Study on the biodiversity - biotope factors’ relations

CORINA SIMIAN  
Department of Computer Science  
Faculty of Sciences  
University “Lucian Blaga”  
Str. I. Ratiu 5-7, 550012, Sibiu  
ROMANIA  
corinafirst@yahoo.com

VLADISLAV GEORGIEV  
University Computer Center  
St. Kliment Ohridski University  
5, J. Bourchier Ave., 1164  
Sofia  
BULGARIA  
v.georgiev@ucc.uni-sofia.bg

ANGELA CURTEAN-BĂNĂDUC  
Department of Ecology  
Faculty of Sciences  
University “Lucian Blaga”  
Str. I. Ratiu 5-7, 550012, Sibiu  
ROMANIA  
banaduc@yahoo.com

Abstract: - The study goal is to highlight the importance of the biodiversity quantified for the assessment of the homeostasis degree and of the support capacity of the lotic systems, also it underline the usefulness of the mathematical models which reveal the biodiversity - biotope factors relations, for the prognoses of these systems evolution. Factorial analysis (correlation and regression analysis) was used to elaborate the mathematical models which show the benthic macro invertebrates diversity variation related with the biotope factors variation.

Key-Words: - Factor analysis, Regression analysis, Biodiversity

1 Introduction and motivation

The relationship between man and nature has been accepted as something essential and irreplaceable. In the last years human pressure on natural ecosystems grew at an even faster rate. The main man’s pressures on biodiversity comes from the changes made in land usage, pollution, changes in the management of agricultural land, waters and forests, the over exploitation of natural resources and tourism. The obvious danger of the disappearance of a significant percentage of plants and animal species became increasingly clear in the last decade.

The aim of this paper is to highlight the importance of the biodiversity quantify for the assessment of the homeostasis degree and of the support capacity of the lotic systems, also it underlines the usefulness of the mathematical models which reveal the biodiversity - biotope factors’ relations, for the prognoses of these systems’ evolution.

The importance of mathematical models and statistic tools in ecology and biology is underlined in many other studies and books ([9]-[13]).

The selected study case is the Cibin River, a second order tributary of Danube localized in the middle of Romania. Cibin River has its sources in the glacial lakes of Cindrel Mountains (1920 m altitude), an 82 km length and a 2210 km² catchment basin. At least due to the biotope characteristics variation and to a variety of human impact types presence, this river is an interesting one concerning the ecological research. It should be added the fact that for this river case exist scientific data starting with 1851. This work is based on macro invertebrates quantitative and qualitative samples, sampled between 1999 - 2005, in 9 sampling stations, also biotope factors (slope, medium water flow, water physical-chemical characteristics) were evaluated. Factorial analysis (correlation and regression analysis) was used to elaborate the mathematical models which show the benthic macro invertebrates diversity variation related with the biotope factors variation. The variables were the biotope factors and the benthic macro invertebrates diversity, expressed through biodiversity indexes Margalef, Shannon - Wiener and equitability.

107 macro invertebrates’ species belonging to 67 genera, 39 families, 16 orders, 10 classes and 6 phylum were identified in Cibin River.

Comparing our data with the historical data it has been ascertained the fact that in Cibin River 19 macroinvertebrates species were disappeared and 13 species reduced their distribution along the river. The majority of these stenovalent species has low resistance at the environmental changes induced by the human impact. This dynamic in time point out the river natural habitats degradation, due to the hydro technical works presence impact (Gura Râului Dam, river channelization, marshes and floodplain drainage, cut of meanders, river banks reshaping and embanking, tributaries deviations, etc.) and due to the pollution.

This situation proved the necessity of study the pollution conditions and their temporal dynamics and the dependence of the biodiversity of these conditions.

The paper is organized as followed: in section 2 we present the biodiversity indexes. In section 3 we characterize the samples used in our analysis. Section 4 contains the main results of this article: the factorial analysis, the biodiversity - biotope factors relations and...
the regression equations used for prediction. Conclusions can be found in section 5.

2 Methods for biodiversity quantification and ecological characterization
2.1 Biodiversity quantification
Biodiversity is an indicator of ecological systems’ stability. The biodiversity of a community can be quantified using many indices.

