Comparison of the use of species abundance and presence-absence data for diversity assessment

M B Kirichenko-Babko¹, Y M Danko², J M Danylkiv¹ and D Majerek³

¹ Schmalhausen Institute of Zoology NAS of Ukraine, Department of Invertebrate Fauna and Systematics, B. Khmelnitsky str. 15, 01030 Kyiv, Ukraine
² Sumy State Pedagogical University, Faculty of Nature Science, Romenska str. 87, 40002 Sumy, Ukraine
³ Lublin University of Technology, Faculty of Fundamentals of Technology, Nadbystrzycka 38, 20-618 Lublin, Poland

d.majerek@pollub.pl

Abstract. The article is devoted to the analysis of empirical data on the distribution of ground beetles in three model sites located in Lublin (Poland). Using Principal Coordinates Analysis (PCoA) and hierarchical cluster analysis, we compared the results of the studies based on the data of species × abundance and binary data (species × presence / absence). It was shown that the hierarchical clustering method and PCoA based on binary data demonstrate the individuality of the studied territories, although they have some common species. While the results of the analysis, based on abundances, did not show a clear separation of the stations within the three studied locations, the similarity between the studied territories is more objectively reflected from a biological point of view.

1. Introduction
Assessment of the diversity of ecosystems or their fragments is an important branch of ecology and is based on the information on the number of species, their spatial distribution and population density. At the same time, the interpenetration of species from one territory or ecosystem to a neighboring one is a natural phenomenon, but it introduces adjustments to the idea of their integrity and individuality. In this regard, ecologists pay considerable attention to quantitative research and the characteristics of the distribution of populations within the studied objects.

In synecological studies, both the quantitative data and the data on the presence or absence of species are widely used. Obviously, the presence/absence assessment method is the least time-consuming and therefore often used to assess species richness and other indicators of diversity [1,2]. However, such data may not be enough, for example, if it is necessary to obtain the long-term forecast. In this case, the quantitative data are expected to provide more reliable information about the diversity of the analyzed groups and their distribution [3]. Therefore, it is important to answer the question of how comparable the results obtained with these methods are. The most objective answer can be obtained based on the analysis of empirical data on abundances, which can be easily converted into binary form (i.e., presence / absence).

When choosing the method of data analysis, we were guided by the fact that the linear ordination method PCoA (principal coordinates analysis, metric multidimensional scaling) is more suitable in our case, since it allows using any distance matrix; also, it is a priori clear that the sites we studied are in an
urbanized landscape and their beta diversity is low [4,5]. Most ecologists, when choosing the diversity indices, use the distribution of species abundance, neglecting the species richness [6]. Indeed, in many cases, species richness is an uninformative measure. In some cases, an estimate based on the relative abundance of taxa is more informative.

The aim of our research was to use statistical methods to compare the results of the analysis of empirical data.

2. Material and Methods

2.1. Statistical analysis

The data were processed using R version 4.0.2 [7]. The data transformations and distance matrices were produced with the decostand and vegdist functions from the vegan package [8]. Principal Coordinates Analysis [9] was performed with the pco function from the ecodist package [10]. 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 the 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. The clusters with AU larger than 95% are highlighted by rectangles, which are strongly supported by data” [11]. The plots were produced using packages ggplot2 [12], directlabels [13].

For the statistical analysis, we used the quantitative data (abundances) on the distribution of 68 species of ground beetles in 19 stations. This data was also used in binary form (presence / absence). In order to assess the diversity of the ecosystem, two types of measures were used. Alpha diversity refers to the diversity within a particular area or ecosystem, and is usually expressed by the number of species (i.e., species richness) in that ecosystem. If we examine the change in species diversity between these ecosystems, then we are measuring the beta diversity. We are counting the total number of species that are unique to each of the ecosystems being compared. Beta diversity is the ratio between the regional and local species diversity.

