African Journal of Microbiology Research Vol. 6(10), pp. 2346-2356, 16 March, 2012
Available online at http://www.academicjournals.org/AJMR
DOI: 10.5897/AJMR11.1346
ISSN 1996-0808 ©2012 Academic Journals

*Full Length Research Paper*

# Total coliforms and data mining as a tool in water quality monitoring

**Ivana D. Radojević[1]*, Dušan M. Stefanović[2], Ljiljana R. Čomić[1], Aleksandar M. Ostojić[1], Marina D. Topuzović[1] and Nenad D. Stefanović[2]**

[1]Institute of Biology and Ecology, Faculty of Science, University of Kragujevac, 34000 Kragujevac, Radoja Domanovića 12, Serbia.
[2]Institute of Mathematics and Informatics, Faculty of Science, University of Kragujevac, 34000 Kragujevac, Radoja Domanovića 12, Serbia.

Accepted 10 February, 2012

**Total coliforms, as a microbiological indicator of water quality, have been tested on the basis of condition, dynamics, as well as on the dependence on other physicochemical and biological parameters, by methods and models of data mining. Using a combination of intelligent approaches, cluster analysis and classification, total coliforms have been analyzed and modeled on the examples of the Gruža and the Grošnica reservoirs. These reservoirs have different morphometric characteristics, different trophic status as well as dominant bacterial communities. The study is based on the existing information system and automated data analyses for the period of 10 years. The system determines the accuracy of analyses by validity percentage. The analyses show that the number of total coliforms is connected to anthropogenic activity, the amount of organic mater, as well as to the presence of bacterial community which is not dominant or characteristic for the specific reservoir.**

**Key words:** Total coliforms, reservoir, water quality, data mining, cluster analysis, classification.

## INTRODUCTION

Water quality is determined according to its physical, chemical and biological parameters (Sargaonkar and Deshpande, 2003). The main problem is the complexity of the analysis of great number of variables as well as their variability due to natural and human influences (Saffran, 2001; Simeonov et al., 2002). Classification, modeling and interpretation of great number of data are an important segment of water quality monitoring (Boyacioglu and Boyacioglu, 2007).

In recent years, various tools and techniques of data mining have become an important part of monitoring of water quality status; they also provide prediction of changes, and are significant in the processes of sustainable monitoring of water resources (Kumar et al., 2006). Data mining also known as “knowledge-discovery in databases” implies automatic or semiautomatic research and analysis of great amount of data in order to discover patterns and relations hidden among the data (Han et al., 2010).

In water quality assessment the microbial community has special significance, especially in terms of protecting public health. Coliform bacteria, normally present in intestinal tract of humans and worm-blooded animals, can secondary be found on plants, in the soil and in waters. Although primarily non-pathogenic, their presence refers to the presence of disease-causing organisms. They reach natural waters mainly during rainfall, through runoff from agricultural and urban lands as well as through drainage (Medema et al., 2003). Total (TC) and fecal coliforms (FC) as indicators of previous and new fecal pollution, are often used as indicators of microbial water quality (Rompré et al., 2002). TC is used

*Corresponding author. E-mail: ivana@kg.ac.rs. Tel: +381 34 336 223. Fax: +381 34 335 040.

**Abbreviations: TC,** total coliforms; **H,** heterotrophs; **Hm,** heterotrophs (mesophile); **FO,** facultative oligotrophs; **$BOD_5$,** 5-day biochemical oxygen demand; **Mn**, manganese; **COD**, chemical oxygen demand; **TP,** total phosphate; **EC,** conductivity; **Fe,** iron; **$NH_4^+$,** ammonia; **Chl-a,** chlorophyll a; **$Cl^-$,** chloride; **TSS,** total suspended solids; **MPN,** Most Probable Number; **cfu,** colony forming units; **CA,** cluster analysis.

as a parameter giving basic information on microbiological quality of surface waters (WHO, 2008).

Different factors influence the number and dynamics of coliform bacteria in natural surface waters. Physicochemical and biological properties of water, such as pH, dissolved oxygen, temperature, phosphates, $BOD_5$, SS, organic and inorganic nutrients, humic substances, predacious microorganisms such as protozoa, also have an important role (McCambridge and McMeekin, 1984; Curtis et al., 1992; Bagde and Rangari, 1999; Youn-Joo et al., 2002; Juhna et al., 2007; Syed Ahmad et al., 2009; Hong et al., 2010). Environmental factors also have great influence: atmospheric conditions (precipitation and solar radiation), surface runoff, human activities causing contamination such as different use of land – agricultural, urban, industrial (Gameson and Saxon, 1967; McCambridge and McMeekin, 1984; Fisher and Endale 1999; Kistemann et al., 2002; Tong and Chen, 2002; George et al., 2004; Mehaffey et al., 2005; Byamukama et al., 2005; Zhang and Lulla, 2006; Derlet et al., 2008).

The Gruža and the Grošnica reservoirs are important sources of water supply for Kragujevac city and its surroundings. In previous period, these reservoirs were subjects to various hydro-biological researches (Ćurčić and Čomić, 2002; Ostojić et al., 2005, 2007), concluding that Gruža is eutrophic reservoir in which the dominant community is the heterotrophic bacteria while Grošnica is oligo-mezotrophic reservoir in which the facultative oligotrophic is the dominant bacteria group. The number of total coliforms was observed in standard microbiological researches but they have never been subject to any further researches.

The analysis, modeling and prediction of the number of total coliforms were performed by various statistical and other tools, among which data mining tools were less frequent (Canale et al., 1973; Mahloch, 1974; Bergstein et al., 2001; Brion et al., 2002; Idakwo and Abu, 2004; Derlet et al., 2008; Iscen et al., 2008; Syed Ahmad et al., 2009). Due to the importance of coliform bacteria in determining quality of natural waters, the aim of this paper is to automatically, with chosen methods and models of data mining, determine the dependence degree and the size of influence of physicochemical and biological parameters on abundance and dynamics of total coliform bacteria, based on the data implied in information system for the two reservoirs with different morphometric characteristics, trophic status and dominant bacterial community.

## MATERIALS AND METHODS

### Study area and water quality data

The city of Kragujevac (in the central part of Serbia) is supplied with water from the Gruža and the Grošnica reservoirs (Figure 1). Characteristics and the values of trophic state parameters of both reservoirs are given in previous paper (Ostojić et al., 2007). The data set used in this study was generated through monitoring of the water quality of the Gruža and the Grošnica reservoir. The data set includes the data of the laboratory for water quality inspection of the public utility company for water supply and sewerage in Kragujevac. Monthly sampling was carried out during the period of ten years (1998 to 2008). Three permanent sampling sites were selected for qualitative and quantitative sampling for Grošnica reservoir and five sampling sites for Gruža reservoir (Figure 1). Samples were taken at every 5 m of depth. Analyses were performed by using standard methods (APHA, 1998). Physicochemical, microbiological and other parameters used for modeling are same for both reservoirs. They were taken from the information system Serbian lakes and reservoirs (SeLaR), and are described in detail in the paper Stefanović et al. (2012).

