Journal of
open research software

Ecoindicators: An R Package for the Identification of Indicator Taxonomic Units

SOFTWARE METAPAPER

]u[ubiquity press

ANDRÉS ESTEBAN DUHOUR
HERNÁN DE LA VEGA
LILIANA FALCO
CARLOS COVIELLA
NICOLÁS VELAZCO
ROSANA SANDLER
MACARENA RIONDA
MÓNICA DÍAZ PORRES
LEONARDO SARAVIA

*Author affiliations can be found in the back matter of this article

ABSTRACT

The sensitivity of some taxonomic groups to environmental conditions allows for their use as ecological state indicators. In ecological communities, this analysis is of interest to know which species or taxa are indicators of particular environment conditions, taking into account their presence in a set of sampling units.

There are different approaches to estimate indicator species, in this work we develop a new tool for this task identifying indicator units defined as species (or taxa) whose observed frequencies are not uniform between environments or groups of sample units. The package provides methods to estimate the indicator species, the belonging of new samples to a given environment, and returns the frequency of taxonomic units in a selected combination of environmental factors as an estimation of the ecological niche. EcoIndicators was tested on a data set provided with the package and on independent samples of soil fauna communities and is freely available in https://github.com/lsaravia/EcoIndicators.

CORRESPONDING AUTHOR:
Andrés Esteban Duhour
Departamento de Ciencias Básicas, Universidad Nacional de Luján, Argentina; Instituto de Ecología y Desarrollo Sustentable (UNLu – CONICET), AR
aduhour@unlu.edu.ar

KEYWORDS:
Indicator species; ecological community analysis; samples grouping estimation; ecological niche; soil fauna

TO CITE THIS ARTICLE:
Duhour AE, de la Vega H, Falco L, Coviella C, Velazco N, Sandler R, Rionda M, Díaz Porres M, Saravia L 2023 Ecoindicators: An R Package for the Identification of Indicator Taxonomic Units. *Journal of Open Research Software,* 11: 11. DOI: https://doi.org/10.5334/jors.452

(1) OVERVIEW

INTRODUCTION

Ecological communities are affected by anthropogenic impacts, and the resulting disturbances produce changes in the abundance, distribution patterns or behaviour of species [10, 11].

These changes can be used to assess the ecological status of an environment in order to provide early warning signals for environmental management. As a consequence, it is essential to develop tools to detect species or groups of species that reflect environmental changes and could be used as bioindicators [1, 8, 9, 15].

Among different techniques to estimate ecological indicators, the analysis of indicator species is a relevant tool to study which species contribute most to the differences in species composition of the respective units [13].

Several methods are mentioned in the bibliography, mainly for the analysis of plant communities. Among them, the Indicator Value, Indval [7] combines the species` relative abundances with their relative frequency. Chytrý et al. [2], propose correlation indices to evaluate species` preference in a group of sample units. Ricotta et al. [12] consider the functional characters of species, while De Cáceres et al. [4], suggest using a combination of indicator species.

The R package labdsv [14], oriented to ordination and multivariate analysis for ecology, has implemented the indicator value method proposed by Dufrêne and Legendre [7]. A generalization of this method and the correlation indices of species preferences have been included in the R package indicspecies [3].

The purpose of this work is to advance in the design of tools and methodologies that could be used for the construction of ecological system state indices. We present the R package Ecoindicators [5], to automatically classify ecosystems descriptors and estimate grouping of samples, using only the presence and absence of taxonomic units as an implementation of the method developed by de la Vega et al. [6] for the selection of indicator species. We show a comparison with the indicator value method implemented by the function `indval` (labdsv package).

IMPLEMENTATION AND ARCHITECTURE

The Ecoindicators package was developed using a data set -provided with the package- consisting of soil physico-chemical properties and soil fauna abundance of three sites in the Pampean plain (Buenos Aires, Argentina). Samples were taken on three different sites with different use intensity: A naturalised grassland (NG), a grazing field that changed to agriculture two years before the start of the samplings (CG), and a site of continuous intensive agriculture for at least 40 years (AG). The data set comprises measurements of fifteen physical and chemical soil parameters and the abundance of forty-three soil fauna taxonomic units [16]. Samples were taken at the same time in each date and sampling unit.