2.2. Object of study

The data on the species composition of ground beetles (Carabidae) from three model locations were used for the analysis: Saski park (PS 1-6), City Park (PM 1-7) and the riparian zone of the Bystrzyca river (RB 1-6) in the city of Lublin (Poland) (figure 1). These locations represent green areas in the city with different conditions that imply differences between the ground beetles assemblages inhabiting them. A detailed description of these territories is given in the papers [13,14].

![Figure 1. Location of the sampled territories in the city of Lublin. Abbreviations: PS – Saski park (dark blue), PM – City park (blue), or Public park and RB – Bystrzyca riverside (dark red double arrow).](image)
The study was carried out in 2011-2012 and 2014-2016. The beetles were sampled with pitfall traps. In these locations, 68 species of ground beetles were identified according to Hůrka [15]. Out of these, 31 species were found in the Saski Park, 37 species occurred in the City Park, and 43 species – in riparian zone of Bystrzyca river (table 1). At the same time, 9 species were found in all these territories. The number of species found exclusively in each of these territories is shown in table 1.

Table 1. The number of ground beetles species found in each of the studied locations.

| Species total | Number of species found only here | Shannon’s $^a$ $H$ | Whittaker’s $^b$ $\beta_w$ |
|---------------|----------------------------------|-------------------|-----------------------------|
| Saski Park, PS | 31                              | 12                | 2.471                       | 1.296                       |
| City Park, PM | 37                              | 6                 | 3.432                       | 1.59                        |
| Riparian zone of the Bystrzyca river, RB | 43                              | 14                | 3.49                        | 1.505                       |

$^a$ – Alpha-diversity; $^b$ – Beta-diversity.

3. Results

When comparing the studied sites using the data on abundances and on the basis of binary data (presence / absence) by the Principal Coordinates Analysis (PCoA) method, it turned out that the distribution of the studied stations shows similar trends (figure 2a, b). On both graphs, the most separated from others were the stations of the Saski Park (PS1-6). At the same time, the stations of the Saski Park formed a single group both on the basis of the data on abundances (figure 2a) and using the binary data on the presence / absence of species (figure 2b). The stations of the City Park, localized in the floodplain of the Bystrzyca river (RM 1-7), and stations of the riparian zone of the Bystrzyca river (RB1-6) are more individualized on the basis of binary data and are localized on the right in the upper and lower parts of the PCoA plot, respectively (figure 2b).

Moreover, when the data on abundance were used, half of the stations on the banks of the river Bystrzyca (RB 1-3) formed a common group with stations from the City Park (PM3 and PM5) in the upper right part of the graph (figure 2a). The same stations (RB 1-3) were positioned closer to the stations of the City Park, also on the graph obtained using the binary data; however, both of these groups remained on the opposite sides of the zero value of the second axis. In general, it can be stated that with significant similarities in the distribution of stations of the three studied locations, they form more distinct and compact groups on a graph based on binary data (figure 2b).

![Figure 2](image_url)

**Figure 2.** PCoA plot of stations studied: a – non-transformed data on the abundances of ground beetles using the Bray-Curtis distance matrix; b – presence / absence data, Jaccard distance matrix. For the designations of the stations of the three studied locations, see table 1.
After transforming the abundance data with the Hellinger method and using the Bray–Curtis distance matrix, the location of the stations on the graph (figure 3) becomes closer to that obtained using binary data. In addition, the stations of the Saski Park are more closely grouped than those of the other two territories.

Figure 3. PCoA plot of stations studied. The data on abundances were Hellinger-transformed; distance matrix of Bray–Curtis. Designations of stations, as in figure 2.

Hierarchical clustering performed by using the Ward method demonstrated that regardless of the form of data — abundance or presence / absence — the three locations studied are divided into two groups (figure 4a, b). The differences when using these two types of data are the grouping of locations and the significance levels of these clusters. At the same time, at the most reliable level, two clusters were distinguished using binary data (figure 4b): the bootstrapped au p-values equal 99 and 98. Partitioning the studied locations into two clusters based on the abundances data (figure 4a) was realized at a significantly lower level of significance (bootstrapped au p-values equal 79 and 77).

