EnviroInfo 2005 (Brno)
Informatics for Environmental Protection - Networking Environmental Information
 ISBN: 80-210-3780-6

Classification of biological communities in biomonitoring programs – suggestion of robust solution

Jiří Jarkovský[1], Ladislav Dušek[1], Petr Pavliš[1], Klára Kubošová[1], Jan Hodovský[2], Jiří Hřebíček[1]

Abstract

Monitoring of water organisms communities become a standard approach in surface water monitoring and the analytic approach is commonly based on comparison with reference sites. There are several problems with this analysis including selection of appropriate methods with respect to nature of the data. We tested two different "non-parametric" methods for multivariate analysis of reference and standard sites association (i.e. method suitable for classification of unknown sites) consisting of i) robust true distances of sites based on several data views – biotic, static and dynamic abiotic properties and ii) non-parametric comparison of differences among sites and/or their groups (i.e. classification). The first classification algorithm is based on nearest neighbor method, the second is newly developed "centroids distance" algorithm based on percentiles of multivariate homogeneous clusters of localities. The relevance of algorithms were first tested on the set of well known reference data and the results were compared to "classical" methods in this field like discriminant analysis or neural networks. The evaluated non-parametrics methods revealed the same or better results than classical approach probably due their non-sensitiveness to problems with outliers and multivariate distribution of data and are proposed to become standard methodology for analysis of biomonitoring data in the Czech Republic. The methods were then applied on real data of biomonitoring network of the Czech Republic (macrozoobenthos communities and sites abiotic description data); marked differences among reference sites and sites under environmental stress were found and these differences are in relation to state of sites evaluated by more conventional methods or expert opinion. The presented methodology is under implementation into the multicentric expert system for analysis and management of biomonitoring data – Triton.

1. Introduction

Monitoring of water organisms communities has become a standard approach in surface water monitoring as well as a part of complex systems for assessing surface water quality. In European countries, the most commonly used organisms are water macroinvertebrates; developed in the UK (Wright et al., 2000, Clarke et al., 2003), RIVPACS has been one of the first complex systems based on macroinvertebrates. Similar systems have been used in many countries worldwide (Barbour et al., 1999) and development of similar systems is also connected to The European Water Framework Directive (Logan and Furse, 2002; Directive 2000/60/EC).

In the Czech Republic, there exists a huge biomonitoring network covering most river biotopes in the Czech Republic. Surface water quality assessment on the national level is based on production of huge amount of data and; thus, we should be able to manage and analyse them with respect to their nature. For this purpose, we are proposing a network-based and centrally-managed system (TRITON) capable of analysis and presentation of biomonitoring data. The main aim of the system is automatic preparation of

[1]Centre of Biostatistics and Analyses, Faculty of Medicine and Faculty of Science, Masaryk University, Kamenice 126/3, 625 00 Brno, Czech Republic

[2] RECETOX, Faculty of Science, Masaryk University, Kamenice 126/3, 625 00 Brno, Czech Republic

summary outputs on surface water quality within routine monitoring based on i) univariate description/comparison with legal standards and ii) robust multivariate analysis and quality assessment. The analytical methods included cover all kinds of abiotic measures as well as an analysis of macrozoobenthos (or other organisms) communities; all data types are combined in a robust multivariate analysis based on comparison with the reference quality. A presentation of the TRITON system and its newly developed robust multivariate methodology for surface water quality assessment are the main aims of this paper.

2. Surface water quality assessment in the triton system

2.1 The TRITON system – a solution for evaluation of biomonitoring data

The development of the first version of the TRITON system (processing only data on abiotic monitoring) started in 1999; in collaboration with the Agricultural Water Management Authority of the Czech Republic, the present version of the system processing both biotic and abiotic monitoring data has been developed since 2002.