2.1.1 Margalef index ($D_a$)

Let be $S$ the number of species from the community, $N$ the total number of community individuals. The Margalef index is:

$$D_a = \frac{(S - 1)}{\log N} \quad (2.1)$$

2.1.2 Menhinick index ($D_b$)

$$D_b = \frac{S}{\sqrt{N}} \quad (2.2)$$

2.1.3 Simpson index ($\lambda$)

Let be $s$ the number of species and $n_i$ the number of individuals from the species $i$. Simpson index is:

$$\lambda = \frac{\sum_{i=1}^{s} n_i (n_i - 1)}{N (N - 1)} \quad (2.3)$$

2.1.4 Shannon – Wiener index ($H_s$)

$$H_s = -\sum_{i=1}^{s} p_i \log\left(\frac{p_i}{\log 2}\right) \quad (2.4)$$

$$p_i = \frac{n_i}{N} \quad (2.5)$$

2.1.5 Equitability ($\varepsilon$)

$$\varepsilon = \frac{H_s}{I} \quad (2.6)$$

with

$$I = \frac{\log s}{\log 2} \quad (2.7)$$

2.2 Ecological characterization

The quantification of the ecological state can be made using synthetically indices. We use two synthetically indices: Belgian biotic index (IBB) and biotic index Hilsenhoff (IBH). Many details related to the computing methods of these indices could be found in [1].

3. Samples determination

Macro invertebrate’s quantitative and qualitative were sampled in the interval 1999 – 2005, in 9 sampling stations chosen taking into account the specific of the biotope and the presence of the specific pollution sources. Using the samples from the 9 stations have been computed biodiversity indices and biotic indices. Biotope factors: slope, medium water flow, water physical-chemical characteristics (T – water temperature, Susp. – suspension, DT – water total durity, Dt – water temporal durity, Rez. fix – constant reziduu, OD – oxygen, Defic. ox. – deficit of oxygen, CBO$_5$ – biochemical consum of oxygen in 5 days, CCO-Mn – consum of oxygen, chemical composition of water: Cl, SO$_4^{2-}$, Ca, Mg, HCO$_3$, NO$_2$–, NH$_4$+, NO$_3$–, PO$_4^{3-}$, Fe).

Fig. 1 Position of the sample stations on Cibin river

4 Factorial analysis

We can use a “dimensional” and a “temporal” analysis.

1. “Dimensional” analysis.

In this case we have many samples corresponding to all measured characteristics. Every sample contains 9 values, corresponding to the 9 stations. The given values are computed as mean of the observations on a station in the 3 years period of observation. This kind of analysis is indicated for establishing relations between biodiversity and biotope characteristics.

2. We can perform 2 type of temporal analysis.

a) Data are given as mean values for 3 years in every month between March and November. Analysis is made separately for the 9 stations. A sample contains 8 data. This type of analysis is recommended for establishing
relations between biotope characteristics and a certain period of year.

b) Samples contains 24 data (3years*8months). Analysis is made separately for the 9 stations. This type of analysis is recommended if we want to make prediction in time of some of the biotope characteristics.

For the aim of our article, the first type of analysis is indicated. For every station we have the values of the biodiversity indexes and the biotope characteristics. Every sample has 9 values. First we must verify the type of sample distribution. Let denote by $X_i$, the variables observed in our study. The 9 stations are not random chosen. However we can test if the data in vector $X_i$ obtained from these 9 stations is a random sample from a normal distribution with mean and variance estimated from $X_i$. For this we can use a chi-square goodness-of-fit test. Chi-square goodness-of-fit tests the default null hypothesis that the data in vector $X$ is a random sample from a normal distribution with mean and variance estimated from $X$, against the alternative that the data are not normally distributed with the estimated mean and variance. The result is 1 if the null hypothesis can be rejected at the 5% significance level. The result is 0 if the null hypothesis cannot be rejected at the 5% significance level.

A $p$-value of the test can also be calculated. The $p$-value is the probability, under assumption of the null hypothesis, of observing the given statistic or one more extreme.

If the distribution is not normal, it is possible to test if it belongs to a parametrical family of distributions like Pearson and Johnson parametric family of distributions. Each of two systems is a flexible parametric family of distributions that includes a wide range of distribution shapes, and it is often possible to find a distribution within one of these two systems that provides a good match to your data. Each member of the Pearson’s and Johnson’s systems is defined by the following four parameters: Mean, Standard deviation, Skewness and Kurtosis. To test if the data in vector $X_i$ is a random sample from a Pearson distribution, a Jarque-Bera, test of the null hypothesis that the sample in vector $X_i$ comes from a normal distribution with unknown mean and variance, against the alternative that it does not come from a normal distribution, might be used. The Jarque-Bera test is a two-sided goodness-of-fit test suitable when a fully-specified null distribution is unknown and its parameters must be estimated.

ANOVA type tests perform hypothesis tests in the context of linear modeling. Before to fit a function to a model that relationships between two measured quantities, it is necessary to determine if a relationship exists between these quantities.