Regardless of the type of data being analyzed (figure 4a, b), the stations of the Saski park are combined into one cluster. However, at the abundance dendrogram (figure 4a), station RB6, located on the banks of the Bystrzyca River, joined the stations in the Saski park.

On the basis of the binary data (figure 4b), the second cluster was subdivided into two subclusters: one included the stations on Bystrzyca river, in another – stations of the City Park. On the dendrogram, built on the basis of the abundance of ground beetles (figure 4a), there is no obvious separation of these two territories: The City Park and the riparian zone of the Bystrzyca river.

Clustering carried out after data transformation using the Hellinger method showed an intermediate variant between the two variants considered earlier (figure 5). The Hellinger-transformed data on the number of ground beetles also divided the stations into two clusters. The first cluster included exclusively the Saski Park stations, similar to the clusterization of binary data. However, the second cluster did not subdivide in a clear way into groups: the City Park and the banks of the Bystrzyca River, similarly to what was observed when clustering the data based on abundances.
Thus, the Ward clustering method based on binary data and clustering data on the abundance of ground beetles after data transformation using the Hellinger method made it possible to distinguish two groups of stations with high reliability (96 and 98): (1) the Saski park, (2) the Bystrzyca river banks and the City park. Moreover, only the clustering of binary data demonstrates the separation of stations on the bank of the Bystrzyca river and stations of the City Park, located in the floodplain of the river, into two independent groups.

It is important to note that in the hierarchical clustering performed using the Ward's method, the use of the Hellinger's transformation of the data on the abundances before calculating the distance matrix (figure 5) gives the results similar to the results of clustering the data on species presence-absence (figure 4b).

**Figure 4.** Hierarchical clustering according to the Wards method: a – non-transformed data on abundances, Bray-Curtis distance matrix; b – data on the presence / absence, Jaccard distance matrix. Near nodes $au$ and $bp$ p-values (see material and methods), bootstrap replications number 1000. Designations of stations, as in figure 2.

**Figure 5.** Ward's clustering of Hellinger-transformed data on abundances of ground beetles (Bray-Curtis distances). Near nodes $au$ and $bp$ p-values (see material and methods), bootstrap replications number 1000. Designations of stations, as in figure 2.
3. Discussion

Biological diversity is an important parameter of environmental research and its indicators are used to assess the quality of the environment and compare territories [16,17]. Such assessments are usually preceded by the studies that can be both qualitative (presence / absence of species) and quantitative. The results of studies can be significantly affected by seasonality, circadian rhythms, accounting methods, and even the qualifications and experience of the researcher.

Ground beetles, as a test object, are often used in assessing the quality of the environment [18-21]. At the same time, the zoological studies are aimed at assessing the diversity of species, while the environmental studies allow identifying the dominant (background) species as well as establishing the quantitative ratio of populations in samples, at individual stations, or compared territories [22]. Given the complexity of quantitative research in the analysis of alpha and beta diversity, they are often based on the presence / absence data [4,23,24,25]. When evaluating beta diversity based on the presence / absence data, the Whittaker measure ($\beta_{w}$) is used [26].

The results of the analysis performed by means of the PCoA method showed that the statistical analysis of the data on the basis of quantitative and binary data gives a similar idea of the situation and the similarity of the studied territories (figure 2a, b). The most compact group in both cases was formed by the locations from the Saski Park, which indicates a low beta diversity of this territory. This was also confirmed by the calculations of the Whittaker measure (table 1): its minimum value is observed in the Saski park (1.296). It should be noted that alpha diversity (based on the Shannon index) is also lower in the Saski park (2.471) than in the other two locations.

In general, this confirms [27] the conclusion that “presence/absence data can yield similar assessment results to those for abundance-based data, despite type-specific deviations. For most metrics, it should be possible to inter calibrate the two data types without substantial efforts.” The authors also argue that “systematic and large-scale studies directly comparing data on the abundance of taxa and presence/absence are few and molecular data could be used to simply replace the classical taxa × abundance matrices with taxa × presence/absence matrices”.