The present version of the system has been developed with the intention to use it for complex analysis of surface water quality based on biological communities, i.e. with the ability to implement analogical methodology as described in outputs of European international projects in this field (e.g. AQEM: Sandin et al. 2000, 2001; Hering et al., 2001).

The monitoring data are collected from regional offices of the Agricultural Water Management Authority using the technology of www forms and stored in a central database system called *Salamander*. The validated data are then used as a data source for TRITON clients installed on computers of the system users. Its operation in the server-client environment requires sophisticated management of users' rights. The TRITON software identifies three main groups of users: i) the Supervisor ("mother") can use the complete set of functions and analyses. Moreover, they can manage rights of all other users, define templates for routine reports and set the reference models. The users' rights and settings of routine analyses are stored on a central server and get updated for all users on their login into the system. ii) "an Explorer" can use all analytical functions of the software: for example, these could be the scientists developing the reference model. iii) "Workers" are standard system users. They use the TRITON system for preparation of routine reports on biomonitoring results and are granted access to selected analyses and filtered data only.

Standard users access their analyses from a simple user-friendly interface; they do not need the knowledge of statistical analysis or programming: the outputs are directly computed as "ready-to-use" presentations based on predefined templates. Behind this simple interface, there are several modules directly accessible only for advanced users (mother, explorer); these provide the functionality of the software including the following most important functions: i) an analytical module providing statistical data analysis (common descriptive and comparative univariate statistics, graphs etc.) ii) a GIS module displaying monitored localities on the map of the Czech Republic; it is used inside presentations or for direct selection of sampling localities. iii) a presentation module providing graphical presentations of outputs as well as design of templates for routine analyses. iv) a multivariate comparative module computing robust multivariate analyses of surface water quality.

The TRITON system has been designed for maximum flexibility and with the vision of further development: it allows simple additions of new equations for water quality assessment and only minor changes are required for incorporation of new data (e.g. new taxonomical groups for biomonitoring).

Since the end of 2003, the system has been used in routine work at the Agricultural Water Management Authority offices. It aids preparation of legislatively required reports on surface water quality – especially data on abiotic monitoring. Concerning biomonitoring data, the use of the whole system potential is still

limited due to absence of a multivariate reference model for complex assessment of biological communities.

2.2 Scientific background of surface water quality assessment

Concerning analysis of surface water quality based on biological communities, there are several methodology approaches: single metrics (Washington, 1994), multimetric (Barbour et al., 1999; Barbour and Yoder, 2000) and a multivariate approach (Norris and Georges, 1993, Wright et al., 2000). The TRITON implements basic single metrics like diversity and biotic indices; nevertheless, the most desirable method is a complex multivariate or multimetric system. We decided to develop the multivariate method of complex quality analysis of sampling localities; we follow a general design of this kind of analysis adding some novel modification to statistical methodology. Although the multimetric approach has not been implemented yet, as stated above, the TRITON system is very flexible and makes it possible for a multimetric approach to be added in the future.

The general concept of quality assessment by comparison with the reference quality is very simple and it is implemented in almost all complex systems of surface water quality assessment; however, its implementation presents a difficult task. The process could be divided into three steps, each of them with its own methodology problems: i) preparation of the reference model, ii) classification of unknown cases (sites) into reference categories and iii) comparison of an unknown site and the reference status, i.e. assessment of unknown site quality.

The principal basis of the whole analysis is the quality of the reference dataset, which should homogeneously cover all environmental conditions in the analysed area (small rivers of the Czech Republic in our case) and contain minimal influence of human activities. Some guidelines for selection of reference localities are given in Hughes, 1994. Unfortunately, in the conditions of Central Europe, it is almost impossible to find real ”natural” sites; thus, localities in the reference dataset originate from long-term knowledge of the considered sites and consensus of hydrobiology experts on what is the site’s “nearest-to-natural-conditions” status.