Data set for Gruža reservoir includes 1608 samplings, out of which 640 values for TC are missing, therefore 968 data have been used for the analysis. For Grošnica reservoir the data set implies 382 sampling, out of which 172 include TC values and they have been used for the analysis.

### Data analysis, methods and models

In our case, data source is relational database of the SeLaR information system. This database stores a wide spectrum of different data such as characteristics of lakes and reservoirs, their geographic positions, characteristics of the surroundings etc. It also stores values of physical, chemical and microbiological parameters measured over the years at different locations and depths of lakes and reservoirs. This way, it integrates all the data required in the data mining process. Data entry process ensures data quality through different validation mechanisms, data model constrains and relationships. Processes of extracting, transformation and loading are realized through the special interface called unified dimensional model (UDM) (Mundy et al., 2011).

Construction of the UDM as an additional layer over the data sources offers more clearly data model, isolation from the heterogeneous data platforms and formats, and improved performance for aggregated queries and data mining processing. UDM also allows business rules to be embedded in the model, as well as option to define actions in relation to query results (that is drill-down reporting). Another advantage of this approach is that UDM does not require data warehouse or data mart. It is possible to construct UDM directly on top of relational database systems, and to combine relational databases and data warehouse systems within a single UDM. UDM allow creation of one data source view (DSV) for use by the system. The DSV is an abstraction layer that is used to extend the objects (relational tables and views) that are exposed by the data source to a collection of objects from which OLAP (On-line analytical processing) server objects are created. Within the data source view we included all of the relational views that were used to create OLAP cubes and data mining models.

Data in the relational database are stored in normalized tables optimized for transactional processing. UDM interface performs several activities: data selection, calculation of certain aggregated values, and transforms data so it can be used in the data mining process. These data transformation forms have both temporal and spatial dimensions.

As a software development environment we used Microsoft SQL server 2008 R2 package: relational database engine for SeLaR data, integration services for data transformation and loading, and analysis services for OLAP and data mining modeling. We used clustering and decision trees algorithms for building the data mining models.

### Clustering

Clustering is the process of grouping a set of data objects into

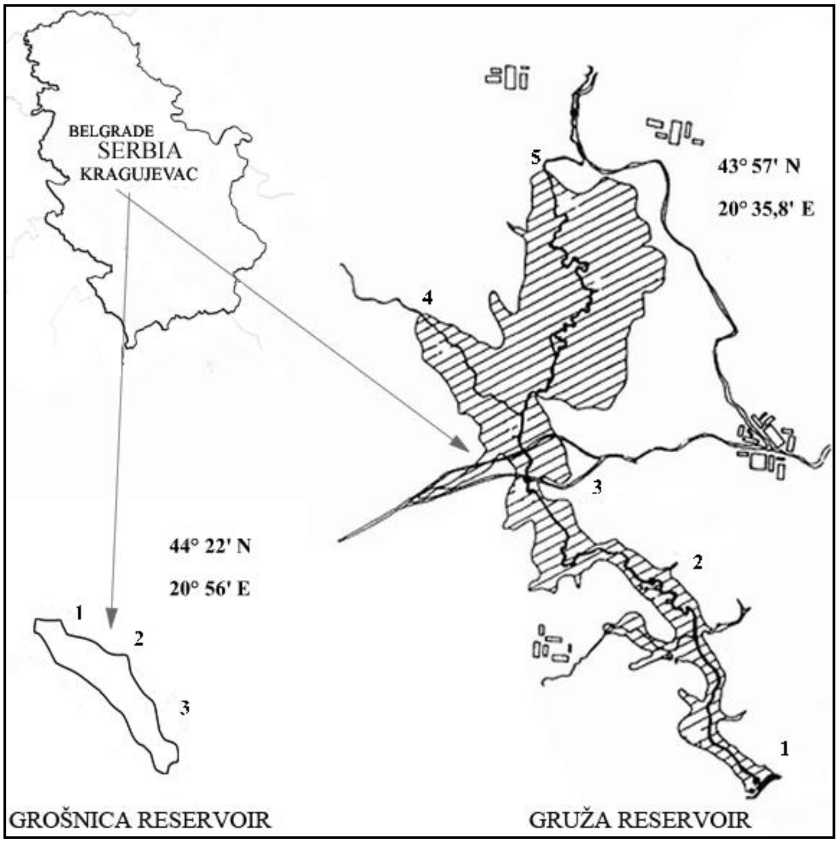

**Figure 1.** The Gruža and the Grošnica reservoir sampling points (1– Dam, 2 – Center, 3 – Bridge, 4 – The mouth of the Borač River, 5 – The mouth of the Gruža River).

multiple groups or clusters so that objects within a cluster have high similarity, but are very dissimilar to objects in other clusters. Similarity or dissimilarity between objects in a single cluster is determined by measurements (or proximities). Calculation of similarity/dissimilarity between attributes is dependent on the attribute type (normal, binary, numeric and ordinal) and can be based on distance between attribute values or probabilities of attribute values. Clusters can have stronger or weaker relationships. Values of particular objects (rows or n-tuples) of some variable (column or attribute) can be common for more than one cluster. The number of this common object values determine the degree of relationship (similarity) between clusters. Basic clustering techniques are organized into the following categories: partitioning methods, hierarchical methods, density-based methods and grid-based methods. Partitioning organizes the objects of a set into several exclusive groups or clusters. First, from *n* set of objects, *k* partitions are formed, and then iterative relocation technique is applied which improves the model by moving objects between partitions. Partitioning methods include k-means, k-medoids, and CLARANS (Jiawei Han et al., 2010).

Clustering algorithm within SQL server analysis offers two clustering methods: Expectation Maximization (EM) and K-means. EM cluster assignment method uses a probabilistic measure to determine which objects belong to which cluster. EM method considers a bell curve for each dimension with a mean and standard deviation. As a point falls within the bell curve, it is assigned to a cluster with a certain probability. Because the curves for various clusters can (and do) overlap, any point can belong to multiple clusters, with an assigned probability for each. This technique is considered soft clustering because it allows clusters to overlap with indistinct edges. The K-means (it belongs to group of partitioning methods) method assigns cluster membership by distance an object belongs to the cluster whose centre it is closest to (which is measured using a simple Euclidean distance). When all objects have been assigned to clusters, the centre of the cluster is moved to the mean of all assigned objects, thus the name K-means -K being the typical denomination for the number of clusters to look for. This technique is considered hard clustering because each object is assigned one and exactly one cluster (MacLennan et al., 2009). The analysis services clustering algorithm provides a scalable framework. The principle of the scalable framework is that particular data points that are not likely to change clusters can be compressed out of the data you are iterating over, providing room to load more data.