The package consists of a set of functions (Table 1) to identify indicator taxonomic units (`select_indicator_species`), and assign according to the species composition (`identify_env`) a new sample to a given environment. It also shows the presence of a species in a hypercube of chosen environmental variables, as an estimate of its ecological niche (`sp_hypercube`). A flow chart describing the input data and the purpose of each function is shown in Figure 1.

Selection of indicator units

As a first step, and based on their frequency of appearance in each system, the package selects "indicator" taxonomic units. The rationale is, if the occurrence (O_i^j) of each taxonomic unit *i* were independent of the environments *j*, it would be expected that the proportion of appearances of each taxonomic unit in each environment was uniform. Then we test this hypothesis using the χ^2 distribution, assuming groups with equal number of samples:

$$\chi^2 = \Sigma_{j=1\ldots n}(O_i^j - 1/n)^2 / (1/n)$$

where n is the number of environments or groups of samples. If the taxonomic unit is not an indicator

FUNCTION TYPE	NAME	DESCRIPTION
Data	`soilandfauna`	Soil physico-chemical properties and soil fauna of three sites in the Pampean plain
Basic use	`select_indicator_species`	Selection of species or taxonomic units that indicate with a given probability of belonging to an environment
Basic use	`identify_env`	Identify the environment from new samples of community species
Auxiliary	`identify_env_test`	Tests the accuracy of the environment identification of a set of samples
Auxiliary	`as_niche_factor`	Converts a vector or matrix of quantitative environmental data into a factorized version with a desired number of partitions.
Basic use	`sp_hypercube`	Returns the frequency of any taxonomic unit in a selected combination of environmental factors

Table 1 Functions included.

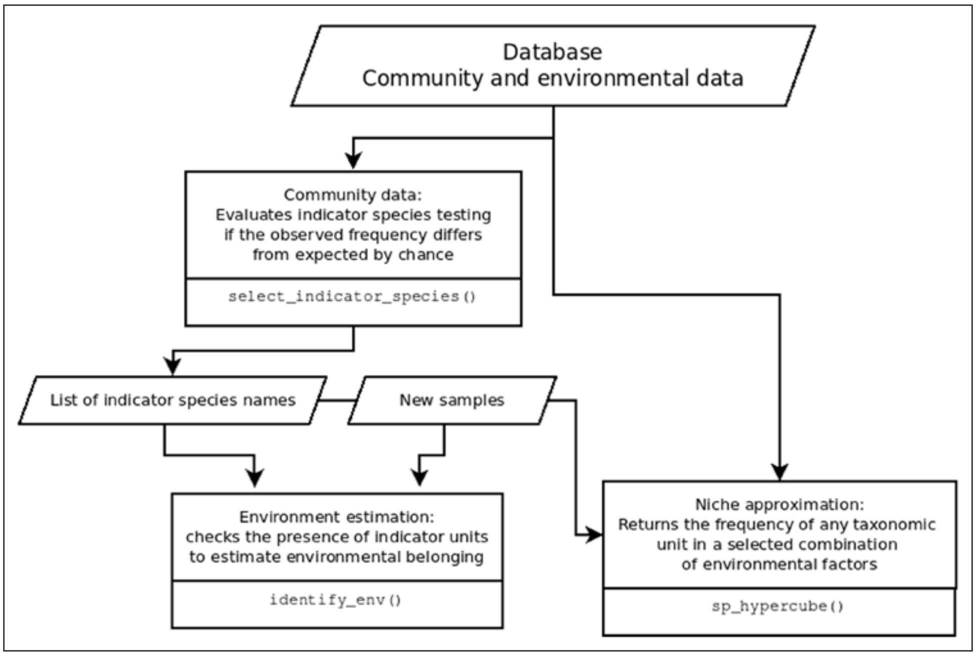
Database
Community and environmental data
Community data:
Evaluates indicator species testing
if the observed frequency differs
from expected by chance
select_indicator_species()
List of indicator species names
New samples
Environment estimation:
checks the presence of indicator units
to estimate environmental belonging
identify_env()
Niche approximation:
Returns the frequency of any taxonomic
unit in a selected combination
of environmental factors
sp_hypercube()

Figure 1 Main functions and flowchart for basic usage.

species, its distribution in the environment does not differ from that expected by chance. The function `select_indicator_species` takes as arguments a community data matrix with species in columns and samples in rows, a vector of the sample's grouping of the community data and the significance level α used for the chi-square test. The function returns a list with two components: a vector of the name of the indicator species selected and a vector of the conditional probability of each species or taxonomic unit in the data.

