Assessment of Ecological State of Surface Waters in ARROW Project: Robust Multivariate Predictive Models

J. Jarkovský[a], J. Rácek[b], D. Némethová[a], P. Brabec[a], J. Hodovský[c]

[a] *Centre of Biostatistics and Analyses, Masaryk University, Brno, Czech Republic*

[b] *Faculty of Informatics, Masaryk University, Botanická 68a, 602 00 Brno, Czech Republic*
racek@fi.muni.cz

[c] *Ministry of Environment of the Czech Republic, Prague, Czech Republic*

Abstract: The ARROW (www.cba.muni.cz/arrow/eng - Assessment and Reference Reports of Water monitoring) project is the project of implementation of EU Water Framework Directive in the monitoring of surface waters of the Czech Republic and covers all aspects of this problem from data sampling to informatics solution (focusing on the ecostat). The project is based on long term development of this field in the Czech Republic and it is carried out under the supervision of the Ministry of the Environment of the CR.
The important part of the ARROW project is the development and implementation of a Czech approach to the evaluation of the ecological state of surface waters using an analysis of monitoring data. The main idea of the system is an approach based on a network of reference sites and robust multivariate modeling of expected environmental and biological conditions (could be called "RIVPACS type"). The approach presented is compatible with the EU WFD and will be implemented in the national-wide information system of the ecological state of surface waters of the CR. The evaluation of ecological state in ARROW project is based on two former projects; the direct predecessor is TRITON project aimed on analysis of biomonitoring data from small watercourses under collaboration with Agricultural Water Management Authority of the Czech Republic which has been developed since 1999. The other source of inspiration and especially reference dataset is also PERLA system which should be considered as a scientific background for biomonitoring activities in the CR. The methodology is flexible and robust and could be adopted for different types of data and biological communities. All levels of the process are covered by objective statistical methodology and monitored by experts; their computations are based on robust multivariate and multimetric methods, some of them newly developed.

Keywords: biomonitoring; multivariate analysis; environmental modeling; RIVPACS; FWD EU.

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 and Yoder [2000]) and development of similar systems is also connected to The European Water Framework Directive (Logan and Furse [2002]; Directive 2000/60/EC).
Nowadays the demands of the EU WFD call for the complex system of evaluation of ecological state integrating all existing surface water monitoring networks and systems in the Czech Republic. The ARROW project of the Ministry of Environment of the CR covers all aspects of this problem from data sampling to informatics solution. The project incorporates both WFD demands and complex solution of sampling, biological, analytical and ICT problems. The integral part of the project is also the definition and calibration of reference network of sites. The project is based on long term

development of this field in the CR and is related to parallel project of Ministry of the Environment of the CR connected to monitoring programs. The solution of the project binds many experts from both scientific and commercial institutions in this field.

The aim of this paper is to present the methodology of evaluation of ecological state as the very important part of the project; the other parts of the project are only noted.

The main idea of the system is an approach based on a network of reference sites and robust multivariate modeling of expected environmental and biological conditions (could be called "RIVPACS type").

The evaluation of ecological state in ARROW project is based on two former projects; the direct predecessor is TRITON project aimed on analysis of biomonitoring data from small watercourses under collaboration with Agricultural Water Management Authority of the Czech Republic which has been developed since 1999. The other source of inspiration and especially reference dataset is also PERLA system which should be considered as a scientific background for biomonitoring activities in the Czech Republic.

2. EVALUATION OF ECOLOGICAL STATE

2.1 Project prerequisites

In the Czech Republic, there exists a huge monitoring network covering most river biotopes in the CR. The sampling is on regular base and covers the chemical, physical and hydrological parameters together with composition of biological communities. The data are stored in central database; the sampling procedures, data storage and reporting of results of evaluation of ecological state are other parts of the ARROW project.

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 and Yoder [2000]) and a multivariate approach (Wright et al. [2000]). The ARROW implements all these metrics on different levels of computation. The main approach based on multivariate 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. In the third step the difference in single metrics between observed and predicted state are used together with multivariate differences between biological communities for partial evaluation of the ecological state. Above the direct comparison with expected natural state there is multimetric combination of partial measures which forms the final measure of ecological state.