Clustering modeling is the process in which selection of variables and determinations of input parameters are performed. Selection of variables depends on the research goal. Parameters used in our cluster model are as follows:

i) The CLUSTERING METHOD parameter indicates which methods (algorithm) are used to determine cluster membership. This parameter can have the following values:

a) Scalable EM (default);
b) Vanilla (non-scalable) EM;
c) Scalable K-means;
d) Vanilla (non-scalable) K-means.
ii) CLUSTER_COUNT tells the algorithm how many clusters to find.

iii) MINIMUM CLUSTER CASES controls when a cluster is considered empty and is discarded and reinitialized.
iv) MODELLING_CARDINALITY controls how many candidate models are generated during clustering.
v) STOPPING_TOLERANCE is used by the algorithm to determine when a model has converged.
vi) SAMPLE_SIZE indicates the number of cases used in each step of the scalable framework.
vii) CLUSTER_SEED is the random number seed used to initialize the clusters.
viii) MAXIMUM_INPUT_ATTRIBUTES controls how many attributes can be considered for clustering before automatic feature selection is invoked.
ix) MAXIMUM STATES controls how many states one particular attribute can have.

Based on our experimenting and general best practice, we selected the most suitable values for these parameters. Clustering results (knowledge) are presented in different view which allows further analysis and decision making. These are: Cluster profiles view, Cluster diagram view, Cluster characteristics view and Cluster discrimination view. The first two views correspond to all clusters, and the last two views refer to particular clusters. The Cluster profiles view displays a column for each cluster in the model and a row for each attribute. This setup makes it easy to see interesting differences across the cluster space. Using this view, it is possible to choose an attribute of interest and visibly scan horizontally to see its distribution across all clusters. When we notice some interesting items, we can look at neighboring cells or other cells of the same cluster to learn more about what that cluster means. Clicking any cell in the grid provides details on the information contained in the mining legend. Cluster diagram gives us a visual representation of all clusters, where clusters with greater number of objects are shaded with darker color. Also, the ticker lines between clusters represent stronger relationships. In Cluster characteristics view, attributes are displayed by probabilities and sorted in descending order. Attributes with highest probabilities determine characteristics of the cluster and its name. Cluster discrimination view gives comparison of a single cluster with complement of whole population, as well as comparison between any two clusters (MacLennan et al., 2009). Probabilities obtained from clustering process are calculated as:

$p = r_c/r_p$,

where $r_c$ is number of rows (objects) in the observed cluster, and $r_p$ is number of rows in the whole population.

### Classification with decision trees

Classification is a process of finding the set of models or functions that describe and differentiate classes of data or concepts. These models are used for prediction of object class whose class label is unknown. Classification involves two main steps. In the first step, model based on the known data is designed. If the model is acceptable, in the second step, model used for classification of new data is developed.

There are several classification methods which use different data mining algorithms: decision trees, logistic regression, naïve bayes and neural networks. The principal idea of a decision tree is to split data recursively into subsets. Each input attribute is evaluated to determine how cleanly it divides the data across the classes (or states) of your target variable (predictable attribute). The process of evaluating all inputs is then repeated on each subset. When this recursive process is completed, a decision tree is formed. Decision tree induction is a top-down recursive tree induction algorithm, which uses an attribute selection measure to select the attribute tested for each nonleaf node in the tree. ID3, C4.5, and CART are examples of such algorithms using different attribute selection measures. Tree pruning algorithms attempt to improve accuracy by removing tree branches reflecting noise in the data (Jiawei Han et al., 2010).

Besides the classification, the decision tree method can be used for regression and estimation. Decision tree used in SQL server analysis services has several input parameters. These parameters are used to control the tree growth, tree shape, and the input/output attribute settings. Classification modeling consists of determining the target variable, influence variables, as well as input parameters. A classification model extracts patterns that predict the individual values of one column based on the values in other columns. Classification process is characterized by Gorunescu (2011): input - a training dataset containing objects with attributes, of which one is the class label; output - a model (classifier) that assigns a specific label for each object (classifies the object in one category), based on the other attributes; the classifier is used to predict the class of new, unknown objects. A testing dataset is also used to determine the accuracy of the model.

Interpretation of results is done through the decision tree viewer and dependency network. The tree is laid out horizontally with the root node on the left side. Each subsequent node in the tree relates to certain condition for input variable. Each node contains a histogram bar with different colors, representing various classes. For each class, at given conditions, occurrence probability of its values is shown.

The dependency network displays the relationships among attributes derived from decision tree model’s content. Each node represents one attribute, and each edge represents the relationship between two nodes. An edge has a direction, pointing from the input attribute (node) to the predictable attribute (node). An edge can be bidirectional, which means two nodes can predict each other. Probabilities in classification representation are calculated as

$p = n_c/n$,

where $n_c$ is number of class instances of target variable, and $n$ is total number of instances of target variable in the observed tree node.

## RESULTS AND DISCUSSION

### Cluster analysis (CA)

Ten clusters have been initialized based on the data for the Gruža reservoir, and seven clusters for Grošnica reservoir. Figures 2 and 3 show the cluster diagrams for both reservoirs which show a graphical representation of the data associations found. The significant clusters or nodes of data are shown as shaded rectangles. Dark lines show the strong intercluster relations. For each cluster the range of parameters is given, as well as the probabilities with which they participate in the cluster. Based on the analysis of cluster characteristics, it is possible to see which variables are dominant in certain clusters and with which probabilities. Since the emphases in this paper lies on the analysis of TC, for the representation we chose the clusters in which TC had the most significant influences. As best way of presenting different influences that TC and other parameters have in a certain cluster in regard to other clusters, we chose the Cluster discrimination.