Although reference data represents the “natural” quality of rivers, they could be (and, in fact, are) very heterogeneous from the hydro-morphological point of view as well as according to the composition of their macrozoobenthos communities; this composition of communities is the criterion of the “natural” quality. Also, there is a link between hydro-morphology and communities composition. First, we have to define a homogeneous group within the reference database based on comparison of biological communities composition, i.e. to define a reference model consisting of several homogeneous categories according to their community composition. These groups should also be defined by hydro-morphological parameters of the respective sites, i.e. by parameters much less influenced by human activities – or even not influenced at all. According to a theory stating that the composition of biological communities is highly influenced by the environment, a link may be identified between standard and reference sites to see if, under the same environmental parameters, these sites have got the same or different community composition. In other words, if there is a shift in community composition in case of a standard locality in comparison to the “natural” (reference) composition of biological community under the same environmental conditions (in the sense of hydro-morphological or hydro-geological properties of the respective sampling site).

There are several methods of building up reference models like cluster analysis, k-means clustering, TWINSPAN or neural networks. This step of analysis is supervised by experts and it is mostly scientific work; the influence of problems concerning data, their distribution etc. could be evaluated and balanced in a model. Concerning the TRITON system, the reference model is still under development.

2.3 Suggested robust method of comparison with the reference quality

The second step of analysis includes the classification of unknown cases into categories of reference data defined by hydro-morphological and hydro-geological parameters of the respective sampling sites (parameters of natural heterogeneity). Unknown cases are classified according to parameters: i) minimally influenced by human activities and ii) influencing the composition of biological communities.

There is a number of methods for multivariate classification of objects, e.g. logistic regression, discriminant analysis or neural networks; however, these also have their problems, e.g. prerequisites of normality and absence of outliers for discriminant analysis (Legendre & Legendre). Moreover, the methods included in this step of analysis may be used in a routine way in monitoring, i.e. without proper analysis of problems concerning the data. Keeping these facts in mind, we attempted to develop a robust multivariate method suitable for classification of unknown cases with minimum sensitivity to data distribution problems; and thus, suitable for routine use in biomonitoring practice.

The most simple and the most objective measure of object association in multivariate space is their distance; thus, we decided to build our method on an analysis of a distance matrix among localities. Now, selection of proper distance metric is the first task in designing the method. We have adopted Gower distance metric (Gower, 1971); however, any multivariate distance metric suitable for given data could be used. Concerning biomonitoring data, there are some advantages in Gower metric:

i) Continuous, binary or categorical parameters may be incorporated in computation: binary data is computed by coefficient – agreement and disagreement of values forming distance 0 or 1 respectively; categorical data is computed in the same way. Distance of objects according to continuous data is weighted to i) a parameter range in the data file or ii) an externally provided parameter range, i.e. difference in parameter values of objects is divided by parameter range to obtain partial metric ranging from 0 to 1.

ii) As noted above, parameters are weighted to their range, i.e. the influence of parameter absolute value is removed.

iii) The final distance metric ranges from 0 to 1 and could be easily interpreted.

iv) Parameters in computation could be weighted according to expert knowledge or results of preliminary analysis. The final metric takes the following form:

$$D\ (x_1, x_2) = \frac{\sum_{j=1}^{p} w_j d_{12\,j}}{\sum_{j=1}^{p} w_j}$$

, where D is distance between objects x_1 and x_2, d_{12j} is partial distance of objects x_1 and x_2 associated with parameter j (there are 1..p parameters; partial metric associated with parameter ranges from 0 to 1) and w_j is weight of parameter j ranging from 0 to 1.

