundja after town Kalofer	**Kalofer**	
3.	**4**	River Tundja in the end of damp Koprinka	**Kazanlak**	- livestock production - illegal dumps - agriculture
4.	**5**	River Tundja before Kazanlak after dam Koprinka	**Kazanlak**	- livestock production - illegal dumps - agriculture
5.	**24**	River Kranska - tributary after Kazanlak	**Kazanlak**	- textile industry - production and processing of metals - livestock production
6.	**6**	River Eninska - tributary before influx in river Tundja	**Kazanlak**	- livestock production - illegal dumps - agriculture
7.	**7**	River Tundja after influx of River Eninska	**Kazanlak**	- chemical industry - livestock production - illegal dumps - agriculture
8.	**8**	River Tundja, village Iagoda	**Maglizh**	- lack of sewerage system - chemical industry - livestock production - illegal dumps - agriculture - no WWTP
9.	**9**	River Tundja in the end of dam Jrebchevo	**Nikolaevo**	- lack of sewerage system - livestock production - illegal dumps - agriculture - no WWTP
10.	**10**	River Tundja - bridge village Bania	**Nova Zagora**	- illegal dumps - agriculture - dairy industry - meat processing industry - no WWTP
11.	**11**	River Belenska before influx in river Tundja - tributary	**Sliven**	- livestock production - illegal dumps - agriculture

Number of sampling sites	Conditional site number	Name and situation of sampling site	Municipality	Impact factors or sources of pollution
12.	**12**	River Asenovska before influx in river Tundja - tributary	**Sliven**	- livestock production - illegal dumps - agriculture
13.	**13**	River Tundja near village Samuilovo after River Asenovska	**Sliven**	- livestock production - illegal dumps - agriculture
14.	**23**	River Tundja near village Gavrailovo	**Sliven**	- textile industry
15.	**25**	River Sotiria near village Kamen - tributary	**Sliven**	- textile industry - livestock production - illegal dumps - agriculture
16.	**15**	River Mochuritza near village Vodenichane - tributary	**Straldja**	- livestock production - illegal dumps - agriculture - no WWTP
17.	**16**	River Mochuritza after town Karnobat - tributary	**Karnobat**	- lack of sewerage system - livestock production - illegal dumps - agriculture - no WWTP
18.	**17**	River Mochuritza before influx in river Tundja - tributary	**Yambol**	- livestock production - illegal dumps - agriculture
19.	**14**	River Tundja before influx of river Mochuritza	**Yambol**	- chemical industry - cocking oil production - dairy production - textile industry
20.	**1**	River Mochuritza near village Mokren - tributary	**Kotel**	- lack of sewerage system - livestock production - illegal dumps - no WWTP
21.	**18**	River Tundja near village Hanovo	**Tundja**	- lack of sewerage system - livestock production - meat processing industry - no WWTP
22.	**19**	River Tundja before town Elhovo	**Elhovo**	- lack of sewerage system - livestock production - illegal dumps - no WWTP
23.	**20**	River Tundja near the bridge of town Elhovo	**Elhovo**	- lack of sewerage system - livestock production - no WWTP
24.	**26**	River Dereorman - tributary	**Elhovo**	- food industry
25.	**21**	River Popovska before influx in river Tundja - tributary	**Boliarovo**	- lack of sewerage system - livestock production - meat processing industry - no WWTP
26.	**22**	River Tundja near the bridge of village Srem	**Topolovgrad**	- livestock production - agriculture - illegal dumps - no WWTP

Note: WWTP means waste water treatment plant

Cluster 1 (left) includes mainly sites from the municipalities Kalofer and Kazanlak where the anthropogenic impact is due to inlet of domestic wastes (Kalofer municipality), agricultural and farming activity, dump sites, chemical industry (both upstream municipalities). Two parameters could be used as tracers (specific indicators) for the water quality assessment in this part of the river catchment. These are oxygen indicators since enhanced concentrations of dissolved oxygen and oxygen saturation are observed for the sites included in cluster 1. Obviously, in the upper stream of the river the pollution level is still relatively low. Additionally, no significant inlets contribute to the total anthropogenic impact of the river catchment.