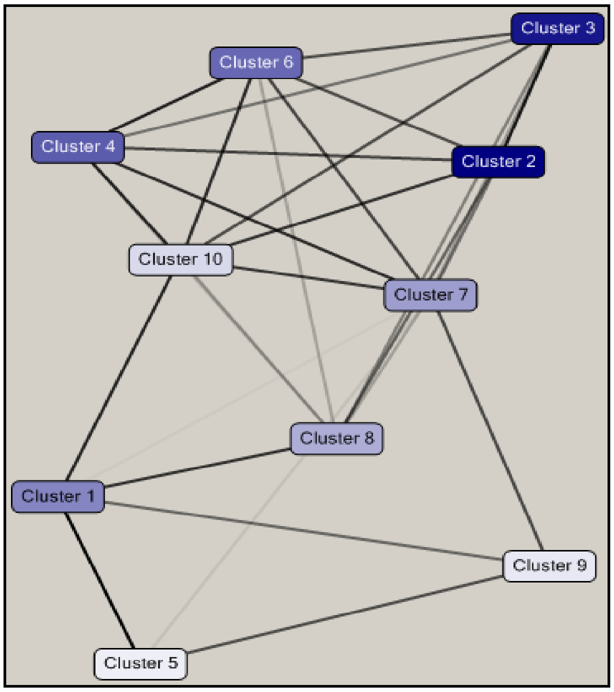

**Figure 2.** Cluster Diagram for TC - The Gruža reservoir.

For the Gruža reservoir (Figure 2) the most important clusters are 2 and 3, and right after 4 and 6. Most objects are focused in these clusters. For representation we chose cluster 2 in which TC are most influential with smaller number, and cluster 6 in which they have significant role with greater values. Difference analysis between cluster 2 and 6 with other clusters is best seen from Cluster discrimination (Tables 1 and 2).

By modeling it has been determined that TC is most influential in cluster 2, with values < 120 MPN/100 ml and probability of 100%. Through Cluster discrimination browser it is noticeable that the likeliness that they could be found in greater number in other clusters is 26%. This indicates the greater data range related to TC in other clusters, and also, in other clusters they do not group around a certain number as in cluster 2. In cluster 2 relatively significant influence have H with values < 1540 cfu/ml and probability of 62%. Relatively significant influences also have Hm < 320 cfu/ml with 56% probability, as well as FO < 1520 cfu/ml with probability of 49%. Relatively significant influence, 46% probability in the cluster 2, has Chl-a with values < 25 µg/l, and EC ranging 300 to 330 µS/cm with 21% probability. With probability of 15%, in cluster 2, there is also Fe < 0.1 mg/l, then 10%, M alkalinity 26 to 29 ml/l and $BOD_5$ < 6.3 mg/l with 5% probability. In which ranges and with whatprobabilities these as well as some less influential parameters determine other clusters in regard to cluster 2, is shown in Table 1.

Contrary to cluster 2, cluster 6 favors higher values for TC, in regard to other clusters (0 to 2340 MPN/100 ml) and with values of 51%. In it, the greatest significance has the depth of 15 m with 100% probability, as well as Mn 0 to 0.1 mg/l with 51%. Significant influence, as in cluster 2, has Chl-a with 25% probability, but with lower values 0.0 to 7.7 µg/l. The depths smaller than 15m, unlike cluster 6, are present in other clusters with 43%, and the values of TC >2340 MPN/100 ml, with 11%. The range of appearance, probabilities and belonging to cluster 6 or the other clusters, of less influential parameters, are shown in Table 2. If we compare the characteristics of clusters 2 and 6, the clear difference regarding to locations is noticeable. While cluster 2 favors locations 1 and 2 (near the dam) with 60%, and locations 3 (bridge), 33%, which is the deepest part of the reservoir, cluster 6 favors locations 4 and 5 (mouth of tributaries) with 66% and 3 (bridge) whit 13%, where is the shallowest part of the reservoir (Table 3).

For the Grošnica reservoir (Figure 3) the most important are clusters 1 and 3. Most data are focused in these clusters. Clusters 2, 4 and 5 come right after. The cluster analysis shows that TC has the greatest influence in clusters 3 and 5. In cluster 3, in regard to other clusters, the values TC < 135 MPN/100 ml come together with the probability of 100%. Contrary to that, the probability that they would be found in greater number in other clusters is 28%. In cluster 3, relative influence has $NH_4^+$ with values < 0.1 mg/l (contrary to other clusters with values 0.1 to 2 mg/l, 8%) and M alkalinity in the range of 29 to 37 ml/l, with 48% of appearance probability (contrary to others 0.8 to 29 ml/l, 3%). Sampling in March and August in this cluster is present with 13 and 10%, and values for EC 313 to 467 µS/cm, 8%, Fe with 0.02 mg/l is 6% same as the values for $BOD_5$ < 3.6 mg/l. Fe, $Cl^-$, TP, months April and May, also belong to this cluster. Their probabilities, ranges of appearance and belonging to cluster 3 or other clusters are shown in Table 4.

Contrary to cluster 3, cluster 5, in regard to other clusters, favors the values TC < 40 MPN/100 ml with the probability of 100%. The values > 40 MPN/100 ml in other clusters are with 25% probability. M alkalinity in cluster 5 is 32 to 37 ml/l (38%) and FO (9%) from 171 to 385 cfu/ml. November with 11% probability and July with 4% also belong to cluster 5. Mn in range 0.0 to 0.8 mg/l and TP 0.02 µg/l belong to the same cluster and with same probabilities. The ranges and probabilities of the same parameters but in other clusters are shown in Table 5.

The analysis of clusters in which TC are the most dominant, indicates that in the Gruža reservoir localities and depths at which certain number of TC appears can be grouped, as well as that there is the relation between the number of other microbiological communities (H, Hm, FO) and the number of TC. There is regularity that greater abundance of any community (by the analysis of certain number for each) indicates greater number of TC. Also noticeable is the dependence of certain numbers of

**Table 1.** Cluster discrimination - cluster 2 and other clusters - The Gruža reservoir.

| Variables (units)* | Values | Favors cluster 2 | Favors complement of cluster 2 |
|---|---|---|---|
| TC (MPN/100 ml) | < 120 | 100 | |
| H (cfu/ml) | < 1540 | 62 | |
| Hm (cfu/ml) | < 320 | 56 | |
| FO (cfu/ml) | < 1520 | 49 | |
| Chl a (µg/l) | < 25 | 46 | |
| TC (MPN/100 ml) | > 120 | | 26 |
| EC (µS/cm) | 300 to 330 | 21 | |
| H (cfu/ml) | > 1540 | | 20 |
| Hm (cfu/ml) | > 320 | | 19 |
| FO (cfu/ml) | > 1520 | | 16 |
| Fe (mg/l) | < 0.1 | 15 | |
| M alkalinity (ml/l) | 26 to 29 | 10 | |
| Chl a (µg/l) | > 25 | | 7 |
| $BOD_5$ (mg/l) | < 6.3 | 5 | |
| EC (µS/cm) | 330 to 815 | | 3 |

*see abbreviations.