At the same time, given the high price of molecular research and the comparative simplicity of the classical zoological research, the latter remain relevant. Moreover, genetic studies often give a greater number of species (genetic) than the number of morphospecies which are actually present [28].

The studied territories differed in the level of alpha (H) and beta-diversity (Whittaker) (table 1). It turned out that the stations of the Bystrzyca river riparian zone (RB) and the City park (PM) in the floodplain of the river had very close values of both the Shannons' and the Whittakers’ indices. Nevertheless, PCoA showed that the three locations are quite individual. Moreover, a graph based on a binary matrix gave the maximum scatter of stations and the absence of their intersection (figure 2b). In turn, the positioning of the locations of the City Park (PM) and the riparian zone of Bystrzyca river (RB) based on quantitative data (the abundance of ground beetles) demonstrates that a number of stations in these territories are very close (overlap). However, the results obtained on the basis of the abundance data are more informative.

The comparison of the binary and quantitative data using hierarchical cluster analysis with the Wards' method also showed some differences. When using the binary data, a more radical separation of the studied locations is observed (figure 4b): the stations of the Saski Park are separated from the City Park and the banks of the Bystrzyca river at the highest level of confidence. In turn, the banks of the Bystrzyca river and the City Park form two subclusters of one cluster. At the same time, clustering based on abundance did not give such unambiguous divisions of the studied areas. Although the stations of three locations were divided into two clusters, their content does not correspond to the distribution of locations and the level of confidence is much lower than based on the binary data. The quantitative data (species × abundance) allow a finer interpretation of the results. However, it is obvious that the three territories have a common genesis.

In the analysis of the biological data, the use of data transformation is quite popular. The results of the analysis of the abundance data after transformation performed by means of the Hellinger method (figure 3) showed that the transformation eliminates the features of the quantitative distribution of
populations and gives a similar result to the analysis on the binary data (figure 2b). The result of the transformation also significantly changed the results of clustering (figure 5). Therefore, like clustering binary data, a set of stations split into two clusters at a high level of reliability. At the same time, as in the case with binary data (figure 4b), only the Saski Park stations were allocated into one cluster (figure 5). However, the clustering of stations on the territory of the City Park and the banks of the Bystrzyca River on the basis of transformed abundance data did not give their separation and demonstrates the similarity of beetle populations of these two territories (figure 4a).

4. Conclusions
In general, from a biological point of view, the use of both binary and quantitative data is justified, but their interpretation is more likely to complement one another than to exclude one of them. Obviously, the results of statistical processing of binary data show that the green areas of Lublin have retained their individuality and have a fairly high biodiversity for the urbanized area. In turn, the analysis based on the abundance of ground beetle species showed that the studied sites have a common genesis and, complementing each other, represent a territory of the same type: in this case, the remains of an intrazonal landscape.