For the bigger part of sampling locations belonging to cluster 2 there are no specific tracers. All of them are characterized by higher concentrations of all other water quality parameters measured (except for DO and OSat). This is an important indication for the higher level of pollution downstream. The careful check of the information included in Table 3 proves this statement. The sampling locations with higher numbers than 6 reveal the anthropogenic impact due to the lack of purification facilities, industrial and domestic wastewater inlets, illegal dumping sites, higher industrial density (chemical enterprises, textile factories, diaries, pig farms etc.). Additional important indicator for the specific formation of cluster 2 is the high number of river inlets (Asenovska River, Kranska River, Mochuritsa River etc.) contributing significantly for the higher pollution level downstream of the major Tundja River.

The identification of the four latent factors responsible for the monitoring data set structure and explaining in detail the pollution patterns along the river catchment made it possible to apply an additional statistical analysis of monitoring results - apportionment model of the pollution. The main goal is to determine the contribution of each one of the identified sources (with conditional names "*urban wastes*" factor, "*agricultural*" factor, "*industrial wastes*" factor and "*acidity*" factor) to the formation of the total concentration of each one of the water quality parameters. This type of analysis known also as principal components regression was carried out according to the approach of Thurston and Spengler [18] described in the experimental part. The results are presented in Table 4.

Table 4

Apportionment model results in [%]

Variable	Intercept	Urban wastes	Agricultural	Industrial wastes	Acidity	R^2
AR	9.8	-	-	-	90. 2	0.85
DO	-	5.7	-	89.4	4.9	0.88
OSat	2.9	8.4	-	81.1	7.6	0.79
COND	21.3	55.5	-	12.3	10.9	0.81
NDMat	22.3	77.7	-	-	-	0.77
NH_4-N	10.6	29.8	60.6	-	-	0.89
NO_3-N	7.8	-	92.2	-	-	0.81
P	5.5	5.1	89.4	-	-	0.78
NO_2-N	29.9	42.3	-	12.9	14.9	0.74
COD	9.5	85.3	-	5.2	-	0.80
BOD	6.6	90.1	-	3.3	-	0.81

In Table 4 the contribution of each latent factor to the total species concentration is indicated. The first column of the table shows the unexplained by the model concentration

(intercept of the regression equation) and the last column shows the value of the multiple correlation coefficient r^2 being an indication for the model validity (comparison between the experimentally measured and calculated by the model concentration levels).

As seen in Table 4 the apportionment regression models show good adequateness and explain in a reliable way the contribution of each one of the identified factors to the formation of the total concentration of any of the water quality parameters involved. Thus, an overall description of the pollution events along the Tundja River stream for the period of monitoring is obtained. The dominant role of the urban wastes and agricultural activity is obvious in accordance with the comments of the "hot spots" in the river catchment.

Conclusions

In the presented study by the environmetric methods is carried out assessment of large environmental data of physicochemical parameters, characterized the river water quality. Created models of variables (parameters) describe the links between water quality indicators, identify the latent factors, which reveal possible sources of water pollution and determine the contribution of each one of identified pollution factors to concentration of each one water quality parameters. The results identify dominant role of the urban wastes and agricultural activities in water pollution.

CA of sampling sites shows tendency to spatial "upstream - downstream" separation regarding the level of pollution of the Tundja River. Similar pattern of spatial separation is observed for other river systems like Struma River, Saale River, Elba River.

Acknowledgements

The authors would like to express their sincere gratitude to the National Science Fund for the financial support (Project DO-02-352).