**Table 2.** Cluster discrimination - cluster 6 and other clusters - The Gruža reservoir.

| Variables (units)* | Values | Favors cluster 6 | Favors complement of cluster 6 |
|---|---|---|---|
| Depth (m) | 15 | 100 | |
| TC (MPN/100 ml) | 0 to 2340 | 51 | |
| Mn (mg/l) | 0.0 to 0.1 | 51 | |
| Depth (m) | 2 to 14 | | 43 |
| Chl-a (µg/l) | 0.0 to 7.7 | 25 | |
| TC (MPN/100 ml) | > 2340 | | 11 |
| Chl-a (µg/l) | > 7.7 | | 7 |
| Mn (mg/l) | 0.1 to 4.6 | | 5 |
| Hm (cfu/ml) | < 7795 | | 2 |
| Hm (cfu/ml) | > 7796 | 2 | |

*see abbreviations.

**Table 3.** Cluster characteristics - cluster 2 and cluster 6 - The Gruža reservoir.

| Cluster 2 | | Cluster 6 | |
|---|---|---|---|
| Location | Probability (%) | Location | Probability (%) |
| 1 to 2 | 63 | 4 to 5 | 66 |
| 3 | 33 | 3 | 13 |
| 4 to 5 | 1 | 1 to 2 | 1 |

TC on certain values of Chl-a, EC, Mn, Fe, M alkalinity and $BOD_5$.

The analyzed clusters in the Grošnica reservoir indicate that there are no differences regarding to spatial distribution in the number of TC, but there is temporal dependence. The Grošnica reservoir is smaller in size and volume than the Gruža reservoir, and its environmental conditions are less heterogeneous. Temporal dependence is represented through months of sampling. The values for May and April, months with lots of rain, indicate that the number of TC would be > 135 MPN/100 ml. In March, when temperatures are low and with no precipitation, as well as in August, usually with drought and low water levels, it is greater probability that the number of TC would be < 135 MPN/100 ml. Similar situation is in July and November when the probability is that the number of TC would be < 40 MPN/100 ml. M alkalinity is most significantly in relation with TC, while TC in this reservoir, instead of relation with Chl-a, shows greater dependence on $NH_4^+$. Among microbiological communities the relation with FO with small probability, is

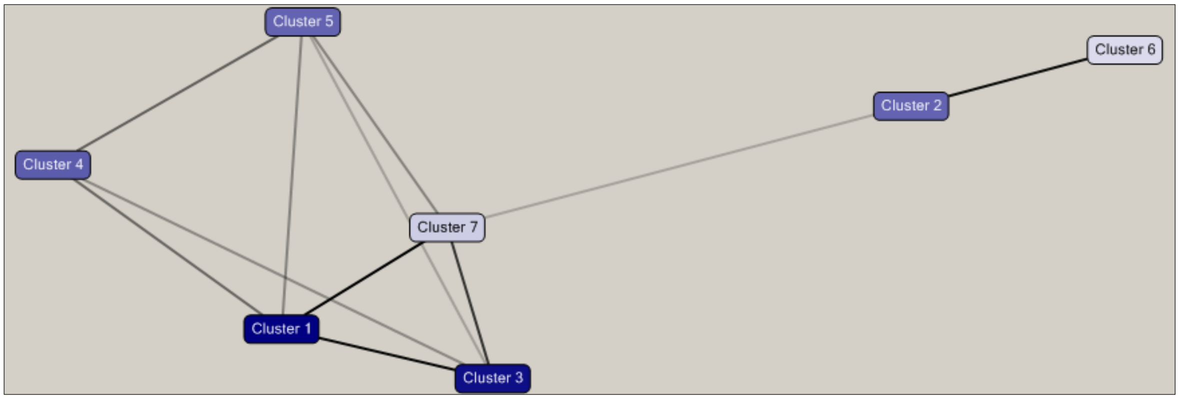

**Figure 3.** Cluster Diagram for TC – The Grošnica reservoir.

**Table 4.** Cluster discrimination - cluster 3 and other clusters - The Grošnica reservoir.

| Variables (units)* | Values | Favors cluster 3 | Favors complement of cluster 3 |
|---|---|---|---|
| TC (MPN/100 ml) | < 135 | 100 | |
| $NH_4^+$ (mg/l) | < 0.1 | 48 | |
| M alkalinity (ml/l) | 29 to 37 | 48 | |
| TC (MPN/100 ml) | > 135 | | 28 |
| month | III | 13 | |
| month | VIII | 10 | |
| $NH_4^+$ (mg/l) | 0.1 to 2 | | 8 |
| EC (µS/cm) | 313 to 467 | 8 | |
| Fe (mg/l) | 0.02 | 6 | |
| $BOD_5$ (mg/l) | < 3.6 | 6 | |
| month | V | | 6 |
| Fe (mg/l) | 0 | 4 | |
| $Cl^-$ (mg/l) | 3.9 to 8.3 | 3 | |
| M alkalinity (ml/l) | 0.8 to 29 | | 3 |
| TP µg/l | 0 | 2 | |
| month | IV | | 2 |
| M alkalinity(ml/l) | > 37 | | 2 |

*see abbreviations.

**Table 5.** Cluster discrimination - cluster 5 and other clusters - The Grošnica reservoir.

| Variables (units)* | Values | Favors cluster 5 | Favors complement of cluster 5 |
|---|---|---|---|
| TC (MPN/100 ml) | 0 to 40 | 100 | |
| M alkalinity(ml/l) | 32 to 37 | 38 | |
| TC (MPN/100 ml) | > 40 | | 25 |
| Month | XI | 11 | |
| FO | 171 to 385 | 9 | |
| M alkalinity(ml/l) | 0 to 32 | | 5 |
| month | VII | 4 | |
| Mn (mg/l) | 0.0 to 0.8 | 4 | |
| TP (µg/l) | 0.02 | 4 | |
| M alkalinity(ml/l) | > 37 | | 2 |

*see abbreviations.

