Bulg. J. Phys. **40** (2013) 325–330

Environmetric Data Interpretation to Assess Surface Water Quality[*]

P. Simeonova, P. Papazova, V. Lovchinov

Laboratory of Environmental Physics, "Georgi Nadjakov" Institute of Solid State Physics, Bulgarian Academy of Sciences, 72 Tzarigradsko Chaussee Blvd., 1784 Sofia, Bulgaria

Abstract. Two multivariate statistical methods (Cluster analysis /CA/ and Principal components analysis /PCA/) were applied for model assessment of the water quality of Maritsa River and Tundja River on Bulgarian territory. The study used long-term monitoring data from many sampling sites characterized by various surface water quality indicators. The application of CA to the indicators results in formation of clusters showing the impact of biological, anthropogenic and eutrophication sources. For further assessment of the monitoring data, PCA was implemented, which identified, again, latent factors confirming, in principle, the clustering output. Their identification coincide correctly to the location of real pollution sources along the rivers catchments.

The linkage of the sampling sites along the river flow by CA identified several special patterns separated by specific tracers levels. The apportionment models 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.

Thus, a better risk management of the surface water quality is achieved both on local and national level.

PACS codes: 87.64.-t, 87.23.-n

1 Introduction

The assessment of the river water quality is usually based on the comparison of analytically determined monitoring values of particular physicochemical parameters with the allowable threshold values defined in national or international legislation packages. A much more sound and reliable approach seems to be the application of environmetric strategies for intelligent data analysis of the river water monitoring results. Using them one considers the environmental system as a multivariate object and treats it respectively [1-10].

Usually, the studies performed a try to assess the river water quality or to optimize the monitoring procedure by multivariate statistics aim classification of the

[*] Talk given at the Second Bulgarian National Congress in Physics, Sofia, September 2013.

sampling locations trying to establish special relationships; detection of specific linkage between water quality parameters looking for identification of polluting sources; clarification of the whole data set structure. Therefore, the major goal of the present study is to reveal all aspects of the data set information (classification, modelling and interpretation) of monitoring data collection from Maritsa River and Tundja catchment in Bulgaria for a long-time period of observations using multivariate statistics in order to assess in a reliable way the river water quality.

2 Experimental

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 EEA laboratories are accredited under the BS EN ISO/IEC 17025.

The data set used for the aims of the present study is part of NESM and involved 21 sampling sites characterized by 8 parameters (Maritsa River) and 26 sampling sites with 12 parameters (Tundja River) – active reaction (pH), water temperature (T) [°C] , dissolved oxygen (O_2) [mg/l], oxygen saturation [%], conductivity [mS/cm], non-dissolved matter [mg/l], ammonia nitrogen (NH_4-N) [mg/l], nitrate nitrogen (NO_3-N) [mg/l], orthophosphates (PO_4) [mg/l], nitrite nitrogen (NO_2-N) [mg/l], biological oxygen demand (BOD) [mg/l], chemical oxygen demand (COD) [mg/l]. The analytical determination of the water indicators was performed according to the respective local and international standard methods.

Cluster analysis is a well-known and widely used classification approach for environmetrical purposes with its hierarchical and non-hierarchical algorithms[16]. The major task of CA is to determine groups of similar objects (e.g. sampling sites) or similar variables characterizing the objects (e.g. water quality indicators). The clusters formed are checked for significance and are subject to specific interpretation.

Principal components analysis (PCA) is a typical display method, which allows to estimate the internal relations in the data set and to model the ecosystem in consideration [11].

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

3 Results and Discussions

The first step in the envirtonmetric analysis of the Maritsa River monitoring data set (420x8) was chemical parameter classification by the use of hierarchical cluster analysis. In the clustering procedure z-transformation of the raw data was performed with squared Euclidean distance as similarity measure and Ward's method of linkage.

It is found that two major clusters are formed. One of them includes the parameters for oxygen content (SAT O_2 and DISS O_2) indicating the impact of a "biological factor" on the water quality. The other cluster involves the nutrition parameters (NH_4, PO_4, NO_3) along with the water turbidity (NDM) showing the influence of an "anthropogenic factor" (result from agricultural activity, waste water inlet, soil abrasion). The parameters BOD and COD are rather independent variables to the water quality revealing pollution processes of eutrophication.