References
[1] Dorazio R M and Royle J A 2005 Estimating size and composition of biological communities by modeling the occurrence of species J Am Stat Assoc 100 389–398
[2] Dorazio R M, Royle J A, Söderström B and Glimskär A 2006 Estimating species richness and accumulation by modeling species occurrence and detectability Ecology 87 842–854
[3] Thompson W L, White G C and Gowan C 1998 Monitoring Vertebrate Populations (Massachusetts: Academic Press) p. 365
[4] Koleff P, Gaston K J and Lennon J J 2003 Measuring beta diversity for presence–absence data J Anim Ecol 72 367–382
[5] Faith D P, Minchin R M and Belbin L 1987 Compositional dissimilarity as a robust measure of ecological distance Plant Ecol 69 57–68. DOI: 10.1007 / BF00038687
[6] Chao A, Chazdon R L, Cotwell R K and Shen T J 2005 A new statistical approach for assessing similarity of species composition with incidence and abundance data Ecol. Lett. 8 148–159. DOI: 10.1111/j.1461-0249.2004.00707.x
[7] R Core Team 2019 R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/
[8] Oksanen J, Blanchet F G, Friendly M, Kindt R, Legendre P, McGlinn D, Minchin P R, O’Hara R B, Simpson G L, Solymos P, Henry M, Stevens H, Szoecs E and Wagner H 2019 vegan: Community Ecology Package. R package version 2.5-6. https://CRAN.R-project. org/package=vegan
[9] Gower J C 1966 Some distance properties of latent root and vector methods used in multivariate analysis Biometrika 53 325–338
[10] Goslee S C and Urban D L 2007 The ecodist package for dissimilarity-based analysis of ecological data J. Stat. Softw. 22(7) 1–19
[11] Suzuki R, Terada Y and Shimodaira H 2019 pvclust: Hierarchical Clustering with P-Values via Multiscale Bootstrap Resampling. R package version 2.2-0. https://CRAN.R-project.org/package =pvclust
[12] Wickham H 2009 ggplot2: elegant graphics for data analysis. Springer New York
[13] Hocking T D 2018 directlabels: Direct Labels for Multicolor Plots. R package version 2018.05.22. https://CRAN.R-project.org/package=directlabels
[14] Kirichenko M, Lagod G and Babko R 2012 Distribution of carabid beetles (Coleoptera, Carabidae) in the urban area of Lublin Proc of ECOpole 6(1) 157–163 DOI:10.2429/proc.2012.6(1)021
[14] Kirichenko-Babko M, Łagód G, Majerek D, Franus M and Babko R 2017 The effect of landscape on the diversity in urban green areas Ecol Chem Engin S 24(4) 613–625. DOI: 10.1515/eces-2017-0040

[15] Hůrka K 1996 Carabidae of the Czech and Slovak Republics (Zlin: Kabourek)

[16] Hooper D U et al. 2005 Effects of biodiversity on ecosystem functioning: A consensus of current knowledge and needs for future research Ecol. Monogr 75 3–35

[17] Brown J H, Ernest S K M, Parody J M, and Haskell J P 2001 Regulation of diversity: maintenance of species richness in changing environments Oecologia 126 321–32

[18] Duelli P and Obrist M K 2003 Biodiversity indicators: the choice of values and measures Agric Ecosyst Environ 98 87–98

[19] Pearce J L and Venier L A 2006 The use of ground beetles (Coleoptera: Carabidae) and spiders (Aranae) as bioindicators of sustainable forest management: a review Ecol Indic 6 780–793

[20] Avgn S S and Luff M L 2010 Ground beetles (Coleoptera: Carabidae) as bioindicators of human impact Munis Entomol Zool 5 209–215

[21] Cameron K H and Leather S R 2012 How good are carabid beetles (Coleoptera, Carabidae) as indicators of invertebrate abundance and order richness? Biodivers Conserv 21 763–779 DOI 10.1007/s10531-011-0215-9

[22] Verberk W C E P, van der Velde G and Esselink H 2010 Explaining abundance-occupancy relationships in specialists and generalists: A case study on aquatic macroinvertebrates in standing waters J Anim Ecol 79 589–601

[23] Martin J et al. 2011 Accounting for non-independent detection when estimating abundance of organisms with a Bayesian approach Methods Ecol. Evol. 2 595–601

[24] Beck J et al. 2013 Undersampling and the measurement of beta diversity Methods Ecol. Evol. 4 370–382

[25] Dorazio R M et al. 2013 Estimating abundance while accounting for rarity, correlated behavior, and other sources of variation in counts Ecology 94 1472–1478

[26] Whittaker R H 1960 Vegetation of the Siskiyou mountains, Oregon and California Ecol. Monogr 30 279–338

[27] Buchner D, Beermann A J, Laini A, Rolauffs P, Vitecek S, Hering D et al. 2019 Analysis of 13,312 benthic invertebrate samples from German streams reveals minor deviations in ecological status class between abundance and presence/absence data PLoS ONE 14(12) e0226547. DOI: 10.1371/journal.pone.0226547

[28] Bert T M, Seyoum S, Tringali M D and Mcmillen-Jackson A 2002 Methodologies for conservation assessments of the genetic biodiversity of aquatic macro-organisms Braz. J. Biol. 62(3) 387–408. DOI 10.1590/S1519-69842002000300002.