References

[1] Simeonova P, Simeonov V, Andreev G. Centr Europ J Chem. 2003;2:121-136.
[2] Simeonov V, Stefanov S, Tsakovski S. Mikrochim Acta. 2000;134:15-21.
[3] Simeonov V, Sarbu C, Massart D, Tsakovski S. Mikrochim Acta. 2001;137:243-248.
[4] Simeonov V, Stratis J, Samara C, Zachariadis G, Voutsa D, Anthemidis A, Sofoniou M, Kouimtzis T. Water Res. 2003;37:4119-4124. DOI:10.1016/S0043-1354(03)00398-1.
[5] Simeonov V, Simeonova P, Tsitouridou R. Ecol Chem Eng. 2004;11:450-469.
[6] Mihailov G, Simeonov V, Nikolov N, Mirinchev G. Water Sci Technol. 2005;51:37-43.
[7] Simeonova P, Lovchinov V, Simeonov V. J Balk Ecol. 2007;10:197-204.
[8] Astel A, Tsakovski S, Barbieri P, Simeonov V. Water Res. 2007;41:4566-4578. DOI:10.1016/j.waters.2007.06.030.
[9] Diadovski I, Atanassova M, Simeonov V. Ecol Chem Eng A. 2009;16:181-200.
[10] Diadovski I, Atanassova M, Simeonov V. J Water Res Protect. 2010;2:455-461. DOI: 10.4236/jwarp.2010.25052.
[11] Tsakovski S, Astel A, Simeonov V. J Chemomet. 2010;24:694-702. DOI: 10.1002/cem.1333.
[12] Diadovski I, Atanassova M, Simeonov V. Ecol Chem Eng A. 2010;17:199-215.
[13] Tsakovski S, Simeonov V, Stefanov S. Fresenius Envir Bull. 1999;8:28-36. DOI: 1018-4619/99/01-02/028-09.
[14] Spanos Th, Simeonov V, Stratis J, Xatzixristou X. Mikrochim Acta. 2003;141:35-40. DOI: 10.1007/ s00604-002-0921-9.
[15] Simeonov V, Simeonova P, Tsakovski S, Lovchinov V. J Water Res Protect. 2010;2:354-362. DOI: 10.4236/jwarp.2010.24041.

[16] Massart DL, Kaufman L. The interpretation of analytical chemical data by the use of cluster analysis. Amsterdam: Elsevier;1983.
[17] Einax J, Zwanziger H, Geiss S. Chemometrics in Environmental Analysis. Weinheim: VCH; 1998.
[18] Thurston G, Spengler J. Atmos Environ. 1985;19:9-26.
[19] Simeonov V, Einax JW, Stanimirova I, Kraft J. Anal Bioanal Chem. 2002;374:898-905. DOI: 10.1007/s00216-002-1559-5.
[20] Simeonova P, Lovchinov V, Dimitrov D, Radulov I. Ecol Chem Eng A. 2008;15:187-198.

DŁUGOOKRESOWA OCENA JAKOŚCI WODY RZEKI TUNDJI

Abstrakt: Dwie główne metody analizy danych środowiskowych (*analiza skupień* (CA) i *analiza składowych głównych* (PCA)) zastosowano do statystycznej oceny jakości wód transgranicznej rzeki Tundja. W badaniach wykorzystano dane otrzymane z monitoringu długookresowego. Próbki pobrano w 26 miejscach i scharakteryzowano za pomocą 12 parametrów fizykochemicznych. Pogrupowanie tych parametrów ze względu na 3 wskaźniki chemiczne pozwoliło na zbudowanie 3 głównych klastrów: pierwszy z nich pokazuje wpływ źródeł antropogennych, drugi - wpływ rolnictwa i działalności rolniczej, a trzeci opisuje rolę parametrów fizycznych i zanieczyszczeń środowiska miejskiego na jakość wody. W celu lepszej oceny danych monitoringowych zastosowano PCA, co pozwoliło na identyfikację czterech ukrytych czynników. Dwa z nich - czynnik „miejskie odpady" i czynnik „rolnictwo" - odpowiadają niemal w całości klastrom 3 i 2 z poprzedniej analizy statystycznej. Trzeci czynnik, nazwany „odpadami przemysłowymi", ukazuje specyficzne zmiany sezonowe w systemie rzecznym. Ostatni czynnik opisuje reakcję wody i jest określany jako czynnik „kwasowość". Powiązania pomiędzy miejscami pobierania próbek wzdłuż przepływu oceniono za pomocą CA. Wskazano istnienie dwóch klastrów z separacją przestrzenną „upstream-downstream". Model podziału zanieczyszczeń określał wkład każdego ze zidentyfikowanych czynników zanieczyszczeń w całkowitym stężeniu każdego z parametrów jakościowych wody.

Słowa kluczowe: monitoring, wody rzeczne, obróbka danych, analiza skupień, analiza składowych głównych