In order to do this, covariance and correlation coefficient must be computed. Correlation measures how variables or rank orders are related. Correlation quantifies the strength of a linear relationship between two variables. We are interested in the degree of relationship between variables; therefore we might calculate correlation coefficients. These coefficients are derived from covariance. The correlation coefficient $r_{X,Y}$ between two random variables $X$ and $Y$ with expected values $\mu_X$ and $\mu_Y$ and standard deviations $\sigma_X$ and $\sigma_Y$ is their covariance normalized by their standard deviations. The correlation coefficients range from -1 to 1. Values close to or equal to 0 suggest there is no linear relationship between the data columns. Values close to 1 suggest that there is a positive linear relationship between the data columns. Values close to -1 suggest that one column of data has a negative linear relationship to another column of data (anticorrelation).

After seeing the existence of correlation between two variables, regression analysis is required in order to find the dependence model. Linear regression models represent the relationship between a continuous response $y$ and a continuous or categorical predictor $x$ in the form:

$$Y=b_0f_0(x)+\ldots+b_nf_n(x)+\varepsilon$$

The response is modeled as a linear combination of (not necessarily linear) functions of the predictor, plus a random error $\varepsilon$. Errors $\varepsilon$ are assumed to be uncorrelated and distributed with mean 0 and constant (but unknown) variance.

We can use Pearson’s correlation coefficient or partial correlation coefficients that describe the linear relationship between two variables while controlling for the effects of one or more additional variables. Two variables can be perfectly related, but if the relationship is not linear, a correlation coefficient is not an appropriate statistic for measuring their association. The bivariate correlations compute the pairwise associations for a set of variables. It is useful for determining the strength and direction of the association between two scale or ordinal variables.

Statistical Data Analysis can be obtained using various software programs: SPSS, MATLAB or free statistics packages.

4. Main results

The aim of this section is to present the experimental results of statistical data analysis for the data presented in [1]. The bivariate correlation between biodiversity indices and biotope characteristics is given.
In this section, we make the analysis and give a rendering of the results in the case of biodiversity index Margalev. Similar analysis was made for all indices presented in section 2 and will be presented in an extended version of this paper. The complete analysis allows us to make a complete characterization of the ecological state of Cibin river.

Descriptive statistics for Margalev index, $D_a$, for our sample is:

\[ \bar{x} = 7.80444, \, \text{Std. Deviation} = 3.502251, \, N = 9 \]

The values for means, standard deviation, Pearson bivariate coefficient, p-value, covariance, significance level of correlation (Sig. 2-tailed), are presented in table 1. Brief interpretation of the correlation significance is also given.

![Table 1](image)

<table>
<thead>
<tr>
<th>Biotope character</th>
<th>Pearson correlation</th>
<th>Covar.</th>
<th>Sig. 2-tailed</th>
<th>Sig. level</th>
<th>Mean</th>
<th>Std. Dev.</th>
</tr>
</thead>
<tbody>
<tr>
<td>pH</td>
<td>-.767</td>
<td>.572</td>
<td>.016</td>
<td>.01</td>
<td>7.10</td>
<td>.213</td>
</tr>
<tr>
<td>Susp. (mg/l)</td>
<td>-.947</td>
<td>-.9847</td>
<td>.000</td>
<td>.01</td>
<td>99.65</td>
<td>29.682</td>
</tr>
<tr>
<td>Dt (°C)</td>
<td>-.831</td>
<td>.7141</td>
<td>.006</td>
<td>.01</td>
<td>2.98</td>
<td>2.454</td>
</tr>
<tr>
<td>Rez. fix. (mg/l)</td>
<td>-.784</td>
<td>-.3756</td>
<td>.012</td>
<td>.05</td>
<td>225.11</td>
<td>36.870</td>
</tr>
<tr>
<td>CBO5 (mg/l)</td>
<td>-.788</td>
<td>-.926</td>
<td>.012</td>
<td>.05</td>
<td>5.00</td>
<td>3.360</td>
</tr>
<tr>
<td>Defic. o (%)</td>
<td>-.813</td>
<td>.40900</td>
<td>.008</td>
<td>.01</td>
<td>25.34</td>
<td>14.362</td>
</tr>
<tr>
<td>SO42 (mg/l)</td>
<td>-.783</td>
<td>-.3948</td>
<td>.012</td>
<td>.05</td>
<td>16.62</td>
<td>14.395</td>
</tr>
<tr>
<td>Ca (mg/l)</td>
<td>-.869</td>
<td>.33606</td>
<td>.002</td>
<td>.01</td>
<td>18.50</td>
<td>11.044</td>
</tr>
<tr>
<td>HCO3 (mg/l)</td>
<td>-.838</td>
<td>.16078</td>
<td>.005</td>
<td>.01</td>
<td>67.66</td>
<td>54.772</td>
</tr>
<tr>
<td>NH4+ (mg/l)</td>
<td>-.745</td>
<td>-.11279</td>
<td>.021</td>
<td>.05</td>
<td>333.64</td>
<td>32.145</td>
</tr>
</tbody>
</table>

Table 1. Correlations between Margalef index and biotope charact.

