Environment and Ecology Research 4(1): 21-29, 2016
DOI: 10.13189/eer.2016.040104 http://www.hrpub.org

Multivariate Statistical Methods in Researching Biocultural Diversity

Joško Sindik, Jelena Šarac*

Institute for Anthropological Research, Croatia

Abstract Nonlinear multivariate statistical methods have proven to be useful tools in research issues dealing with biocultural diversity. Namely, these methods have less restrictions in their use, as compared with compatible linear methods. This research is the example of using some of these methods. The three indices of biocultural diversity by the variables of biological and cultural diversity have been predicted, based on population size (POP), areal size (AREA) and overall biological and cultural richness (RICH). Then, we have determined: clusters in which different countries can be grouped based on biocultural diversity indices (POP, AREA, RICH), the latent dimensions of the biocultural diversity in the space of biocultural diversity indices (POP, AREA, RICH) and finally, the association between the indices of biodiversity and cultural diversity (POP, AREA, RICH). General conclusion is that nonlinear multivariate methods (together with cluster analysis), in spite of their robustness, can provide useful information to the researchers, on the issues they are interested in (in this example, about biocultural diversity). Of course, these methods are more convenient than linear methods only in the context of absence of clear linear relationships between the variables of interest, while the overfitting problem couldn't be avoided in both cases.

Keywords Association, Biodiversity, Biocultural and Cultural Diversity, Nonlinear Methods

1. Introduction

Nonlinear multivariate analysis (NMA) has proven to be very useful in research issues dealing with biocultural diversity, because these methods are not strictly dependent about types of data or distributions and can be applied on small sized samples as well. In this article, the examples of using three of NMA methods will be presented, along with one linear multivariate method (K-means clustering), following the example of the indices that describe biocultural diversity, introduced by Loh and Harmon [15]. To improve the understanding of the results obtained, the issue of biocultural diversity will be explained.

1.1. An Overview about the Theoretic Framework for This Example: Biocultural Diversity

In this study, nonlinear multivariate methods are applied on the indices of biocultural diversity, defined by Loh and Harmon [15]. Biocultural diversity can be simply determined as the sum of all the world's differences and the fundamental expression of the variety upon which all life is founded [16]. It includes all levels of biological diversity (from genes, over populations and species to ecosystems) and cultural diversity in all its manifestations (from individual ideas to entire cultures), together with the interactions among all of them [15]. The paper published by Loh and Harmon [15], was one of the first attempts to quantify the level of global biocultural diversity by introducing the IBCD (index of biocultural diversity), a simple proxy to indicate the status of a complex phenomenon. Implications of such a global index are both theoretical and practical: IBCD provides a global context as a starting-point for in-depth analyses and gives a framework for strategic investments in biocultural diversity conservation [15]. IBCD is a sum of three components: aerial (AREA), population (POP) and biocultural diversity richness (RICH). For defining cultural diversity score (CD) are used five indicators: number of languages, religions and ethnicities. Biological diversity score (BD) comprises the number of mammals/birds and plants, giving equal weight to both diversities. A country's overall biocultural diversity score is calculated as the average of these two scores (CD and BD). For example, the index identified three areas of exceptional biocultural diversity: the Amazon Basin, Central Africa and Indomalaysia/Melanesia [15].

1.2. Multivariate Techniques and Nonlinear Analyses

Since the patterns of diversity are undoubtedly shaped by multiple factors, their research must move beyond single-factor correlative studies, and multicausal approaches should be pursued through multivariate statistical methods.

In general, multivariate quantitative methods and techniques strengthen the indicative and predictive value of factors or variables and allow for cross-cultural comparison of data between and among different groups and communities [24]. Hoeft et al. (1999) classified statistical applications into two broad categories: 1) sets of data where the measurements are taken only to one attribute or response variable allowing for univariate analysis techniques; and 2) sets of data where the measurements are taken simultaneously on more than one variable allowing for multivariate analysis techniques. As the multivariate analysis technique is generally used to make large data sets accessible, recognize structures, and explain and predict patterns among variables, Johnson and Wichern [13] have identified five basic applications of these techniques: 1) data reduction or structural simplification; 2) sorting and grouping; 3) explaining relationships among variables; 4) prediction, and 5) testing of hypotheses. The selection of the most appropriate methodology to achieve maximum results depends on both the objectives of the research and the type of study.

However, most multivariate models tend to focus on linear relationships between variables. Nonlinear analysis increases the value of the relevant variables, like in the case of increasing the value of biocultural conservation factors that include the 'knowledge-practice-belief' complex. Although a linear model could be useful method in research about biodiversity, e.g. for biodiversity-ecosystem functioning experiments [2], the most of the relationships between biodiversity variables are nonlinear [3, 12, 19, 29]. For example, multitrophic interactions are expected to make biodiversity–ecosystem functioning relationships more complex (multivariate) and nonlinear, while the monotonic (linear) changes could be predicted for simplified systems with a single trophic level [27]. These nonlinear multivariate techniques can be regarded as two-step techniques. First step is the nonlinear transformation of variables into optimally scaled variables. Second step is application of the multivariate analysis to the optimally scaled variables. In multivariate analyses, depending of a purpose of certain method, relevant factors have been grouped and divided into blocks of variables [24].

The technique named optimal scaling (OS) describes the algorithm of converting nominal and ordinal variables into interval variables [23]. An alternating least-squares OS algorithm can be divided into two major stages. First stage is estimating the parameters of the linear model, which are used to create the predicted values (target) for each variable that can be transformed, by minimizing squared error [23]. The definition of the target depends on many factors, such as type of the variables, algorithm, etc. The second major stage, i.e. optimal scaling, could be defined as a possibly constrained, least-squares regression problem. This OS phase finds the vector which is a linear combination of the columns of this matrix that is closest to the target (in terms of minimum squared error), respecting all the constraints defined by the transformation family. Optimal scaling methods are independent of the data analysis method that generated the target [23].

1.3. Methods Used in This Example

In this paper, the following methods have been used in order to assess the global biocultural diversity: Categorical Regression (CATREG), Categorical Principal Components Analysis (CATPCA), Nonlinear Canonical Correlation Analysis (OVERALS) and Cluster Analysis (CA). CA or clustering is the task of grouping a set of objects in such a way that objects in one cluster are more similar to each other than to those in other clusters. K-means clustering, the method of cluster analysis used in this study, is a method of vector quantification, where Euclidean distance is used as a metric and variance is used as a measure of cluster scatter. It is rather easy to implement and apply even on large data sets. Then, we are presenting three nonlinear multivariate methods [20]. CATREG is used instead of standard linear (multiple) regression analysis. Standard linear regression analysis involves minimizing the sum of squared differences between a response (dependent) variable and a weighted combination of predictor (independent) variables. Variables are typically quantitative, with (nominal) categorical data recoded to binary or contrast variables. As a result, categorical variables serve to separate groups of cases, and the technique estimates separate sets of parameters for each group. The estimated coefficients reflect how changes in the predictors affect the response. Prediction of the response is possible for any combination of predictor values. CATREG extends the standard approach by simultaneously scaling nominal, ordinal, and numerical variables. The procedure quantifies categorical variables so that the quantifications reflect characteristics of the original categories. The procedure treats quantified categorical variables in the same way as numerical variables. Using nonlinear transformations allow variables to be analyzed at a variety of levels to find the best-fitting model [20]. CATPCA simultaneously quantifies categorical variables while reducing the dimensionality of the data. The goal of principal components analysis (PCA) is to reduce an original set of variables into a smaller set of uncorrelated components that represent most of the information found in the original variables [14]. The technique is most useful when a large number of variables prohibit effective interpretation of the relationships between objects (subjects and units). By reducing the dimensionality, the user interprets a few components rather than a large number of variables. Standard PCA analysis assumes linear relationships between numeric variables. On the other hand, the optimal scaling approach allows variables to be scaled at different levels. Categorical variables are optimally quantified in the specified dimensionality. As a result, nonlinear relationships between variables can be modeled [14]. OVERALS is aimed to determine how similar sets of categorical variables are to one another. Standard canonical correlation analysis is an extension of multiple regression

analysis, where the second set does not contain a single response variable but instead contain multiple response variables. The goal is to explain as much as possible of the variance in the relationships among two sets of numerical variables in a low dimensional space. Initially, the variables in each set are linearly combined such that the linear combinations have a maximal correlation [20]. Given these combinations, subsequent linear combinations are determined that are uncorrelated with the previous combinations and that have the largest correlation possible. The optimal scaling approach expands the standard analysis in three crucial ways. First, it allows more than two sets of variables. Second, variables can be scaled as either nominal, ordinal, or numerical. As a result, nonlinear relationships between variables can be analyzed. Finally, instead of maximizing correlations between the variable sets, the sets are compared to an unknown compromise set that is defined by the object scores [18].

Thus, the general goal of this research is to elaborate the application of the abovementioned nonlinear multivariate analyses (and a single linear one – cluster analysis) in the example of global biocultural diversity, re-analyzing the findings obtained by Loh and Harmon [15]. Specifically, the goals are several: first, to predict the index of global biocultural diversity by the variables of biological and cultural diversity, based on population size (POP), areal size (AREA) and overall biological and cultural richness (RICH). Second, to determine the clusters in which different countries can be grouped based on biocultural diversity indices (POP, AREA, RICH). Third, to determine the latent dimensions of the biocultural diversity, in the space of biocultural diversity indices (POP, AREA, RICH). Fourth, to determine the association between the indices of biodiversity and cultural diversity (POP, AREA, RICH).

2. Materials and Methods

2.1. Materials and Variables

In this article we have used data and indices defined in the article by Loh and Harmon [15]. The IBCD gives equal weight to cultural and biological diversity, so a country's overall biocultural diversity score is calculated as the average of its cultural diversity score (CD) and its biological diversity score (BD) [15].

$$\text{IBCD} = \frac{\text{CD} + \text{BD}}{2}$$

In measuring a country's CD, equal weight is given to linguistic, religious and ethnic diversity. Therefore CD is calculated as the average of a country's language diversity (LD), religion diversity (RD), and ethnic group diversity (ED) [15].

$$\text{CD} = \frac{\text{LD} + \text{RD} + \text{ED}}{3}$$

In measuring biodiversity (BD), equal weight is given to animal species diversity (using birds and mammals as a proxy for all animal species; marine mammals are excluded for the analysis) and plant species diversity. Therefore, BD is calculated as the average of a country's bird and mammal species diversity (MD), and plant species diversity (PD) [15].

$$\text{BD} = \frac{\text{MD} + \text{PD}}{2}$$

Each indicator is given an equal weighting as this is the simplest way of calculating the index. As an aggregated index, the IBCD could be calculated using different weightings, to give greater or lesser importance to any of the five component indicators. To derive country scores for each of the five component indicators, Loh and Harmon [15] compared each country's richness value with the global value. For example, for language diversity, LD is calculated as the log of the number of languages spoken in a country (Li) divided by the log of the number of languages spoken worldwide (Lworld) [15].

$$\text{LD} = \frac{\log L_i}{\log L_{\text{world}}}$$

The calculation was repeated for the other four indicators to derive BCD-RICH [15]. Data sources were as follows: languages [8], religions [1], ethnic groups [1], bird/mammal species [9], plant species [9], country area [26]; countries smaller than 1000 km2 are excluded), and country population ([4]; countries with a population of less than 10,000 are excluded) [15]. To compensate for the fact that large countries tend to have a greater biological and cultural diversity than small ones simply because of their greater area (or greater population), Loh and Harmon [15] calculated two additional diversity values for each country by adjusting first for land area (BCD-AREA) and second for population size (BCD-POP). They measured how much more or less diverse a country is in comparison with an expected value based on its area or population alone. The method used is a modified version of that used by Groombridge and Jenkins [9]. The expected diversity was calculated using the standard formula for the species-area relationship log S = c + z log A where S = number of species, A = area, and c and z are constants derived from observation. Because the distributions of the five indicators against land area and populations size are similar, Loh and Harmon [15] applied the same formula to indicators of cultural diversity. Hence, for BCD-AREA expected log Ni = c + z log Ai where Ni = number of languages, religions,

ethnic groups, or species in country i, and Ai = area of country i. The same formula was used for BCD-POP, except that Pi (population of country i) replaces Ai. To find the values of constants c and z for each of the indicators, we scatter-plotted Ni (where Ni = number of languages, religions, ethnic groups, or species in country i) against log Ai for all countries, and drew the best-fit straight line through the points. To calculate the deviation of each country from its expected value, Loh and Harmon [15] subtracted the expected log Ni value from the observed log Ni value. The index is calibrated such that the world, or maximum, value is set equal to 1.0, the minimum value is set equal to zero and the average or typical value is 0.5 (meaning no more or less diverse than expected given a country's area or population).

2.2. Data Sources

The sources of the data were: Table 4 IBCD-RICH 20 highest ranked countries, pp 234; Table 5 IBCD-AREA 20 highest ranked countries, pp 235; Table 6 IBCD-POP 20 highest ranked countries, pp 236 (all from [15]). In all analyses, statistical software package SPSS 20.0. has been used.

3. Results

In Figure 1, Means and standard deviations for certain biodiversity indices – AREA are presented (similar situation is with RICH and POP indices, hence both for descriptive values and for correlations, only AREA indices are presented). The test for related samples (Wilcoxon) revealed that for the same countries, biological (Bird & mammal and Plant diversity) indices and cultural (Language, Religion and Ethnic group diversity) indices are statistically significantly different ($p<0.01$). In other words, countries varied depending of the type of biocultural diversity.

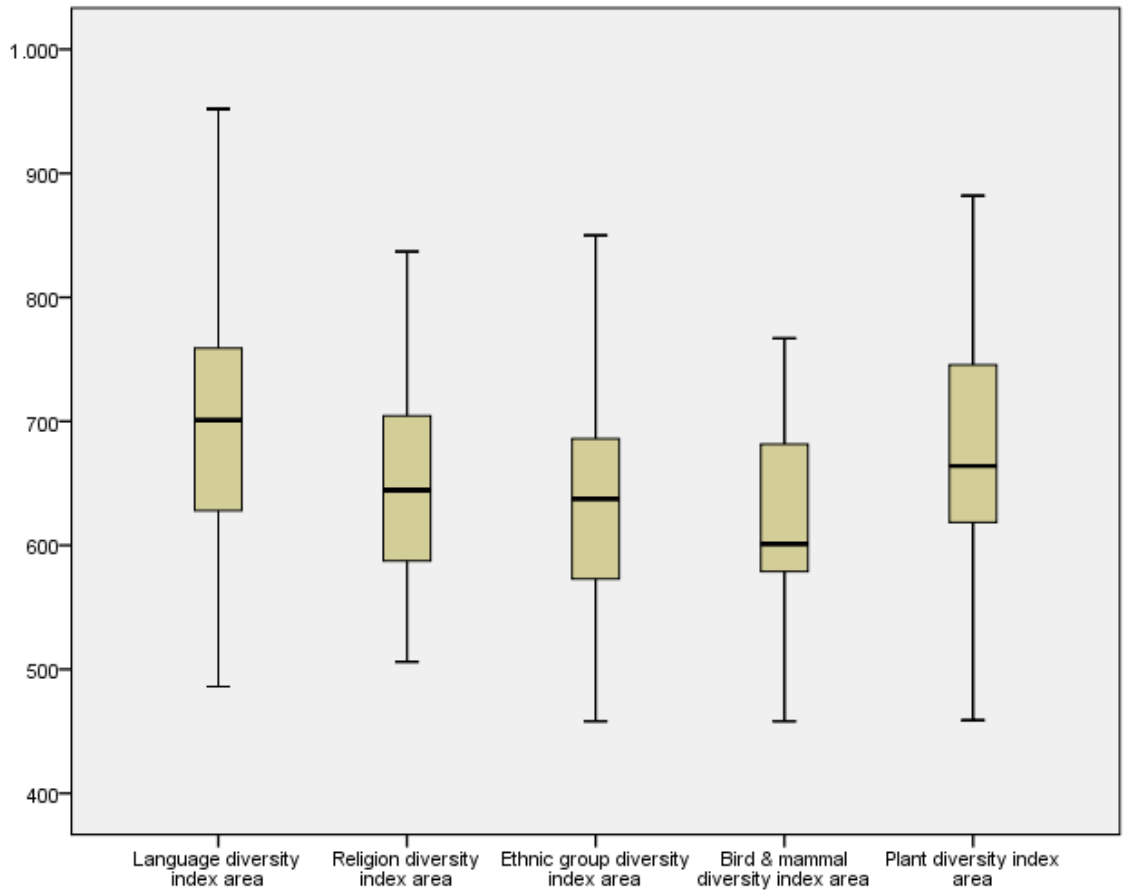

Figure 1. Means and standard deviations for certain biodiversity indices – AREA

In Table 1, from 28 correlations, 20 were mostly high-sized, statistically significant and positive. Higher correlation sizes are found among three cultural diversity indices and among two biodiversity indices, than when comparing cross-correlations (between cultural and biodiversity indices). The least number of statistically significant correlations are found for Bird & mammal diversity index, as well as Biodiversity index (in both cases, very low correlations with cultural diversity indices). Moreover, Index of biocultural diversity highly and statistically significantly correlates with all individual and complex indices.

Table 1. Spearman correlations between AREA biological and cultural diversity indices, with belonging complex biodiversity indices

	Language diversity	Religion diversity	Ethnic group diversity	**Cultural diversity**	Bird & mammal diversity	Plant diversity	**Biodiversity**	**Biocultural diversity**
Language diversity	1	.903**	.974**	.981**	.055	.477*	.008	.703**
Religion diversity		1	.939**	.965**	.233	.628**	.161	.789**
Ethnic group diversity			1	.992**	.200	.624**	.146	.799**
Cultural diversity				1	.163	.586**	.105	.778**
Bird & mammal diversity					1	.749**	.930**	.703**
Plant diversity						1	.737**	.883**
Biodiversity							1	.706**
Biocultural diversity								1

Legend: the names of more complex indices are bolded
**correlation statistically significant at p<0.01; *correlation statistically significant at p<0.05

Table 2. Categorical regression (optimal scaling): prediction of the variables population size, area and index of biocultural diversity IBCD-RICH, with belonging biodiversity indices – POP, AREA and RICH

POP 2000	Standardized Coefficients		F	R	R^2
	Beta	Bootstrap (1000) Estimate of Std. Error	(df=1)	(F=1.440, p>0.20; df=5,14)	
Language diversity index	-.694	.900	.595		
Religion diversity index	1.259	1.111	1.284		
Ethnic group diversity	-.542	1.406	.149	0.583	0.340
Bird & mammal diversity	-.498	.265	3.537		
Plant diversity	.161	.262	.379		
AREA	Standardized Coefficients		F	R	R^2
	Beta	Bootstrap (1000) Estimate of Std. Error	(df=1)	(F=1.214, p>0.20; df=12,7)	
Language diversity index	-1.436	1.209	1.409		
Religion diversity index	.632	1.355	.218		
Ethnic group diversity	1.189	1.651	.519	0.822	0.675
Bird & mammal diversity	-.539	.968	.310		
Plant diversity	1.177	.977	1.449		
RICH	Standardized Coefficients		F	R	R^2
	Beta	Bootstrap (1000) Estimate of Std. Error	(df=1)	(F=33.828, p<0.01; df=5,14)	
Language diversity index	.381	.318	1.437		
Religion diversity index	.117	.300	.151		
Ethnic group diversity	.330	.347	.905	**0.961****	**0.896**
Bird & mammal diversity	.213	.108	3.873		
Plant diversity	.332	.126	**6.916***		
POP 2000	Standardized Coefficients		F	R	R^2
	Beta	Bootstrap (1000) Estimate of Std. Error	(df=1)	(F=1.440, p>0.20; df=5,14)	
Language diversity index	-.694	.900	.595		
Religion diversity index	1.259	1.111	1.284		
Ethnic group diversity	-.542	1.406	.149	0.583	0.340
Bird & mammal diversity	-.498	.265	3.537		
Plant diversity	.161	.262	.379		
AREA	Standardized Coefficients		F	R	R^2
	Beta	Bootstrap (1000) Estimate of Std. Error	(df=1)	(F=1.214, p>0.20; df=12,7)	
Language diversity index	-1.436	1.209	1.409		
Religion diversity index	.632	1.355	.218		
Ethnic group diversity	1.189	1.651	.519	0.822	0.675
Bird & mammal diversity	-.539	.968	.310		
Plant diversity	1.177	.977	1.449		
RICH	Standardized Coefficients		F	R	R^2
	Beta	Bootstrap (1000) Estimate of Std. Error	(df=1)	(F=33.828, p<0.01; df=5,14)	
Language diversity index	.381	.318	1.437		
Religion diversity index	.117	.300	.151		
Ethnic group diversity	.330	.347	.905	**0.961****	**0.896**
Bird & mammal diversity	.213	.108	3.873		
Plant diversity	.332	.126	**6.916***		

Legend: POP=size of human population in year 2000 per country; AREA=country area (in km2); RICH=index of biocultural diversity IBCD-RICH

In Table 2, we have predicted the variables population size, area and total number of (biological and cultural) diversities in countries. Only one of the predictors was statistically significant (Plant diversity), only in case of one significant multiple regression coefficient (for index of biocultural diversity IBCD-RICH).

Table 3. Cluster analysis (K-means clustering): countries grouped by Index of biocultural diversity – POP, AREA and RICH by biodiversity indices

Index of biocultural diversity - POP	Cluster		
	1	2	3
Language diversity index	**720.00**	668.79	677.50
Religion diversity index	**741.00**	661.86	641.75
Ethnic group diversity	**679.00**	623.36	603.50
Bird & mammal diversity	641.50	**795.00**	714.50
Plant diversity	756.50	757.36	**781.00**
Number of cases in clusters	2	14	4
Index of biocultural diversity - AREA	**Cluster**		
	1	2	3
Language diversity index	703.36	**868.00**	558.60
Religion diversity index	645.27	**787.00**	553.00
Ethnic group diversity	637.09	**777.00**	516.40
Bird & mammal diversity	573.55	621.25	**728.40**
Plant diversity	655.73	618.25	**776.80**
Number of cases in clusters	11	4	5
Index of biocultural diversity - RICH	**Cluster**		
	1	2	3
Language diversity index	582.73	**725.50**	585.80
Religion diversity index	522.09	**667.00**	486.40
Ethnic group diversity	557.00	**679.75**	548.40
Bird & mammal diversity	711.64	740.25	**775.80**
Plant diversity	737.45	762.50	**837.80**
Number of cases in clusters	11	4	5

Note: the highest mean values of indices in one line are bolded

Results of the cluster analysis are presented in Table 3, with belonging number of cases in each cluster. According to IBCD-POP, only Indonesia and Brazil are grouped in the first cluster, as countries with the most numerous populations. Four countries are grouped in the third cluster: Colombia, Malaysia, Peru and Australia. All other 14 of 20 highest ranked countries (according to IBCD-POP) are grouped in the second cluster (Papua New Guinea, French Guiana, Suriname, Cameroon, Brunei, etc.). First cluster is the most distant from the other two clusters.

According to IBCD-AREA, 11 of 20 the highest ranked countries are grouped in the first cluster: Malaysia, India, Nepal, Brazil, Mexico, Philippines, Viet Nam, Tanzania, Laos, Congo, Solomon Islands. Four countries are grouped in the second cluster: Indonesia, Papua New Guinea, Cameroon and Nigeria and five constitute the third cluster: Peru, Ecuador, Colombia, Brunei and Panama. Second and third clusters are the most distant from each other.

According to IBCD-RICH, 11 of 20 the highest ranked countries are grouped in the first cluster: United States of America, Cameroon, Congo, Australia, Malaysia, Tanzania, Russia, Myanmar, Sudan, Philippines, and Ethiopia. Four countries are grouped in the second cluster: Indonesia, Papua New Guinea, India and Nigeria and the third cluster constitutes of five countries: China, Colombia, Mexico, Peru and Brazil. Second and third clusters are the most distant from each other, while first is closer to the third cluster, than second.

In Table 4, we have showed the results of CATPCA on biodiversity indices. In all three cases, CATPCA revealed the same finding: two-component solution appeared as the most convenient to explain the space of biodiversity indices. First component always showed very high reliability and high saturation with belonging variables and it was named *Cultural Diversity*. Second component always showed very low (not satisfying) reliability but high saturation with belonging variables, too: it was named *Biological Diversity*.

Table 4. Principal Components Analysis for Categorical Data (CATPCA) on biodiversity indices – POP, AREA and RICH

POP components	**cultural**	**biological**	**Total**
Language diversity index	**.978**	.185	
Religion diversity index	**.988**	.117	
Ethnic group diversity	**.984**	.163	
Bird & mammal diversity	-.432	**.762**	
Plant diversity	-.139	**.889**	
Cronbach's Alpha	.848	.384	.975
Eigenvalue	3.106	1.445	4.551
% of Variance	62.125	28.904	91.028
AREA components	**cultural**	**biological**	**Total**
Language diversity index	**.976**	.213	
Religion diversity index	**.967**	.239	
Ethnic group diversity	**.972**	.227	
Bird & mammal diversity	-.399	**.827**	
Plant diversity	-.396	**.829**	
Cronbach's Alpha	.853	.430	.983
Eigenvalue	3.149	1.526	4.674
% of Variance	62.971	30.510	93.481
RICH components	**cultural**	**biological**	**Total**
Language diversity index	**.939**	.112	
Religion diversity index	**.907**	-.086	
Ethnic group diversity	**.968**	-.002	
Bird & mammal diversity	-.004	**.884**	
Plant diversity	-.025	**.900**	
Cronbach's Alpha	.777	.474	.956
Eigenvalue	2.642	1.611	4.253
% of Variance	52.836	32.227	85.062

Note: the highest mean values of indices in one line are bolded

In Table 5, the application of OVERALS clearly indicates the existence of the association between sets of indices that represent biodiversity and cultural diversity. However, the structure of two components that describe the association between biological and cultural diversity is different, depending on the types of diversity indices - POP, AREA and RICH.

Table 5. Nonlinear Canonical Correlation Analysis (OVERALS): biodiversity and cultural diversity indices - POP, AREA or RICH

Set - POP		Weights		Component Loadings	
		1	2	1	2
1	Bird & mammal diversity	-.838	.548	-.886	.115
	Plant diversity	-.107	-.969	-.482	-.724
2	Language diversity index	.769	-.041	.834	.279
	Religion diversity index	1.673	-4.019	.834	.183
	Ethnic group diversity	-1.579	4.292	.793	.318
Fit	1.510	Eigenvalue		.792	.719
Set - AREA		Weights		Component Loadings	
		1	2	1	2
1	Bird & mammal diversity	1.680	1.264	.534	-.417
	Plant diversity	-1.366	-.697	-.117	-.445
2	Language diversity index	-.110	-1.465	.517	-.829
	Religion diversity index	-1.163	.384	-.594	.805
	Ethnic group diversity	.986	.727	.315	.949
Fit	2.000	Eigenvalue		1.000	1.000
Set - RICH		Weights		Component Loadings	
		1	2	1	2
1	Bird & mammal diversity	1,860	-,108	,245	,468
	Plant diversity	-,497	-,644	-,248	,226
2	Language diversity index	-1,381	1,165	-,108	,546
	Religion diversity index	-,495	,849	,128	,758
	Ethnic group diversity	1,048	-,152	,754	,353
Fit	1.246	Eigenvalue		.825	.684

Note: the highest values of loadings for some indices are bolded

4. Discussion and Conclusions

The main reason why we have emphasized the role of NMA methods in studying biodiversity and ecosystem is complex and in general the most often nonlinear, with plenty of factors that have influence on each other. Results of the analyses which are conducted in this article illustrate that NMA methods could be informative when re-analyzing issues concerning biocultural diversity, in small data sets. NMA methods could be convenient tools for similar types of analyses because these methods are not oversensitive on the presumptions for using compatible linear methods. For example, NMA methods are less restricted by the sample size, measurement scales (or types of data) and types of data distributions. Namely, the only reasonable request that is needed to perform these types of analyses is numerical expression of the variables. Later, these variables can be transformed to other measuring scales.

Nonlinear methods are in this article applied on a small number of cases: 20 countries, from the source tables [15]. When applying CATREG, statistically significant multiple regression coefficient for the criterion (Index of biocultural diversity IBCD-RICH), with its only one statistically significant predictor (Plant diversity), could be a direct consequence of the fact that we were not able to use measures for absolute size of country area or absolute number of inhabitants in certain country, such as for indices of biocultural diversity AREA or POP. We have used the index IBCD-RICH, which is in fact composite measure of belonging biological and cultural diversity indices (RICH). Thus, it is the spurious correlation. The correlation between ratios of absolute measurements arises because of using ratios, rather than because of any actual correlations between the measurements [22]. This issue (spurious correlation) is especially important for the field of compositional data analysis, which deals with the analysis of variables that carry only relative information, such as proportions, percentages and parts-per-million [21]. Therefore, we can conclude that CATREG in fact does not show any statistically significant forecasting, perhaps because of the small sample sizes. However, the overall size of multiple regression coefficients, as well as the sizes of the weights indices (predictors), can indicate possible good ability to forecast indices of biocultural diversity, in the case of larger sample sizes. The results of centroid-based K-means clustering indicate that sets of indices of biocultural diversity – POP have been classified in different cluster than the indices of biocultural diversity AREA and RICH, which have practically the same features. The results of CATPCA indicate the same trends in two-component structure for all three sets of indices of biocultural diversity: POP, AREA and RICH: similar structure of components, similar level of variances explained and similar levels of reliability. The small number of variables (only two) which saturate component [5] could directly influence low reliability of the biodiversity component. Finally, two components obtained by OVERALS, which describe the associations between the sets of biological and cultural variables indicate different structures of correlations.

All the above-mentioned results, obtained using NMA methods, can provide a researcher with more information on the issues of interest. Besides cluster analysis, which is often used method in statistical analyses, other methods (CATREG, CATPCA and OVERALS) belong to the optimal scaling methods. Using linear methods, having numerical variables with a considerable amount of different

values is not efficient and can be quite time-consuming. This problem can be solved be the incorporation of B-splines [28], adding a vast amount of flexibility in terms of score transformations [17]. Gifi [6] offers a comprehensive collection of NMA methods, based on optimal scaling. As we have discussed before, the basic idea of optimal scaling is to transform the observed variables (categories) in terms of their quantifications, but this approach offers higher level of flexibility [18]. It is the opposite procedure than usual statistical methods, which often transform the variables of some type, only at a lower level of measuring, when the distributions are deviating from the Gausses curve, or when sample size is low. Optimal scaling enables converting nominal and ordinal variables into interval variables. Therefore, due to their flexibility, NMA methods could provide more information about the issues of interest (in this example, biocultural diversity). The logical framework of optimal scaling is described as follows [6]. Starting point of the analysis is a 0-1 dummy matrix, based on the data, which are considered as categorical. Subsequently, during the iterations, a loss function stretch/squeeze the variables and compute category scores such that they are optimal in the sense of a minimal loss function (procedure of optimal scaling) [17, 25]. This idea adds a vast amount of flexibility in terms of score transformations. The simplest NMA method is homogeneity analysis, while by imposing restrictions on the quantification ranks, we get CATPCA and by defining sets OVERALS [6].

Therefore, the main advantage of NMA methods is their basic assumption (idea) that each variable can basically be considered as categorical, while different scale levels can be incorporated by means of level restrictions (monotone, linear) on the scores [17]. Second advantage is offering extensions in terms of rank restrictions (CATPCA) and restrictions on sets of variables (OVERALS), as compared with compatible linear methods. Third, the resulting scores can be represented graphically in various ways.

On the other hand, even NMA methods are not 'immune' on the problem of overfitting, which occurs when a statistical model describes random error (noise), instead of real underlying relationship [7], especially in small data sizes (as same as in our biodiversity example). Namely, overfitting occurs when a model is very complex and has too many parameters, relative to the number of observations. Such overfitting model will generally have poor predictive performance, because minor fluctuations in the data could produce large deviations [7]. Finally, even mathematically optimal scaled (transformed) variables are always approximation, while objective functions that are applied in ordinal (or categorical) data analysis must be carefully adapted to the structure of the observed data. Likewise, any analysis of data that is based upon objective functions must lead to interpretable results [11].

REFERENCES

[1] Barrett, D. B., Kurian, G. T., Johnson, T. M. 2001. World Christian Encyclopedia: A Comparative Survey of Churches and Religions in the Modern World. 2nd ed. Oxford University Press, Oxford.

[2] Bell, T., Lilley, A. K., Hector, A., Schmid, B., King, L., Newman, J. A. 2009. A linear model method for biodiversity-ecosystem functioning experiments. The American Naturalist, 174, 836–849.

[3] Burkett, V. R., Wilcox, D. A., Stottlemeyer, R., Barrow, W., Fagre, D., Baron, J., Price, J., Nielsen, J. L., Allen, C. D., Peterson, D. L., Ruggerone, G., Doyle, T. 2005. Nonlinear dynamics in ecosystem response to climatic change: case studies and policy implications. Ecological Complexity, 2, 357-394.