Environment estimation

Once the indicator units have been obtained, their presence in a new set of samples is used to determine to which particular environment it could belong.

The function `identify_env()`, verifies whether the observed proportion of indicator units in each sample group of the original community differs from expected. Then, construct a matrix of coefficients that add or subtract probabilities that a new sample belongs to a certain environment. This coefficients matrix and the vector of frequencies of indicator units in the new sample are multiplied, returning a vector of group estimations whose highest element indicates the environment the procedure is looking for. The function takes as arguments a community matrix, the result of the function `select_indicator_species`, a grouping vector and a significance level for the test.

The function `identify_env_test` uses a subsampling method to test the accuracy of the group estimation.

Ecological niche representation

Regarding the link of biological and environmental data, the interest is more often focused on relating only a limited number of physical and chemical parameters with only some of the taxonomic units. The function `sp_hypercube` makes a niche approximation returning the abundance of a species in a *n* by *n* grid of environmental variables. Each numerical variable is converted into a categorical variable, classifying each element according to the interval in which it is found.

Finally, a contingency table is constructed that indicates the frequency of samples in which a given taxonomic unit is present in each combination of environmental variables initially defined.

The function takes four arguments, a matrix or data frame of environmental variables, a vector or matrix of species abundances, the number of partitions in which to divide the range of each environmental variable, and a logical variable indicating, in case of param *sp* has two or more species, whether the joint or alternative presence of the species should be considered.

The objective consisted in creating a tool to relate the presence of taxonomic units (indicator units) in the samples with the levels of certain physical and chemical parameters of interest.

EXAMPLE ANALYSIS AND USAGE

The package can be installed from its github repository running the command: `devtools::install_github('lsaravia/EcoIndicators')` in the R prompt.

Selecting indicator species

```
data(soilandfauna)
com <- soilandfauna[,18:60] # Select
community (species) data
group <- soilandfauna[,1] # Select grouping
factor
```

```
indicsp <- select_indicator_
species(com,group,0.01)
paste(indicsp$names, collapse = ", ")
# [1] "Hypogastruridae, Onychiuridae,
Rhodacaroidea, Parasitoidea, Veigaioidea,
Euphthiracaroidea, Aporrectodea_caliginosa,
Microscolex_dubius, Eukerria_stagnalis"
```

Identifying environment or sample group

```
  subcom <- com[3:10,] # Select a subset of
  samples to test the estimation
# Estimate the indicator species present
  indicsp <- select_indicator_species(com,
  group)
  idenv <- identify_env(subcom, indicsp,
  group)
  idenv$belonging.env
# [1] "AG"
# Test the accuracy of environment estimation

identify_env_test(com, group) #group
accuracy
# NG 0
# CG 0.9379379
# AG 0.998999
```

The analysis performed in the *soilandfauna* dataset found nine indicator species (α = 0.01). Regarding the environment estimation the function gives an accuracy higher than 90% for the agriculture (AG) and grazing sites (CG).