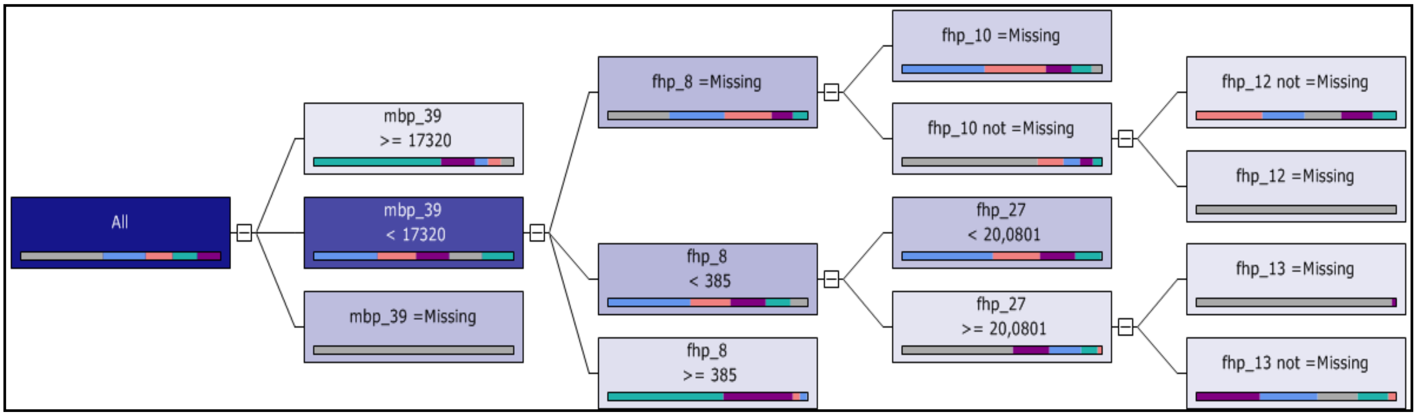

**Figure 4**. Decision Tree for TC - The Gruža reservoir.

**Table 6.** Probabilities of TC values on a root level - The Gruža reservoir.

| Value | Cases | Probability | Color |
|---|---|---|---|
| < 22 | 224 | 13.99 | Pink |
| ≥ 1500 | 203 | 12.70 | Green |
| 150 to 1500 | 195 | 12.20 | Purple |
| 22 to 150 | 346 | 21.50 | Blue |
| Missing | 640 | 39.61 | Gray |

noticeable. Also with small probability, the relation with

## Classify

The classification for TC was made regarding to other physical-chemical and biological parameters. Figure 4 presents decision tree showing the identified pattern upon which the dependency net with FO (mbp_39), EC (fhp_8), COD (fhp_10), turbidity (fhp_27), TP (fhp_13) and $Cl^-$ (fhp_12) was determined for the Gruža reservoir.

For the Gruža reservoir the pattern identified five classes which have been formed based on the grouping of values of TC (Table 6). For each node the probabilities of TC values have been calculated by categories. Each category corresponds to one color on the decision trees diagram, which is proportional to its probability. On the root node (All) which refers to the whole sample irrespective of FO, EC, COD, turbidity, TP and $Cl^-$, probabilities of TC according to categories (Figure 4) are calculated.

The values of TC in the range of 22 to 150 MPN/100 ml are the most common. The level 2 emphasizes the primary influence on FO. Generally, extremely high values of FO cause high values of TC. For FO are ≥ 17320 cfu/ml, the probability for values of TC to be ≥ 1500 MPN/100 ml is 63%, and 150 to 1500 MPN/100 ml, 18%. If the values of FO are < 17320 cfu/ml, the analysis shows that the greatest number of TC is < 150 MPN/100 ml. If FO is < 17320 cfu/ml, then the third level shows the influence of EC on TC. If values for EC exist and are smaller than 385 μS/cm, the number of TC in 61% of cases will be < 150 MPN/100 ml, and if higher than 385 μS/cm, the number of TC will be higher than > 150 MPN/100 ml in 92% of cases. The fourth level shows the influence of COD, and only if there are no values for EC. On the fourth level the equal influence of turbidity is notisable, but only if the values of EC are smaller than 385 μS/cm. The analysis offers two categories of data. The first category emphasise the values of turbidity under 20 NTU, whereas the values of TC < 150 MPN/100 ml are favored in 69%. If the turbidity values are higher than 20 NTU (the second category of data), then TC have the range of appearance 22 to 1500 MPN/100 ml in 75%. If COD not missing the fifth level offers the relation to TP and if turbidity values are higher than 20 NTU then the fifth level offers the relation to and $Cl^-$ (Figure 4).

Figure 5 presents the decision tree showing the pattern by which dependence network of TC with Hm (mbp_38) and $BOD_5$ (fhp_25) has been found for the Grošnica reservoir. By classification, three classes of data regarding to TC have been identified, whereby one class comprises the data with non existing values for TC. For Grošnica reservoir it can be excluded from further analysis, which leaves two classes of data for TC. Table 7 presents classes of data regarding TC on root level. The second level presents classes of values for Hm (if there are no values for Hm there won't be any values for TC, those are the data excluded from the analysis on basic node), while the third level shows the strongest relation between TC and $BOD_5$.

The Grošnica reservoir, unlike the Gruža reservoir, it is evident that beside strong relation with Hm and $BOD_5$, the analysis does not offer ranges of values of variables according to which certain quantity or class of TC appears. It means that classification is applicable from

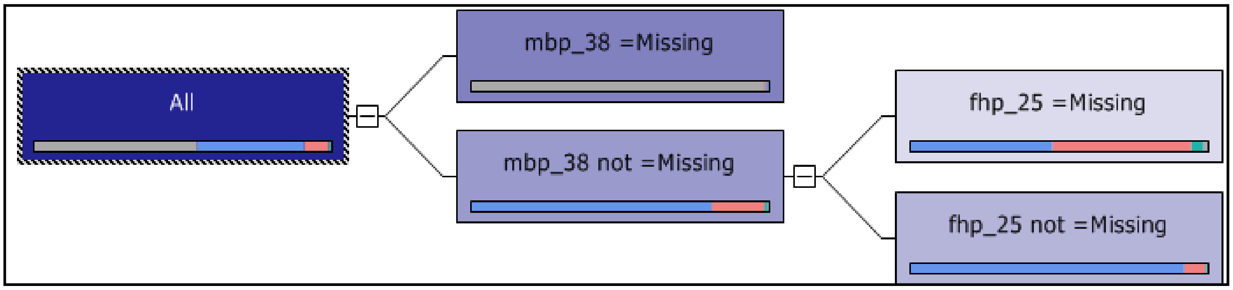

**Figure 5.** Decision Tree for TC – The Grošnica reservoir.

**Table 7.** Probabilities of TC values on a root level - The Grošnica reservoir.

| Value | Cases | Probability | Color |
|---|---|---|---|
| < 130 | 141 | 37 | Blue |
| > 130 | 31 | 9 | Pink |
| Missing | 210 | 54 | Gray |

entered minimum to entered maximum for a certain parameter, and that within that range there are no other regularities. In the Grošnica reservoir, for TC they are 0 to 38000 MPN/100 ml; for Hm 4 to 4533 cfu/ml and for $BOD_5$, 0.06 to 4.18 mg/l.