Every homogeneous category of reference data could be characterised by its position in the multivariate space; and also, by its multivariate variability. Position of the reference category centroid (based on the median of continuous data and modus of binary/categorical data) exhibits representative of this group; multivariate radius of group provides the measure of its variability (in fact 95% percentile of radius is used in our computation to remove the influence of outliers). The distance of an unknown case to the centroid (**D**) is compared to the percentile of the reference category range (**R**). This ratio measures the extent to which an unknown case differs from objects incorporated in the reference category – see figure 5. Due to the fact that reference categories are not probably multivariate spheres we had to add a safety measure reflecting the real multivariate shape of the reference data. There are two parameters incorporated in the computation: the distance of an unknown case to the nearest neighbour in the reference group (**N**) and the measure of intragroup distances (**I**) within the reference group. The measure of intragroup distances is taken as median length of the MST branches (minimal spanning tree, Prim (1957)) of objects in the refer-

ence group. The following formula gives the measure of distance of an unknown case to the reference group x (U_x) in multiplies of the reference group x radius weighted for multivariate shape of this group.

$$U_x = \frac{abs(D + N - I)}{R}$$

This computation could be also expressed as a probability of case U belongs to group x: $P(U_x) = \frac{1}{U_x} \times 100$ where values over 100% (i.e. objects inside the reference group) are truncated to 100%. In the first step of the analysis (classification of an unknown case into reference groups according to natural heterogeneity), $P(U_x)$ is computed for all reference groups x=1..n and probability of unknown case belongs to a particular group is weighted as follows:

$$PW(U_x) = \frac{P(U_x)}{P(U_1) + P(U_2) + \cdots + P(U_n)}$$

In the second step of the analysis, U_x or $P(U_x)$ based on abiotic/biotic parameters (i.e. parameters linked to sampling site quality) is adopted for assessing distance/similarity of an unknown case from/to a particular reference group (quality assessment).

There are two outputs of the analysis based on the above-mentioned non-parametric computations:

- The probability of assigning a locality into the reference class based on natural heterogeneity, i.e. to which reference class the evaluated locality belongs.
- The difference between the evaluated locality and the given reference class based on abiotic and biotic parameters with respect to human influence and pollution, i.e. quality of the evaluated locality. Quality of the evaluated locality is either expressed as a percentage difference from the reference quality or it may be recoded into several quality classes; it serves as the final output of multivariate quality assessment in the biomonitoring system.

One of the main aims of this paper is to test the predictive power of this methodology against the most common classification methods like discriminant analysis, classification trees and neural networks.

2.3.1 Validation analysis of suggested methodology

The presented methodology was tested on real datasets of 300 reference localities thorough the whole Czech Republic. First, the localities were divided into 8 homogeneous cluster using k-means clustering. The clusters were based on parameters of natural heterogeneity (ecoregion, Strahler order, main river basin, width and depth of the stream, distance from well and altitude), the importance of factors and mutual clusters position was validated using principal component analysis. The analyses were performed by Statistica for Windows (StatSoft, Inc., 2005).

The classification methods applied on data were:

- The novel method ("centroid distance") mentioned above
- Discriminant analysis
- Classification tree
- Neural network

The dataset with the given groups of localities was divided into two files which were used for cross validation in these analyses and the following results of application of the models on independent cross validation datasets were obtained:

- Centroid distance: 91.3%

- Discriminant analysis: correct classification 87.6%
- Classification tree: 94.7%
- Neural network: 93.7

The results suggest that the developed methodology has similar predictive power as the commonly used methods or even better than some of them (discriminant analysis).