In order to confirm this parameter classification and to get more information about the data set structure principal components analysis was applied to the same data set.

In Table 1 the factor loadings values are given.

Table 1. Factor loadings (statistically significant loadings are marked in bold)

Variable	PC 1	PC 2	PC 3
DISS O2	**0.966**	-0.037	0.137
SAT O2	**0.916**	0.195	0.156
NDM	0.098	**0.721**	-0.085
NH4	0.016	**0.493**	0.408
NO3	-0.124	**0.878**	0.207
PO4	0.062	**0.848**	0.046
COD	-0.029	0.023	**-0.926**
BOD	**-0.778**	0.065	0.261
Expl. Var. %	30.1	28.7	16.0

It is clear that three latent factors determine the data structure describing nearly 75% of the total variance of the system. The first latent factor PC1 could be conditionally named "*biological factor*" since it reflects the impact of aerobic biological processes on the water quality. The second latent factor PC2 could be conditionally named "*anthropogenic factor*". It indicates the role of various anthropogenic pollution sources (agriculture, waste water pollutants, soil abrasion particles, etc.) to surface water quality. The last latent factor PC3 reveals the specific role of COD and could be conditionally named "*eutrophication factor*".

It was important to analyse the special relationships (linkage between sampling locations) along the flow of Maritsa River. Four major clusters are formed:

$$K1(21, 19, 18, 20, 17) \qquad K2(15, 14, 8, 12, 7)$$
$$K3(13, 6, 3, 5, 4, 10, 11, 2) \qquad K4(16, 9, 1).$$

It might be concluded that the first cluster K 1 is formed entirely from sites located in the lower stream of Maritsa River characterized by significant municipal and industrial activity. The other three clusters are characterized by mixed features (with respect to their special location) and include sampling sites both from the upper and lower stream of the river.

K1 is characterized by the highest levels of dissolved oxygen, saturation with oxygen, ammonia nitrogen, non-dissolved matter, nitrate nitrogen and phosphate. It forms the pattern of sampling sites with the most significant pollution by biological and anthropogenic sources. The second pattern formed (K2) has as specific tracers the highest levels of BOD and the lowest levels of NH_4, NO_3 and PO_4. As expected the agricultural activity is lower and, thus, the indicators for agricultural pollution are non-significant. Although the vicinity is known for major industrial activity, no indication for industrial pollution is found due to the function of waste water treatment plant of City of Plovdiv. Still, eutrofication sources are observed (highest levels of BOD), which determine the specificity of this pattern of sampling sites.

An intermediate position is detected for the biggest group of sites (K3) where the averages for all water quality parameters lie between those of the other groups. The sampling sites are dominantly from the upper stream of the river and some tributaries to the major river flow. The pollution level is not high. Finally, the fourth small cluster K4 offers a pattern with high level of eutrophication parallel to agricultural activity.

If cluster analysis is applied to Tundja River monitoring data (standardized data set, squared Euclidean distance as similarity measure, Ward's method of linkage of the variables) three major clusters are formed (Figure 1).

Cluster 1 includes the indicators OSat, DO and AR and forms a pattern showing the impact of anthropogenic sources (e.g. industrial wastes) causing the oxidation properties of the water body. Cluster 2 contains another three parameters (P, NO3 and NH4) which probably get into one group of similarity due to their common origin, e.g. agricultural and farming activities. The last identified cluster 3 unites the rest of the water quality parameters. 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.

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 indicates the strong impact of biological pollution parameters. It could be conditionally named *"urban wastes"* factor.