By classification using the decision tree it is possible to detect the relation of TC and some other physicochemical or biological parameter in the reservoirs. In both reservoirs the strongest relation of TC is to other bacterial community. The specificity shown from the results of analysis in both reservoirs is that TC realizes the strongest relation with community that is non-dominant for specific reservoir. In the Gruža reservoir TC realize the greatest dependence on FO, while in the Grošnica reservoir on Hm. In the Gruža reservoir there is also partial concurrence in CA (higher values of FO indicate higher values of TC). In Grošnica reservoir there is clear relation of number of TC with $BOD_{5.}$ The same relation has been obtained by both, CA and decision tree, which is also confirmed by other authors who used this parameter as one of the basic in modeling of the number of TC (Syed Ahmad et al., 2009). In the Gruža reservoir, the decision tree analysis is more complex and apart from FO there are also concurrences of other parameters with CA. It refers to EC as one of important factors influencing the number of TC. In CA, together with EC, it is also noticeable that the influence of M alkalinity is significant, which indicates that mineral budget of the reservoir has great influence on the number of TC in these reservoirs. Further analysis show the relation of TC with range of parameters regarding to the amount of organic matter in water. In the Gruža reservoir it is noticeable by the relation of TC with COD and turbidity, while previously mentioned relation of $BOD_5$ and TC in the Grošnica reservoir indicates the same thing. The relation between the amount of organic material and the number of coliforms was confirmed by other authors as well. Hong et al. (2010) established the influence of SS, organic and inorganic nutrients on total coliforms. They point out the relation of total amount of carbon with TC, and TSS with FC, wherein the increased amount of TSS is the result of factors from the external environment. A number of authors confirm the existence of strong relation between human activities, their ways of using surrounding land, with the number of coliforms in water. That relation can be direct and under the influence of surface runoff, but also indirect due to the changes of physicochemical and biological factors which eventually cause the changes in number of coliforms (Fisher and Endale, 1999; Tong and Chen, 2002; Kistemann et al., 2002; Mehaffey et al., 2005; Zhang and Lulla, 2006; Derlet et al., 2008; Hong et al., 2010). Byamukama et al. (2005) point out the constant presence of TC in the soil surrounding water and the great influence they have on the number of TC in water. They also point out the significant relation of TC with EC and TSS in waters under the great influence of anthropogenic activities and contamination.

Results indicate that it is possible, with help of CA, to separate localities and depths showing the difference in number of TC. The Gruža reservoir is bigger regarding the area it covers and volume, the number of sampling localities is greater, and the number of samples per locality more balanced. In the Gruža reservoir the number of TC varies regarding to depths, it is different in the deepest and in the shallowest parts. This can be explained by smaller anthropogenic influence in the deepest parts due to the existence of a vegetation zone along the reservoir shore, by greater average depth, as well as by the use of hypolymnetic aerator. In the shallowest part of reservoir, especially in mouths of tributaries as well as in the tributaries, the anthropogenic influence is great (recreational activities, cultivated fields, use of pesticides, fertilizers, industrial activity on tributaries etc.). Here CA directly indicates that the number of TC is influenced by anthropogenic activity in the area. In Grošnica reservoir CA does not offer the same possibility primarily due to small number of data as well as their structure. The Grošnica reservoir is small regarding the area it covers and volume, most samples are from one locality only (the dam), so it is not possible to determine regularly the influence of other localities. In the Grošnica reservoir it gives the influence of temporal dimension on the number of TC, which is not noticeable in Gruža. The reason for this is found in different geographical position and elevation. In this reservoir CA is directed to prediction and it gives the information about the ranges in which some characteristics follow the certain number of TC, which decision tree in this case does not present. CA with *K-means* algorithm gives satisfying results about the analyses of physicochemical and biological parameters in water monitoring (Areerachakul and Sanguansintukul, 2010).

All noted indicates that parameters of water quality collected by standard hydrobiological researches could be used for making models which would efficiently enable monitoring of dynamics and prediction of the state of microorganisms such as total coliforms (Brion et al., 2001, 2002). The application of created models could be significant for the improvement of water quality, reduction of the expenses of monitoring and management of water resources.

## Conclusion

Analysis and modeling of microbiological parameters is an important aspect in evaluation of state and quality improvement of freshwater ecosystems. The tools of data mining are becoming even more significant in this area of research, and are being applied in data analysis. In this paper a combination of cluster analyses and classification has been used. Data mining models are built upon data source views which represent an abstraction layer over existing SeLaR information system. This approach is more flexible and effective comparing to traditional data warehouse approaches, and it provides a single metadata model for creation of data mining models. By water quality modeling and prediction of state of total coliforms, a good presentation of one dynamic system could be obtained. Designed data mining models allow identification of previously unknown relationships and provide predictions, so it is possible to make more informed water management decisions. In this way, a new dimension of monitoring of reservoirs and lakes is provided. The approach presented in this study could be one of the valuable techniques for managing water resources.

## ACKNOWLEDGEMENTS

This study was supported by the Ministry of Science and Education of the Republic Serbia (grants No. III 44010 and III 41010).

## REFERENCES

APHA (1998) Standard Methods for the Examination of Water and Waste Water, 20th Edition. Part 2000:1-92; 4000:1-180; 9000:1-140; 10000:1-28; American Public Health Association, Washington DC.

Areerachakul S, Sanguansintukul S (2010). Clustering Analysis of Water Quality for Canals in Bangkok, Thailand. In Taniar et al. (eds) Computational Science and Its Applications, ICCSA 2010 Lecture Notes in Computer Science, LNCS, Part III, 6018: 215–227.

Bagde US, Rangari AK (1999). Periodicity of coliform bacteria in an aquatic environment. Water. Sci. Techn., 40(7): 151-157.

Byamukama D, Mach RL, Kansiime F, Manafi M, Farnleitner AH (2005). Discrimination Efficacy of Fecal Pollution Detection in Different Aquatic Habitats of a High-Altitude Tropical Country, Using Presumptive Coliforms, *Escherichia coli*, and *Clostridium perfringens* Spores. Appl. Environ. Microbiol., 71(1): 65–71.

Bergstein Ben-Dan T, Shteinman B, Kamenir Y, Itzhak O, Hochman A (2001). "Hydrodynamical Effects on Spatial Distribution of Enteric Bacteria in the Jordan River—Lake Kinneret Contact Zone". Water Res., 35(1): 311-314.

Boyacioglu H, Boyacioglu H (2007). Surface Water Quality Assessment by Environmetric Methods. Environ. Monit. Assess., 131:371–376.

Brion GM, Neelakantan TR, Lingireddy S (2001). Using Neural Networks to Predict Peak Cryptosporidium Concentrations. J. Am. Water Works Ass. (AWWA)., 93(1): 99-105.

Brion GM, Neelakantan TR, Lingireddy S (2002). A neural network based classification scheme for sorting sources and ages of fecal contamination in water. Water Res., 36(15): 3765–3774.