Abundance of a species in a grid of environmental variables

The output of the `sp_hypercube` function is a table that shows the samples' frequencies in each cell of the grid. Figure 2 shows the dispersion of the *Onychiuridae* species in a space of three environmental variables: available phosphorus (ppm, P), organic matter (%, OM) and Kjeldahl nitrogen (%, N).

```
# Frequency of any taxonomic unit in a
selected combination
# of environmental factors
# Select environmental data
env <- soilandfauna[,3:17]
# Obtaining the presence of the Onychiuridae
species in a 3 × 3 grid
sp_hypercube(env[,c("P","OM","N")],
com[,"Onychiuridae"],3)

##    N [0.14,0.263] (0.263,0.387] (0.387,0.51]
## P  OM

## [0,25.3] [1.51,4.08]     16        15        4
## (4.08,6.66]              16        40        6
## (6.66,9.23]              0         6         4
## (25.3,50.5] [1.51,4.08]  3         7         0
## (4.08,6.66]              1         0         0
## (6.66,9.23]              1         0         0
## (50.5,75.8] [1.51,4.08]  0         0         0
## (4.08,6.66]              0         1         0
## (6.66,9.23]              0         0         0
```

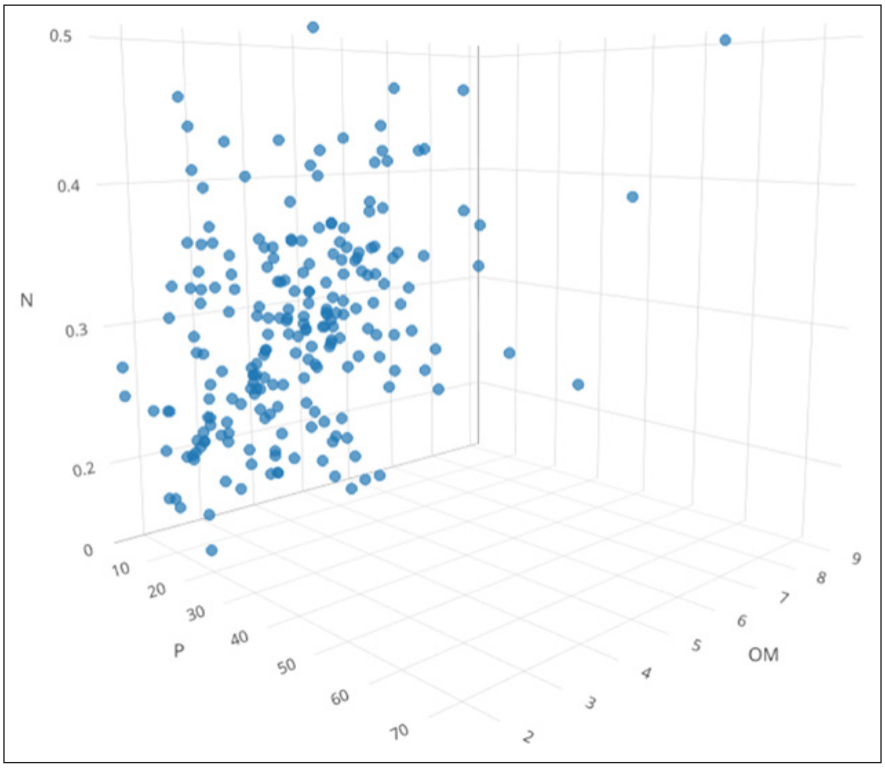

Figure 2 3D scatter plot of the presence of the *Onychiuridae* taxon regarding the environmental variables N, P, and OM.

The *Onychiuridae* taxon is more frequent in intermediate values of N and OM and in the lower values of P.

Then, the simultaneous appearance of the units "*Onychiuridae*", "*Isotomidae*", "*Eupodoidea*", and "*Aporrectodea rosea*" is detected two times in the cube delimited by 0 ≤ P ≤ 25.26, 4.08 ≤ OM ≤ 6.66 and 0.26 ≤ N ≤ 0.39 (Figure 3).

```
# sp_hypercube(env[,c("P","OM","N")],
      com[,c(com[,c("Onychiuridae","Isotomidae",
      "Eupodoidea",
      "Aporrectodea_rosea")],5)

## N [0.14,0.263] (0.263,0.387] (0.387,0.51]
## P         OM
## [0,25.3]    [1.51,4.08]   0       0       0
##             (4.08,6.66]   0       2       0
##             (6.66,9.23]   0       0       0
## (25.3,50.5] [1.51,4.08]   0       0       0
## showing only the first four lines
```

With the help of the functions provided in the EcoIndicators package it is possible to evaluate the indicator species of a community data matrix, and to take a step forward by providing a methodology to, from this information, predict the possible type of environment to which a new set of samples might belong. In addition, by making it possible to obtain the frequencies of a taxon or combination of taxa in an *n*-dimensional space of environmental parameters, the tools developed allow for a more precise analysis of the relationships between biological and environmental components.