3. Conclusions

Although there are several organisations in the Czech Republic dealing with biomonitoring of surface waters, TRITON is the first system working in cooperation with the central biomonitoring database and providing complex services for biomonitoring data on the national level – including routine reporting on abiotic monitoring and biomonitoring results as well as complex multivariate assessment of water quality.
TRITON includes a newly-developed robust multivariate method of water quality assessment; this methodology, due to its non-parametric nature, is less sensible to outliers and other problems concerning data distribution than more conventional methods (e.g. discriminant analysis in the RIVPACS system); moreover, it is a very promising method suitable for comparison of biomonitoring network localities and reference classes. On the other hand, this method is also dependent on a reference model available; yet, the model for routine usage is still under development.
Last year, the TRITON system was introduced into routine work on the national level in the Agricultural Water Management Authority of the Czech Republic where it is used for analyses and reporting on data collected during 10 years of abiotic monitoring (more than 1000 localities) and 2 years of biomonitoring (about 300 localities) of small surface waters.
There are three main tasks in the future of TRITON: i) in cooperation with hydrobiologists, a reference model should be developed for quality assessment allowing full utilisation of the SW multivariate module in routine biomonitoring; ii) at present, the system works only with data on small surface waters; nevertheless, there are also data on other types of surface waters and our ambition is to design a database binding all abiotic monitoring and biomonitoring data in one complex system; iii) nowadays, biomonitoring in the Czech Republic is based on macrozoobenthos communities; this should be extended to analysis of other taxonomical groups like fishes or phytobenthos.

4. Acknowledgements

The study was supported by the Ministry of Education of the Czech Republic (project No. MSM 0021622412 INCHEMBIOL). We would like to thank the Department of Zoology and Ecology, Masaryk University, and the Agricultural Water Management Authority of the Czech Republic for providing us with biomonitoring datasets.

Bibliography

Barbour, M.T. et al., 1999. Rapid Bioassessment Protocols for Use in Streams and Wadeable Rivers: Periphyton, Benthic Macroinvertebrates and Fish, Second Edition. EPA 841-B-99-002. U.S. Environmental Protection Agency; Office of Water; Washington, D.C.

Barbour, M.T. and Yoder, C.O. 2000. The multimetric approach to bioassessment, as used in the United States of America. In: Assessing the biological quality of fresh waters. RIVPACS and other techniques (editors J.F. Wright, D.W. Sutcliffe and M.T. Furse). Freshwater Biological Association, Ambleside, Cumbria, UK.

Clarke, R.T. et al. 2003. RIVPACS models for predicting the expected macroinvertebrate fauna and assessing the ecological quality of rivers. Ecological Modelling 160, pp. 219-233

Directive 2000/60/EC - Establishing a Framework for Community Action in the Field of Water Policy

Gower, J.C., 1971. A general coefficient of similarity and some of its properties. Biometrics 27, pp. 857–871

Hering, D. et al. 2001. AQEM 2nd deliverable: Reference biocoenoses and deviations: structure and tools for description. AQEM consortium (www.aqem.de)

Hughes, R.M., 1994. Defining acceptable biological status by comparing with reference conditions. - In: Biological assessment and criteria. Tools for water resource planning and decision making.- In: Davies, W. S. and T. P. Simon (eds.), Lewis Publishers, Boca Raton, Florida, pp. 31-48.

Legendre, P. and Legendre, L., 1998. Numerical ecology. Elseviere Science BV, Amsterdam.

Logan, P., Furse, M., 2002. Preparing for the European Water Framework Directive - making the links between habitat and aquatic biota. Aquatic Conservation-Marine And Freshwater Ecosystems 12 (4), pp. 425-437.

Prim, R.C., 1957. Shortest connection networks and some generalizations. Bell Syst Tech J 36, pp. 1389–1401.

Sandin, L. et al. 2000. AQEM 1st deliverable: Stream assessment methods, stream typology approaches and outlines of a European stream typology. AQEM consortium (www.aqem.de)

Sandin, L. et al. 2001. AQEM 3rd deliverable: Experiences with different stream assessment methods and outlines of an

integrated method for assessing streams using benthic macroinvertebrates. AQEM consortium (www.aqem.de)

StatSoft, Inc. (2005). STATISTICA (data analysis software system), version 7.1. www.statsoft.com.

Washington, H. G. 1984. Diversity, biotic and similarity indices. A review with special relevance to aquatic ecosystems. Water Research 18, pp. 653-694.

Wright, J.F. et al., 2000. Assessing the biological quality of freshwaters: RIVPACS and similar techniques. Freshwater Biological Association.