The experimental results prove the existence of a anticorrelation between Margalef index and pH, Dt, rez fix, OD, CBO5, def ox, SO42-, Ca, HCO3, NH4+, suspensions.

The most important anticorrelation is between Margalef index and suspensions.

That means that if the pollution increases (the amount of rez fix, OD, CBO5, def ox, SO42-, Ca, HCO3, NH4+, suspensions increases) the Margalef index decreases, that is the biodiversity decreases in a linear way.

The existence of the correlations implies that has sense to make a regression analysis for prediction the Margalef index as dependent variable from the independent variables, situated in column 1 of the table 1.

The coefficients of the regression line for the pH independent variable are given in table 2.

We perform ANOVA tests for characterization of the acceptability of the model from a statistical perspective. The Regression row displays information about the variation accounted for by the model. The Residual row displays information about the variation that is not accounted for by the model. The regression and residual sums of squares are approximately equal, which indicates that about half of the variation of Margalef index is explained by the model.

The significance value of the F statistic, 0.016 is less than 0.05, which means that the variation explained by the model is not due to chance. While the ANOVA table is a useful test of the model's ability to explain any variation in the dependent variable, it does not directly address the strength of that relationship. For characterization of the strength of that relationship we use the coefficients R and $R^2$.

R, the multiple correlation coefficient, is the linear correlation between the observed and model-predicted values of the dependent variable. Large value for R indicates a strong relationship.

$R^2$, the coefficient of determination, is the squared value of the multiple correlation coefficient.

In our case $R^2 = 0.58$.

It shows that about 58% of the variation in time is explained by the model.

![Table 2](image)

<table>
<thead>
<tr>
<th>Coefficients $^a$</th>
<th>Unstandard. Coeff.</th>
<th>Standard. Coeff.</th>
<th>t</th>
<th>Sig.</th>
</tr>
</thead>
<tbody>
<tr>
<td>B</td>
<td>Std. Error Beta</td>
<td>t</td>
<td>Sig.</td>
<td></td>
</tr>
<tr>
<td>97.321</td>
<td>28.311</td>
<td>3.438</td>
<td>.011</td>
<td></td>
</tr>
<tr>
<td>12.610</td>
<td>3.987</td>
<td>-.767</td>
<td>-3.163</td>
<td>.016</td>
</tr>
</tbody>
</table>

Table 2 - Regression analysis of Margalef index.
Table 3 – ANOVA tests

Table 4 summarize the results of the regression analysis for Margalef index as dependent variable. The labels of the columns of Table 4 are: RL (regression Line), ASSR1(ANOVA Sum of Squares for Regression), ASSR2(ANOVA Sum of Squares for Residual), ASig(ANOVA Sig. 2 tails), R2 (coefficient of determination)

Table 4 – Regression analysis for Margalef index

From table 4 we observe that the stronger dependence is between Margalef index and suspension. Also the linear model fit better for the dependence between Margalef index and suspensions. Almost 90% of the variation of Margalef index due to suspensions is explained by model.

Ecological characterization of Cibin river can be made using a graphical representation of physical-chemical characteristics on the 9 stations.

Table 5 – Physical-chemical characteristics

It is to see that stations 8 and 9 have all the characteristics greater or equal to the mean value. More, many of the values obtained for these stations are situated near to maximum value. There is an increasing trend of almost all the characteristics from station 1 to station 8. That proves that the pollution is more accentuated in a neighborhood of station 8.

5 Conclusions and further directions of study

Achieving the correlation analysis for the considered data set, it was found that significant statistical correlations exist between the benthic macroinvertebrates diversity and the biotope parameters from table In the paper were reproduced the regression equations. Conferring values to the independent variables, in the regression equations, the values of the dependent variables can be determined, with a known certain level of error. In this way, possessing a data set which describes the habitat quality, it is possible to estimate the biodiversity like indexes values, like an expression of the analyzed river homeostasis degree. The obtained mathematical models, allow prognosis with a certain level of probability, of the benthic macroinvertebrates diversity variation in the condition of some biotope parameters modification, being useful to establish the river sustainable management program. The descriptive statistics for the 9 stations show that the highest macroinvertebrates diversity is in the river mountainous sector, there where the human impact is insignificant. Downstream the diversity is decreasing in parallel with the human impact increasing, reaching a minimum downstream the confluence of the river with the Sibiu locality waste water plant water. Our, including a complete data and regression analysis for biodiversity indices, synthetically indices and biotope characteristics will be presented in an extended article.
References: