Selection of the optimal unit of analysis in assessing the structure of terrestrial arthropods assemblages

M Kirichenko-Babko1, Y Danko2 and D Majerek3

1 National Academy of Sciences of Ukraine, , Schmalhausen Institute of Zoology, Department Fauna and Systematics of Invertebrates, B. Khmelnytsky 15, 01030 Kyiv, Ukraine
2 Sumy Makarenko State Pedagogical University, Department of General Biology and Ecology, Romenska 87, 40002 Sumy, Ukraine
3 Lublin University of Technology, Faculty of Fundamentals of Technology, Nadbystrzycka 38, 20-618 Lublin, Poland.

d.majerek@pollub.pl, kirichenko@izan.kiev.ua, yaroslavdanko@gmail.com

Abstract. The article shows the use of statistical methods of multivariate analysis on the example of environmental data – the species composition of terrestrial arthropods. It is shown that the research results largely depend on the scale of observations made – whether it is a study conducted in a whole ecosystem or in a separate landscape element, type of habitat or, finally, an analysis of elementary samples. The results and their interpretations directly depend on the scale chosen by the researcher. We were mainly concerned with the reliability of using of multidimensional statistics in practical situations.

1. Introduction

Modern ideas about the quality of the environment rely on the study of its biological components. Attention is increasingly paid to indicators characterising the structure of biocenoses. Today, the most popular are index estimates based on species diversity [1; 2]. At the same time, problems with research coverage, choice of an adequate sample size, which can be considered representative with respect to one or another object, remain quite controversial [3]. Very often, different samples from the same ecosystem can produce very different results. This is quite fair, because ecosystems or their fragments are very complex, not to mention processes that keep them in dynamically stable condition [4]. The environmental and biodiversity assessment could be performed on territories of different size and materials with various depth of detailing [5]. Its interpretation depend on scale of compared objects and their deep study. Some researchers [6;7;8] emphasise that the scale of observations made — this study conducted in the landscape, the whole ecosystem, habitat or, finally, an elementary sample — significantly affect the results.

We analysed the possibility of interpretation of quantitative data (abundance) obtained at various levels of research. Thus, the subject of our work was to select the most appropriate scale of the territory covered by the research, to obtain the most representative results and their adequate interpretation.
2. Material and methods

2.1. Statistical analysis

Data were processed using R, version 3.6.1 [13]. Hierarchical clustering was performed with hclust (function ward.D2) from core R package stats. Two types of p-values (AU and BP) for the nodes of hierarchical clustering dendrograms were calculated with R package pvclust. According to the package documentation, “AU p-value, which is computed by multiscale bootstrap resampling, is a better approximation to unbiased p-value than BP value, computed by normal bootstrap resampling. Clusters with AU larger than 95% are highlighted by rectangles, which are strongly supported by data” [15]. K-mean clustering was performed with function kmens from package cluster [12] and plotted on the plain of principal components (PCA) with fviz_cluster from factoextra [11]. Plots were produced using packages ggplot2 [16], factoextra [11], ggrepel [14], directlabels [10]. Species data analysed with and without Hellinger transformation [18].

2.2. Region of study and collected data

The study was carried out in a steppe landscape at the regional landscape park (RLP) of Kleban-Byk, which is located about 10 km from Konstantinovka town (eastern part of Ukraine, Donetsk Oblast). The territory of the RLP Kleban-Byk is situated within the western outskirts of the Donetsk Ridge and is a wavy plateau, dissected by deep rivers and a ravine-beam system. This region has undergone many changes in the last 150 years. Petrophytic and crushed stipa steppes, real hygrophytic stipa steppe and stony outcrops occupy the largest area. Ravine and artificial forests occupy considerably smaller area.

The material was being collected from June to August 2005-2006 in seven locations of RLP (Tab. 1) [19]. Sampling has focused on a wide range of habitats, including steppes in gully and in plakor, shrubby vegetation on the steppe, forests, meadows, arable field and shore of ponds (Tab. 1). Carabid beetles were sampled using pitfall traps which were filled with preservative fluid – a propylene glycol [20]. In each sampling site 15 traps were used, arranged in a row with a spacing of 3 m between the traps. 70 species of ground beetles (Coleoptera, Carabidae) were identified at different habitats in the territory of the RLP Kleban-Byk.

| Code and Name of locations | Habitats & the area (ha) of locations |
|----------------------------|--------------------------------------|
| Loc-1 Druzhkovski skamjanili trees | Steppe, shore of pond, field; 110,5 ha |
| Loc-2 Gorilyj forest | Shrub steppe (“caragana”), wet meadow; 60 ha |
| Loc-3 Olenivskyj forest | Deciduous forest, meadow; 161,3 ha |
| Loc-4 Bila gora | Steppe, shore of pond; 134,7 ha |
| Loc-5 Ridkodub | Steppe with oak trees, meadow; 52,3 ha |
| Loc-6 Ravine Kravets’ka | Steppe, shore of pond; 28,6 ha |
| Loc-7 Kleban-Byk reservoir | Reeds on shore, forest belt, meadow; 650 ha |

3. Results

Environmental studies involve the study of all components of ecosystems. Statistical methods provide an arsenal of techniques that allow to analyse data presented as “presence-absence” of species in samples or habitats or their changing abundances [21].
Due to the heterogeneity of the relief, the steppe landscape is represented by different types of habitats: the steppe proper localized, as a rule, on the open plateau; slopes of ravines overgrown with shrubs; ravine forests limited to low areas of relief; shore of the reservoir on the Kleban-Byk River and shores of small ponds. Materials were collected in 7 locations (Table 1) remote from each other but representing one type of ecosystem – the steppe and were analysed at two levels: (1) at the level of individual samples and (2) at the level of samples grouped by habitat. At the level of the working hypothesis, it could be assumed that different types of habitats should have a similar species composition of ground beetles. Therefore, according to this hypothesis, the presence of structures at the level of habitat types: open and shrubby steppe, ravine forest, shores of water bodies, etc. was expected. Obviously, expected was the presence of associations at the level of samples that were taken in various habitats. Using different levels of generalisation of data and methods of statistical analysis, an attempt is made to assess objectivity of the existence of the structure of the ground beetles’ assemblage. And also, to evaluate which of the methods of analysis, and at what level will give a more objective idea of the structure of the assemblage.

The results of Ward’s hierarchical clustering of the matrix with Kulczynski distances are presented in (Figure 1 A, B). The distance matrix was obtained on the basis of a Q-mode table in which the rows are individual samples, named for the corresponding habitats, and the columns are the abundance species present in these habitats. The results of cluster analysis at the level of local samples showed that our working hypothesis is not confirmed when using this method.

![Figure 1](image_url)

Figure 1. Ward’s hierarchical clustering of a Kulczynski distances matrix among samples. A — untransformed, B — Hellinger-transformed data. Samples are named for their respective habitats. First number after the name of habitat corresponds to the number of locations. AU (Approximately Unbiased) and BP (Bootstrap Probability) p-values are shown next to the nodes. Reliability of six clusters is confirmed by the AU (Approximately Unbiased) p-values obtained by bootstrap method (Figure 1 A). Although AU p-values are considered more reliable than BP (Bootstrap Probability) p-values, and it is on the basis of the first that a conclusion is made on reliability of the cluster, it should be noted that BP values are extremely low. The obtained cluster dendrogram showed that both these “reliable” clusters and higher and lower level clusters cannot be adequately interpreted biologically. Thus, hierarchical clustering on the basis of non-transformed data at the level of individual samples does not allow to draw conclusion about the structure of the ground beetles’ assemblage. At this generalisation level, all studied plots and habitats are presented as a single territory on which the species composition is not structured.
Quite often, transformation of data, by one method or another, significantly improves the results of clustering. The dendrogram shows that transformation data by Hellinger has changed the composition of the clusters (Figure 1 B). Despite the fact that the application of the Hellinger’s transformation before hierarchical clustering yielded two “reliable” according to p-values clusters, from a biological point of view, this separation of samples into two groups is just as uninterpretable as clustering obtained without transformation (Figure 1 A).

Further, a non-hierarchical K-means clustering was applied to previously non-transformed sample data. This method requires pre-setting the desired number of clusters. To determine the optimal number of clusters, we used two methods: silhouette and Calinski-Harabasz [22]. Both of these methods consider two as the optimal number of clusters. Since two are, in fact, the minimum acceptable number of clusters, a priori, most likely, all clustering results will be banal. To visualise the clusters, they were depicted on the plane of the first two principal components of the PCA (Figure 2 A). Clustering by K-means method without data transformation, as well as hierarchical clustering, showed that all samples make up a one unit. Only a sample taken on the shore of one of the reservoirs stood out in a separate cluster.

**Figure 2.** K-means clustering of samples from various habitats on the plain of the first two principal components. A — untransformed, B — Hellinger-transformed data. The number after the habitat name indicates number of samples.

The methods for determining the optimal number of clusters for Hellinger-transformed data gave the following results. The silhouette method recommends 15 clusters; such a number of clusters seems senselessly large. The Calinski-Harabasz method recommends defining two clusters, alike in the case of untransformed data. Figure 2 B shows the results of clustering for a given number of clusters two. In this case, unlike non-transformed data, the second cluster is already represented not by one sample, but by twelve, which at first glance looks encouraging. Nevertheless, from a biological point of view, the interpretation of the allocation of groups is very difficult and cannot be associated with objective factors that determine the structure of the ground beetles’ assemblage [23]. Transformed data demonstrate the presence of certain structures, the principle of separation of which, obviously, has a purely mathematical basis. Thus, we can conclude that the PCA also does not exhibit a biologically meaningful structure in our samples. In general, the first two principal components account for an extremely small fraction of the variability in data.

Examination of the cluster dendrogram obtained based on data grouped by habitat showed, that without transformation only a field and a wet meadow were found in the “reliable” cluster (Figure 3 A).
After Hellinger’s transformation, two reliable clusters emerged (Figure 3 B). Such a division has an obvious biological meaning. Thus, the structure of the assemblages of ground beetles of the studied sites is represented by two components: one group of species is associated with xerophytic and petrophytic habitats of the steppe, and the other – with habitats along the shores of reservoir (“reeds”) and ponds, meadows in ravines and fields. Fields as artificial habitats are usually grouped with adjacent habitats [24]. The studied fields were really close to the water bodies and adjoined to the meadows.

The K-mean method applied to the habitat data without transformation did not show their structuredness in Figure 4 A. According to the silhouette method, the optimal number of clusters is two, the Calinski-Harabasz method – 7. Seven are obviously too many clusters, therefore, we chose two. On the plane of the first two principal components of the PCA (Figure 4 A), K-mean clustering without data transformation showed that all habitats, as well as samples form a whole. Only a sample taken on the shore of one of the reservoirs again stood out in a separate cluster.

Methods for determining the optimal number of clusters for Hellinger-transformed data for habitats gave the following results. The silhouette method recommends three clusters; the Calinski-Harabasz method recommends four clusters. More sensible is the option with three clusters (Figure 4 B). The shores open and covered with a reed stood out separately, which is quite logical. The grouping of fields and meadows is probably due to their artificial origin. Natural steppe areas are separated into a separate cluster, which makes biological sense, since these various habitats accumulate mainly xerophilic fauna. This article analyses faunistic data – samples collected in several fragmented areas of the steppe landscape. Since these samples were collected at the same time under almost identical conditions, the obtained results seem to be consistent with the neutral theory [25; 26]. Neutral theory considers rich and diverse biocenoses, such as tropical forests or coral reefs, where all events are stochastic. Environmentalists argue that factors limiting the number of species in communities, and even more so the question of whether they are truly stochastic or deterministic, have not yet been clarified [27]. It is possible that the answers to these questions can be obtained as a result of the study of different types of biocenoses.

Thus, we can conclude that the methods of K-mean [28] and PCA at the level of data habitat after Hellinger transformation exhibit a biologically meaningful, logical structuring. To establish and analyse the structure of assemblages, many researchers often use multivariate analysis such as principal component analysis (PCA) and cluster analysis [29; 30; 9].
Figure 4. K-mean clustering of habitat types (without considering their location in seven territories of the Kleban-Byk RLP) on the plain of the first two principal components. A — untransformed, B — Hellinger-transformed data.

4. Conclusion

As an analysis of the data obtained during the study showed, the structure of the ground beetle’s assemblages in the steppe landscape cannot be detected using one particular method. It is shown that the identification of structures of mobile biological objects assemblages involves a preliminary transformation of the data.

An important aspect for the systematisation of data are two points — the averaging of samples from similar habitats, and preliminary transformation of the data when analysing the structure at the habitat level. In our case, the application of Hellinger transformation was very successful.

5. References

[1] Clarke K R and Warwick R M 2001 A further biodiversity index applicable to species lists: variation in taxonomic distinctness Marine Ecology Progress Series 216 265-78
[2] Shimadzu H and Darnell R. 2015 Attenuation of species abundance distributions by sampling R. Soc. Open Sci. 2 140219 http://dx.doi.org/10.1098/rsos.140219
[3] Melo A S, Pereira R A S, Santos A J, Shepherd G J, Machado G, Medeiros H F and Sawaya R J 2003 Comparing species richness among assemblages using sample units: why not use extrapolation methods to standardize different sample sizes? Oikos 101 398–410
[4] Banks-Leite C, Ewers R M and Metzger J P 2012 Unraveling the drivers of community dissimilarity and species extinction in fragmented landscapes Ecology 93(12) 2560–69
[5] Anderson M J, Ellingsen K E and McArdle B H 2006 Multivariate dispersion as a measure of beta diversity Ecology Letters 9 683–93
[6] Ellwood M D F, Manica A and Foster W A 2009 Stochastic and deterministic processes jointly structure tropical arthropod communities Ecol. Lett. 12 277–84
[7] Ribera I, Foster G N and Vogler A P 2003 Does habitat use explain large scalespecies richness patterns of aquatic beetles in Europe? Ecography 26 145–52
[8] Pakulnicka J, Górski A, Bielecki A, Buczyński P, Tończyk G and Cichocka J M 2013 Relationships within aquatic beetle (Coleoptera) communities in the light of ecological theories Fundam. Appl. Limnol 183/3 249–58
[9] Kirichenko-Babko M, Łagód G, Majerek D, Franus M and Babko R 2017 The effect of landscape on the diversity in urban green areas *Ecological Chemistry and Engineering* S **24**(4) 613-625. DOI: 10.1515/eces-2017-0040

[10] Hocking T D 2018 directlabels: Direct Labels for Multicolor Plots. R package version 2018.05.22. https://CRAN.R-project.org/package=directlabels

[11] Kassambara A, Mundt F 2017 factoextra: Extract and Visualize the Results of Multivariate Data Analyses. R package version 1.0.5. https://CRAN.R-project.org/package=factoextra

[12] Macchler M, Rousseeuw P, Struyf A, Hubert M and Hornik K 2019. cluster: Cluster Analysis Basics and Extensions. R package version 2.0.8.

[13] R Core Team 2019 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[14] Slowikowski K 2018 ggrepel: Automatically Position Non-Overlapping Text Labels with 'ggplot2'. R package version 0.8.0. https://CRAN.R-project.org/package=ggrepel

[15] Suzuki R, Shimodaira H 2015 pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling. R package version 2.0-0. https://CRAN.R-project.org/package=pvclust

[16] Wickham H 2009 ggplot2: elegant graphics for data analysis. Springer New York.

[17] Huisman J, Olff H and Fresco L F M 1993 A hierarchical set of models for species response analysis *Journal of Vegetation Science* **4** 37–46, https://doi.org/10.2307/3235732

[18] Ruokolainen L and Blanchet G 2014 Introduction to ecological multivariate analysis Univ. Helsinki **79**

[19] Kirichenko M B and Babko R V 2007 Species diversity of ground beetles of the Kleban-Byk Regional Landscape Park. *Abstracts of 7th IUT Congress* (September 14-18, 2007, Nizhyn) **52** [in Ukrainian]

[20] Jarosik V 1992 Pitfall trapping and species-abundance relationships: a value for carabid beetles (Coleoptera, Carabidae) Acta *Entomologica Bohemoslovaca* **89** 1–12.

[21] Green J L and Plotkin J B 2007 A statistical theory for sampling species abundance *Ecol. Lett.* **10** 1037–1045 (doi:10.1111/j.1461-0248.2007.01101.x)

[22] Kingrani S K, Levene M and Zhang D 2018 Estimating the number of clusters using diversity *Artificial Intelligence Research* **7**(1) 15–22.

[23] Trivellone V, Bougeard S, Giavi S, Krebs P, Balseiro D, Dray S and Moretti M 2017 Factors shaping community assemblages and species co-occurrence of different trophic levels. *Ecol Evol.*, **7** 4745–54.

[24] Wamser S, Dauber J, Birkhofer K and Wolters V 2011 Delayed colonisation of arable fields by spring breeding ground beetles (Coleoptera: Carabidae) in landscapes with a high availability of hibernation sites *Agriculture, Ecosystems and Environment* **144** 235–40

[25] Bell G 2001 Neutral macroecology *Science* **293** 2413–18.

[26] Chave J 2004 Neutral theory and community ecology *Ecol. Lett.* **7** 241–53

[27] Zhou S R and Zhang D Y 2006 Allee effects and the neutral theory of biodiversity *Funct. Ecol.* **20** 509–513

[28] Hill M O, Harrower C A and Preston C D 2013 Spherical k-means clustering is good for interpreting multivariate species occurrence data *Methods in Ecology and Evolution* **4** 542–51

[29] Jurkiewicz-Karnkowska E 2011 Effect of habitat conditions on the diversity and abundance of molluscs in floodplain water bodies of different permanence of flooding *Pol. J. Ecol.* **59** 165–78

[30] Lenda M, Skórk P, Moron D, Rosin Z M and Tryjanowski P 2012 The importance of the gravel excavation industry for the conservation of grassland butterflies *Biol. Conserv.* **148** 180–90.