[4] FAO – Food and Agriculture Organization of the United Nations. 2004. FAOSTAT (FAO statistical databases). FAO, Rome, Italy. Available from: http://apps.fao.org.

[5] Field, A. 2005. Discovering statistics using SPSS. 2nd Edition. Sage, London.

[6] Gifi, A. 1990. Nonlinear Multivariate Analysis. Wiley, Chichester, England.

[7] Good, P. I., Hardin, J. W. 2006. Common errors in statistics (and how to avoid them). Wiley-Interscience, Hoboken, N.J.

[8] Grimes, B. 2000. The ethnologue: languages of the world. 14th edition. Vol. 1. SIL International, Dallas.

[9] Groombridge, B., Jenkins, M. D. 2002. World Atlas of Biodiversity: Earth's Living Resources in the 21st Century. University of California Press, Berkley.

[10] Harmon, D., Loh, J. 2004. The IBCD: A measure of the world's biocultural diversity. Policy Matters, 13, 271–280.

[11] Herden, G., Pallack, A. 2005. Adequateness and interpretability of objective functions in ordinal data analysis. Journal of Multivariate Analysis, 94(1), 19–69.

[12] Höft, M., Barik, S. K., Lykke, A. M. 1999. Quantitative Ethnobotany: Applications of Multivariate and Statistical analyses in Ethnobotany. People and Plants Working Paper 6, WWF/UNESCO/Kew, Paris.

[13] Johnson, R. A., Wichern, D. W. 1988. Applied Multivariate Statistical Analysis. Prentice Hall, Englewood Cliffs, New Jersey.

[14] Linting, M., Meulman, J. J., Groenen, P. J. F., Van der Kooij, A. J. 2007. Nonlinear Principal Components Analysis: Introduction and Application. Psychological Methods, 12, 336-358.

[15] Loh, J., Harmon, D. 2005. A global index of biocultural diversity. Ecological Indicators, 5(3), 231-241.

[16] Maffi, L. 2001. Introduction: On the interdependence of biological and cultural diversity. In: Maffi, L. & J. A. McNeely (eds): On Biocultural Diversity. Linking Language, Knowledge and the Environment. Smithsonian Institution Press, Washington and London.

[17] Mair, P., de Leeuw, J. 2009. Rank and Set Restrictions for Homogeneity Analysis in R: the homals Package. In: JSM 2008 Proceedings, Alexandria, VA, pp. 2142-2149. American Statistical Association.

[18] Mair, P., de Loeew, J. 2010. General Framework for Multivariate Analysis with Optimal Scaling: The R Package aspect. Journal of Statistical Software, 32(9), 1-23.

[19] Mahecha, M.D., Martinez, A., Lischeid, G., Beck, E. 2007. Nonlinear dimensionality reduction: Alternative ordination approaches for extracting and visualizing biodiversity patterns in tropical montane forest vegetation data. Ecological informatics, 2, 138-149.

[20] Meulman, J. J., Heiser, W. J. 2007. PASW® Categories 17.0. SPSS Inc., Chicago, US.

[21] Pawlowsky-Glahn, V., Buccianti, A. (eds). 2011. Compositional Data Analysis: Theory and Applications. John Wiley & Sons, Chichester, UK.

[22] Pearson, K. 1897. Mathematical Contributions to the Theory of Evolution—On a Form of Spurious Correlation Which May Arise When Indices Are Used in the Measurement of Organs. Proceedings of the Royal Society of London 60, 489–498.

[23] SAS/STAT(R) 9.2. 2009. User's Guide, Second Edition. SAS Institute Inc., Cary, NC, USA.

[24] Slikkerveer, J. 2005. A Multivariate Model of Biocultural Conservation of Medicinal, Aromatic and Cosmetic (MAC) Plants in Indonesia. Ethnobotany Research and Applications 3, 127-138.

[25] Takane, Y. 2005. Optimal scaling. In: Everitt, B., and Howell, D. (eds.). Encyclopedia of Statistics for Behavioral Sciences (pp. 1479-1482). – Wiley, Chichester.

[26] The Times. 2000. Comprehensive Atlas of the World, 10th ed. Times Books, London.

[27] Thébault, E., Loreau, M. 2006. The relationship between biodiversity and ecosystem functioning in food webs. Ecological Research, 21, 17-25.

[28] van Rijckevorsel, J. L. A., de Leeuw, J. 1988. Component and correspondence analysis: Dimension reduction by functional approximation. Wiley, New York.

[29] Yamanaka, T., Raffaelli, D. White, P. C. L. 2013. Non-Linear Interactions Determine the Impact of Sea-Level Rise on Estuarine Benthic Biodiversity and Ecosystem Processes. - PLoS ONE 8(7), e68160. doi:10.1371/journal.pone.0068160