The principal basis of the whole analysis is the quality of the reference dataset, which should homogeneously cover all environmental conditions in the analyzed area 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.

2.3 Building up the model

According to an ecological 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 sampling site). The usage of reference datasets in evaluation of expected natural conditions consists of several steps.

(I) 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. Hierarchical agglomerative clustering on distance matrix of biological communities followed by algorithm for definition of optimal number of clusters and expert opinion will be used for definition of reference groups.

(II) These groups should 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; these parameters are used for classification of unknown sites (standard monitoring) into reference groups

according to theory that sites with similar environmental conditions should have similar biological communities.
(III) For the comparison of observed and predicted conditions the natural environmental state should be defined on reference groups; the natural conditions should be defined for biological communities, environmental metrics and chemical pollutants.
(IV) The sites of monitoring network with unknown ecological state should be classified into reference categories. There are a number of methods for multivariate classification of objects, e.g. logistic regression, discriminant analysis or neural networks; however, these also have their problems. 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 method is discussed in details in separate chapter.
(V) The final goal of the system is the measure of environmental disturbance of sites of monitoring network; it is defined as difference between observed ecological state of site and its expected state defined by probability of site classification into reference categories and the description of natural conditions based on reference dataset. The evaluation of state is computed for different partial measures of ecological state: Similarity measure (Jaccard coefficient) between predicted and observed composition of biological community and the EQI between measured value and median of predicted distribution of chemical pollutants or environmental metrics. The Ecological Quality Indices (EQI), according to RIVPACS methodology (Wright et al., 2000), is defined as the ratio of observed to expected values (O/E) of each parameter being used. The intention of using such ratios is that it provides a means of standardizing the biotic indices, so that a particular value of the EQI ratio implies the same ecological quality for that index.
(VI) The last step of the analysis is combination of partial measures of ecological states into one final measure which could be categorized into five categories according to WFD EU (from natural to most disturbed ecological state) and used for environmental reporting and management.

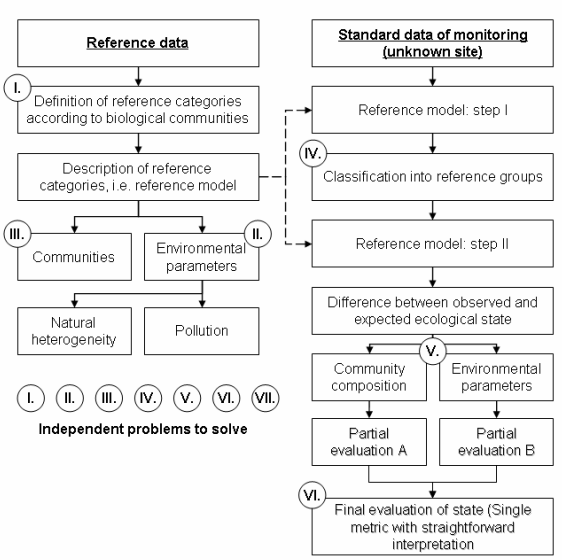

Figure 1 Reference model and its usage (see the text for details)

3. SUGGESTED ROBUST METHOD OF ANALYSIS

There are two parts of the methodology where the original approach of comparison with reference state is modified with aim to enhance the statistical robustness of the analysis.

3.1 Suggested robust method of classification into reference classes

The simplest 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 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:

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

,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 with range 0-1.
Every homogeneous category of reference data could be characterized 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 neighbor 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 reference 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.

$$U_x = \frac{abs(D+N-I)}{R} \quad (2)$$

This computation could be also expressed as a probability of case U belongs to group x:

$$P(U_x) = \frac{1}{U_x} \times 100 \quad (3)$$

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 (4):

$$PW(U_x) = \frac{P(U_x)}{P(U_1)+P(U_2)+\cdots+P(U_n)} \quad (4)$$

The output of the classification method is the probability of assigning a locality into the reference class based on natural heterogeneity, i.e. to which reference class the evaluated locality belongs; these probabilities are crucial for further evaluation of ecological state.
The presented methodology should provide more robust classification of sites into reference categories than parametric methods together with simple and straightforward interpretation of results. The methodology was tested on real datasets of 300 reference localities thorough the whole Czech Republic and provided the similar predictive power (neural networks; classification trees) or even better (discriminant analysis) than standard classification methods.

3.2 Suggested methodology of comparison of observed and predicted ecological state

As already mentioned, the comparison of predicted and observed ecological state is computed differentially for i) biological communities, ii) biotic indices and environmental parameters and iii) complex measure difference in multiple indices and/or environmental parameters between reference dataset and analyzed site.
For the comparison of observed and predicted community composition, the RIVPACS approach for prediction of biological community composition, which incorporate three parameters (S_{gz}: weight of reference group z according to its size; $P(x)_{gz}$: probability of classification of unknown site into reference group z; $pSpc(y)_{gz}$: probability of occurrence of species y in reference group z) in computation of probability of occurrence of all species sampled in the monitoring (5):

$$pSpc\,(y,x) = \frac{S_{g1} \times P(x)_{g1} \times pSpc(y)_{g1} + \ldots.. + S_{gn} \times P(x)_{gn} \times pSpc(y)_{gn}}{S_{g1} \times P(x)_{g1} + \ldots.. + S_{gn} \times P(x)_{gn}}$$

The next step of the computation is the cut-off of species with low probability of occurrence (the threshold value is according to experiences of RIVPACS and similar systems set to 50% probability of occurrence). The expected number of species is computed as (6):

$$S_{\max}(x) = round\left(\sum_{y} \mathrm{I}_{[\mathrm{p}Spc(\mathrm{y},x)>0.5]}\ \mathrm{p}Spc(\mathrm{y},\mathrm{x})\right)$$

The difference between observed and predicted state is computed

$$IP(x) = \frac{ns}{S_{\max}(x)} \quad (7)$$

where ns is the number of species occurred both in predicted and observed community.

Whereas the prediction of species abundances which are important for computation of biotic indices is connected with high level of noise and low precision (Guisan, Zimmermann [2000]), the prediction of expected values of biotic indices and also chemical parameters is based on another approach. The distribution of predicted values is based on frequency tables of values in reference data, where the probability of classification into reference group is used as the frequency weight of the given value. The predicted value is computed as median of predicted distribution and compared to observed value.

The EQI measure (ecological quality index according to RIVPACS methodology (Wright et al. [2000]) between measured value and median of predicted distribution is computed as the final value for evaluation of disturbance in this particular parameter

$$EQI_{DI(x)} = \frac{DI_{observed}}{DI_{\exp ected}}. \quad (8)$$

The intention of using such ratios is that it provides a means of standardizing the biotic indices, so that a particular value of the EQI ratio implies the same ecological quality for that index, no matter what the type of site.

It is largely because of the success of the RIVPACS type approach, using O/E ratios, and its acceptance as a robust tool for standardization within the scientific freshwater community, that the WFD prescribes the calculation and use of O/E ratios for reporting monitoring results.

The last measure of ecological state is the complex analysis of difference in observed and reference condition of set of biotic indices and/or environmental parameters. The computation has got two steps. In the first step it is the same as the computation of differences between unknown site and the reference groups, nevertheless these distances are computed on parameters which could be considered as metrics of environmental disturbance (biotic indices, environmental parameters). The differences converted to scale 0-100% are combined with probability of classification into reference groups (reference groups can have different natural levels of evaluated parameters and thus it is important which of reference categories is the most similar to unknown site) (9),

$$D(x) = \mathrm{D(x)}_{g1} \times \mathrm{P(x)}_{g1} + \mathrm{D(x)}_{g2} \times \mathrm{P(x)}_{g2} + ... + \mathrm{D(x)}_{gn} \times \mathrm{P(x)}_{gn}$$

where $\mathrm{D(x)}_{gz}$ is the distance of unknown site from reference group z and $\mathrm{P(x)}_{gz}$ is probability of classification of unknown site into reference group z. The final measure takes values 0-1.

3.3 Particular results of analysis and the final measure of ecological state

The above mentioned computations produce the set of partial measures of ecological state:

- Comparison of observed and expected composition of biological community (0-100%)
- EQI ratio for any continuous biotic index or environmental parameter (ratio observed vs. predicted ~ expected)
- Overall difference between observed and predicted ecological state based on set of environmental parameters and/or biotic indices

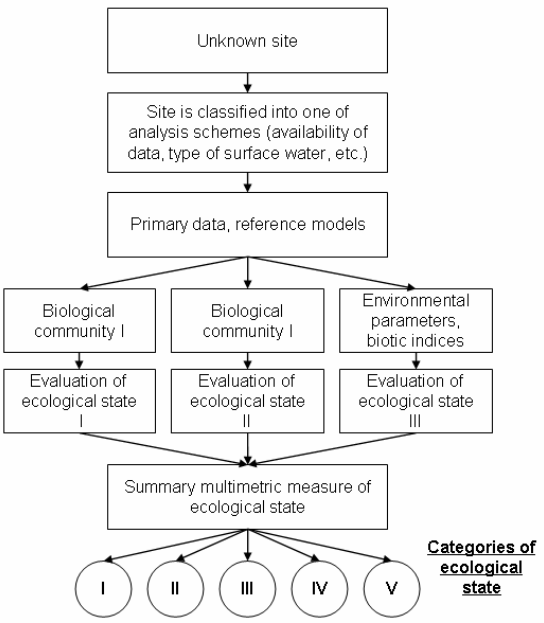

Figure 2 Final results of evaluation of ecological state

All of these measures represent different points of view on ecological state of localities and provide very useful information for scientists and experts on biological or chemical part of environmental disturbance. Additional information is gained from analysis of several types of biological communities (there should be two different biological communities in evaluation of ecological state according to WFD).

Nevertheless for standardized reporting and management of ecological state of surface waters there should be single metric with straightforward interpretation. According to demands of WFD the

particular results are combined into one metric which is split into five ordered categories from "natural" conditions to the most disturbed ecological state.

4. INFORMATIC SOLUTION OF METHODOLOGY OF EVALUATION OF ECOLOGICAL STATE

The presented methodology will be implemented into specialized application which will cooperate with central information system of the project through the standardized interface, compute the final measures of ecological state from the primary data of monitoring and biomonitoring and store them in the central data warehouse.

5. CONCLUSIONS

The presented methodology of evaluation of ecological state is useful for evaluation of surface waters with respect to general methodology of comparison of biological communities with reference state; moreover it can include also some methodological approaches from other types of evaluation of ecological state (biotic indices, multimetric approach) or modification according to demands of different biological communities.

The demands of experts for different types of biological communities, hydromorphology, chemical and physical parameters were included in the methodology development together with recent scientific knowledge in this field.

The methodology is universal for any type of date, i.e. any biological communities and respects the problems of data distribution and variability; the results of analysis are simple measures with straightforward interpretation for environmental reporting and management of both national institutions of the Czech Republic and EU environmental reports.

The problems of monitoring of surface waters, its analysis, reporting and interpretation reports have got the crucial role in the water management and planning according to WFD EU; for example in remedial processes, their control and also investment in this field.

Our current tasks according to WFD EU are i) the development of informatics solution of the methodology for cooperation with central information system, ii) the analysis of sufficient level of taxonomical determination of biological communities for routine biomonitoring (cost vs. gained information on ecological state) iii) to build up the reference model for biological communities utilizable in the routine biomonitoring.

6. REFERENCES

Barbour, M.T. and Yoder, C.O., 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, 2000.

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

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

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

Guisan, A., Zimmermann, N. E., Predictive habitat distribution models in ecology. *Ecological Modelling 135*, 147-186, 2000.

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

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

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

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

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