A second specific source PC2 in the river catchment is strongly related to those parameters which are linked to the concentration of the nutritional components

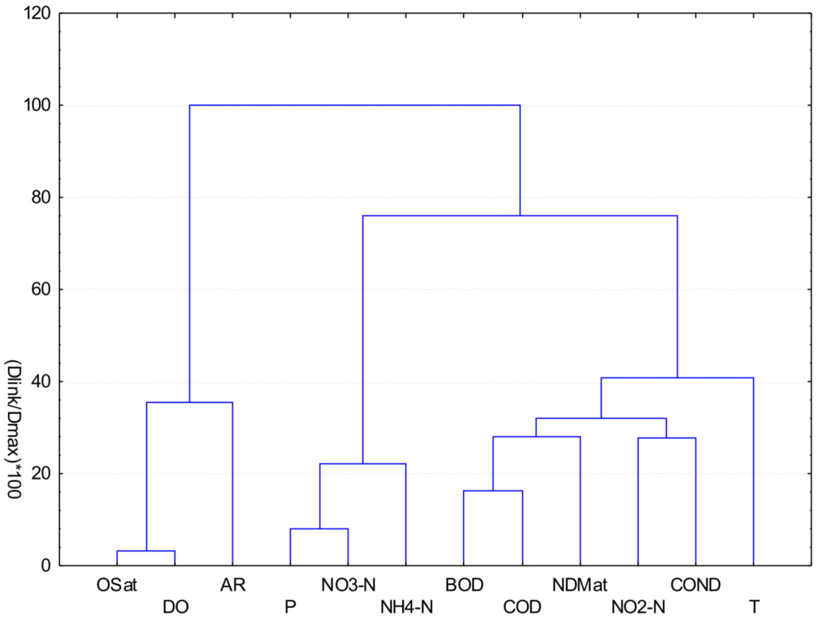

Figure 1. Hierarchical dendrogram for linkage of 12 water parameters.

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
NH4-N	0.27	**0.75**	0.11	-0.03
NO3-N	-0.06	**0.92**	-0.05	0.01
P	-0.11	**0.90**	-0.01	-0.01
NO2-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

in the water body – nitrogen containing species and phosphates withconditional name “*agricultural*” factor. 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 reveals a specific seasonal behavior of the river system. The active reaction (AR) of the water body is separated in PC4 (explanation of 10.1% of the total variance) and does not correlate with any other water quality indicator. This specific latent factor is probably related to the natural water acidity as *"acidity"* factor.

The linkage of the sampling locations (26) leads to the formation of two major clusters. One of the clusters includes sites with conditional numbers 2, 3, 4, 5, 6, all of them upstream sites and the other one– the rest of the sites (most of them located downstream). Again, the spatial separation "upstream – downstream" is proved.

4 Conclusion

In the presented environmetric study assessment of monitoring data of parameters, characterizing the river water quality is performed.

The multivariate statistical models obtained: describe the links between water quality indicators; identify the latent factors, which reveal possible sources of water pollution; reveal spatial patterns of sampling locations contributing in this way a better understanding of the role of "hot spots" and traditional urban and agricultural pollutants along the river stream. The results identify the dominant role of the industrial wastes and agricultural activities in water pollution.

References

[1] P. Simeonova, V. Simeonov, G. Andreev (2003) *Centr. Eur. J. Chem.* **2** 121.

[2] V. Simeonov, S. Stefanov, S. Tsakovski (2000) *Mikrochim. Acta* **134** 15.

[3] V. Simeonov, C. Sarbu, D. Massart, S. Tsakovski (2001) *Mikrochim. Acta* **137** 243.

[4] V. Simeonov, P. Simeonova, R. Tsitouridou (2004) *Ecol. Chem. Eng.* **11** 450.

[5] P. Simeonova, V. Lovchinov, V. Simeonov (2007) *J. Balkan Ecol.* **10** 197.

[6] A. Astel, S. Tsakovski, P. Barbieri, V. Simeonov (2007) *Water Res.* **41** 4566.

[7] S. Tsakovski, A. Astel, V. Simeonov (2010) *J. Chemomet.* **24** 694.

[8] S. Tsakovski, V. Simeonov, S. Stefanov (1999) *Fres. Environ. Bull.* **8** 28.

[9] T. Spanos, V. Simeonov, J. Stratis, X. Xatzixristou (2003) *Mikrochim. Acta* **141** 35.

[10] V. Simeonov, P. Simeonova, S. Tsakovski, V. Lovchinov (2010) *J. Wat. Res. Prot.* **2** 354.

[11] J. Einax, H. Zwanziger, S. Geiss (1998) *Chemometrics in environmental analysis*. VCH, Weinheim.