COMPARISON OF ECOINDICATORS AND INDVAL METHODS FOR INDICATOR SPECIES ANALYSIS

We analysed a data set collected in the northeast of the Buenos Aires province, in experimental plots of the National University of Luján (UNLu) and in rural plots of the localities of Open Door, Cortines and General Rodríguez (Buenos Aires, Argentina). Sampling was carried out seasonally from 2008 to 2014.

Different sites were chosen according to their use history: (A), with a history of use between 17 and 31 years applying conventional tillage and reduced tillage; (G) livestock soils with a use history of 15 years with sheep and cattle grazing: (N) Naturalized soils, with no agricultural or livestock use for at least 30 years. A total of 99 sampling units of each type of use were analysed for the evaluation of indicator species.

The indicator species analysis was carried out with EcoIndicators. In turn, the calculation of the indicator value [7] was performed with the `labdsv` package [14]. Those species that showed a *p* value < 0.05 were selected as indicator species.

Applying both methods, results coincided in nine indicator species, finding a total of 10 species with a value of $p < 0.05$. The data and code for this analysis could be found as supplementary material to the source code of the package.

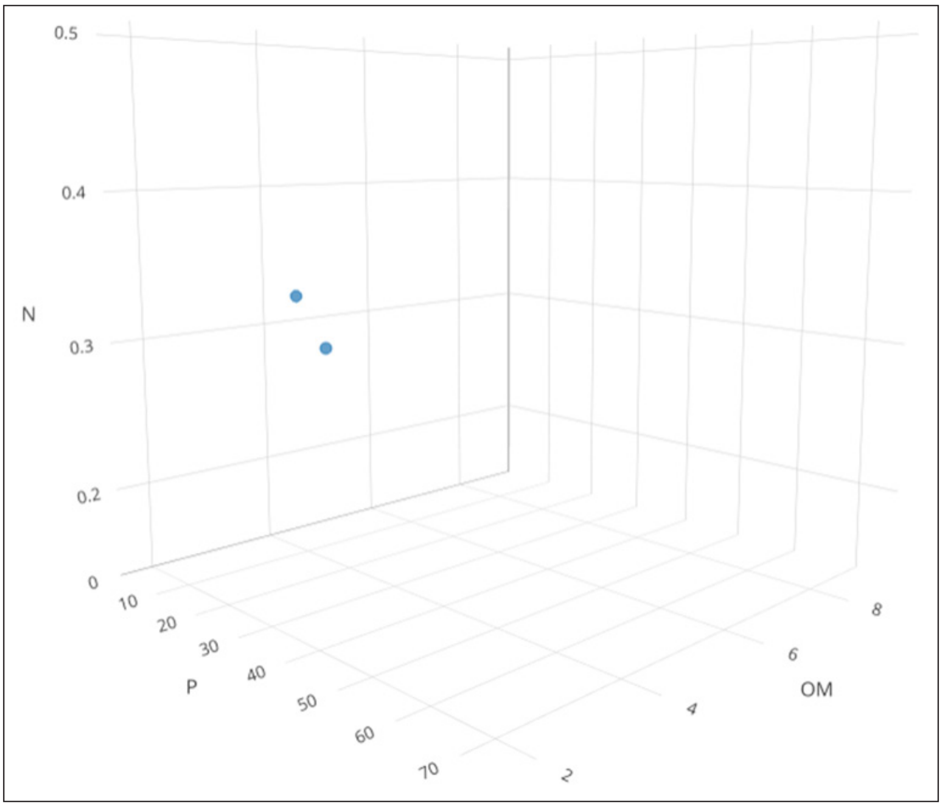

Figure 3 3D scatter plot of the simultaneous occurrence of taxonomic units *Onychiuridae*, *Isotomidae*, *Eupodoidea*, and *Aporrectodea rosea*.