Canale RP, Patterson RL, Gannon JJ, Powers WF (1973). Water Quality Models for Total Colform. Water Pollut. Control Fed., 45(2): 325-336.

Curtis TP, Mara DD, Silva SA (1992). Influence of pH, oxygen, and humic substances on ability of sunlight to damage fecal coliforms in waste stabilization pond water. Appl. Environ. Microbiol., 58(4): 1335–1343.

Ćurčić S, Čomić LJ (2002). A microbiological index in estimation of surface water quality. Hydrobiologia., 489(1-3): 219-224.

Derlet RW, Ali Ger K, Richards JR, Carlson JR (2008). "Risk Factors for Coliform Bacteria in Backcountry Lakes and Streams in the Sierra Nevada Mountains: A 5-Year Study". J. Wilderness Environ. Med., 19(2): 82-90.

Kumar P, Alameda J, Bajcsy P, Folk M, Markus M (2006). Hydroinformatics: Data Integrative Approaches in Computation, Analysis and Modeling. CRC Press, pp. 1-534.

Fisher DS, Endale DM (1999). Total *Coliform*, *E. coli*, and *Enterococci* Bacteria in Grazed and Wooded Watersheds Of The Southern Piedmont. Proceed. Georgia Water Res. Conf., In Hatcher KJ (ed), Georgia. CRC Press.

Gameson ALH, Saxon JR (1967). Field studies on effect of daylight on mortality of coliform bacteria. Water Res., 1: 279-295.

George I, Anzil A, Servais P, (2004). Quantification of fecal coliform inputs to aquatic systems through soil leaching. Water Res., 38: 611–618.

Gorunescu F (2011). Data Mining: Concepts, Models and Techniques. Intelligent Systems Reference Library, Springer-Verlag Berlin Heidelberg, p.12.

Han J, Kamber M, Pei J (2010). Data mining: concepts and techniques, 3rd ed, Elsevier Inc., pp. 327-349.

Hong H, Qiu J, Liang Y (2010). Environmental factors influencing the distribution of total and fecal coliform bacteria in six water storage reservoirs in the Pearl River Delta Region, China. J. Environ. Sci., 22(5): 663–668.

Idakwo PY, Abu GO (2004). Distribution and Statistical analysis of bacteria in Lake Alau in the arid northern Nigeria, J. Appl. Sci. Environ. Mgt., 8(1): 5-9.

Iscen CF, Emiroglu Ö, Ilhan S, arslan N, Yilmaz V,ahiska S. (2008). Application of multivariate stastical techniques in the assessment of surface water quality in Uluabat lake, Turkey. Environ. Monit.

Assess., 144:269-276.

Juhna T, Birzniece D, Rubulis J, (2007). Effect of phosphorus on survival of *Escherichia coli* in drinking water biofilms. Appl. Environ. Microbiol., 73(11): 3755 – 3758.

Kistemann T, Claen T, Koch C, Dangendorf F, Fischeder R, Gebel J (2002). Microbial load of drinking water reservoir tributaries during extreme rainfall and runoff. Appl. Environ. Microbiol., 68(5): 2188–2197.

MacLennan J, Tang ZH, Crivat B (2010). Data Mining with Microsoft SQL Server 2008, Wiley Publishing, Inc., pp. 291-318.

Mahloch JL (1974). Comparative Analysis of Modeling Techniques for Coliform Organisms in Streams. Appl. Microbiol., 27(2): 340-345.

McCambridge J, McMeekin TA (1984) 'Effect of solar radiation and predacious microorganisms on survial of fecal and other bacteria', Appl. Environ. Microbiol., 41: 1083–1087.

Medema GJ, Shaw S, Waite M, Snozzi M, Morreau A, Grabow W (2003). Catchment characterization and source water quality. In Dufour et al. (eds) Assessing Microbial Safety of Drinking Water: Improving Approaches and Methods, WHO, OECD, London, pp. 111–158.

Mehaffey MH, Nash MS, Wade TG, Ebert DW, Jones KB, Rage RA (2005). Linking land cover and water quality in New York City's water supply watersheds. Environ. Monit. Assess., 107: 29–44.

Mundy J, Thornthwaite W, Kimball R (2011). The Microsoft Data Warehouse Toolkit: With SQL Server 2008 R2 and the Microsoft Business Intelligence Toolset, Wiley Publishing, Inc., pp. 29-78.

Ostojić A, Ćurčić S, Čomić Lj, Topuzović M (2005). Estimate of the Eutrophication Process in Gruža Reservoir (Serbia and Montenegro). Acta Hydroch. Hydrob., 33: 605-613.

Ostojić A, Ćurčic S, Čomić Lj, Topuzović M (2007). Application of the Peg Model to Two Reservoirs with Different Trophic Levels. Ekologia (Bratislava), 26(4): 409-429.

Rompré A, Servais P, Baudart J, de-Robine MR, Laurent P (2002). Detection and enumeration of coliforms in drinking water: Current methods and emerging approaches. J. Microbiol. Meth., 49(1): 31-54.

Saffran K (2001). Canadian water quality guidelines for the protection of aquatic life, CCME water quality Index 1,0, User`s manual. Excerpt from Publication no.1299.

Sargaonkar A, Deshpande V (2003). Development of an overall index of pollution for surface water based on a general classification scheme in Indian context. Environ. Monit. Assess., 89: 43–67.

Simeonov V, Einax JW, Stanimirova I, Kraft J (2002). Environmetric modeling and interpretation of river water monitoring data. Anal. Bioanal. Chem., 374: 898–905.

Stefanović D, Radojević I, Čomić Lj, Ostojić A, Topuzović M, Kaplarević-Mališić A (2012). Management Information System of Lakes and Reservoirs. Water Resour., 39(4): In press.

Syed Ahmad SM, Turki MB, Malek S (2009). Intelligent Computational Modeling and Prediction of Coliform Growth in Tropical Lakes based on Hybrid Self Organizing Maps (SOM) and Fuzzy Logic Approaches. eJCSIT Electron. J. Computer Sci. Inform. Techno., 1(1): 18-22.

Tong STY, Chen W (2002). Modeling the relationship between land use and surface water quality, J. Environ. Manage., 66: 377-393.

Youn-Joo A, Kampbellb DH, Breidenbach GP(2002). *Escherichia coli* and total coliforms in water and sediments at lake marinas, Environ. Pollut, 120: 771–778.

Zhang X, Mukesh L (2006). Evaluation of Pathogenic Indicator Bacteria in Structural Best Management Practices. J. Environ. Sci. Health Part A, 41: 2483–2493.

WHO (2008). Guidelines for Drinking-water Quality (3rd ed., incorporating first and second addenda). World Health Organization Press, Switzerland. 1: 281-294.