QUALITY CONTROL

The functions provided with EcoIndicators were evaluated to test if they produce the desired output. The workflow was tested on soil environment and faunal data -provided as a dataset with this package-. Additionally, functions were tested with samples taken in a different experiment from those used to build the package.

The structure of the package successfully passed the CRAN R CMD check with no errors, warnings, or notes.

(2) AVAILABILITY

OPERATING SYSTEM

The package was tested on Linux and Windows.

PROGRAMMING LANGUAGE

R version 4.1.2 or higher

ADDITIONAL SYSTEM REQUIREMENTS

There are no additional requirements.

DEPENDENCIES

R package: stats.

LIST OF CONTRIBUTORS

The package was created by Andrés Duhour, Hernán de la Vega and Leonardo Saravia.

SOFTWARE LOCATION

Archive

Name: Zenodo
Persistent identifier: https://doi.org/10.5281/zenodo.7026224
Licence: GPL-2
Publisher: Leonardo Saravia
Version published: 1.0.0
Date published: 26/08/2022

Code repository

Name: Github (https://github.com/lsaravia/EcoIndicators)
Identifier: https://doi.org/10.5281/zenodo.7026224
Licence: GPL-2
Date published: 26/08/22

Language

R

(3) REUSE POTENTIAL

The functions use Roxygen to generate documentation that can be asked with the R help function. EcoIndicators can be integrated with other ecosystem data analysis frameworks, enabling a new approach to ecosystem studies.

The EcoIndicators package was originally developed using a soil biota database [16], and it will work just the same in any other environment or ecosystem for which there is a database of species abundances associated with an environmental dataset. The package provides a new method for the selection of indicator species and allows to directly use this result to identify the belonging of new samples to one of the environments identified in the original database. In addition, the ecological niche approach helps to represent the interrelationships between ecological communities and environmental variables.

The package is hosted in a public repository with version control, allowing contributors to add new features. It is possible to report issues and suggest improvements via Github or by contacting the corresponding authors. The package is a starting point for the improvement of the methods presented here and for comparison with the other available methods.

ACKNOWLEDGEMENTS

The authors wish to acknowledge the collaboration of Agr. Eng. Eduardo Penon and Loreta Gimenez for their field and laboratory assistance. A special acknowledgment goes to Edgardo Ferrari, Pablo Peretto, Carlos Carey, and Romina de Luca for allowing the use of their properties as sampling sites.

FUNDING INFORMATION

This work was partially funded by doctoral grants to Rosana Sandler from Agencia Nacional de Promoción Científica y Tecnológica (ANPCyT), and to Víctor Velazco from Consejo Nacional de Investigaciones Científicas y Técnicas (CONICET). Additional funding was provided by Universidad Nacional de Luján.

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR AFFILIATIONS

Andrés Esteban Duhour orcid.org/0000-0002-5783-9224
Departamento de Ciencias Básicas, Universidad Nacional de Luján, Argentina; Instituto de Ecología y Desarrollo Sustentable (UNLu – CONICET), AR

Hernán de la Vega orcid.org/0000-0002-2483-139X
Departamento de Ciencias Básicas, Universidad Nacional de Luján, AR

Liliana Falco orcid.org/0000-0001-9097-2572
Departamento de Ciencias Básicas, Universidad Nacional de Luján, Argentina; Instituto de Ecología y Desarrollo Sustentable, (UNLu – CONICET), AR

Carlos Coviella orcid.org/0000-0001-9200-5826
Departamento de Ciencias Básicas, Universidad Nacional de Luján, Argentina; Instituto de Ecología y Desarrollo Sustentable, (UNLu – CONICET), AR

Nicolás Velazco orcid.org/0000-0002-5546-0616
Departamento de Ciencias Básicas, Universidad Nacional de Luján, Argentina; Instituto de Ecología y Desarrollo Sustentable, (UNLu – CONICET), AR

Rosana Sandler orcid.org/0000-0003-0702-5094
Instituto de Ecología y Desarrollo Sustentable, (UNLu – CONICET), AR

Macarena Rionda orcid.org/0000-0002-3342-0672
Instituto de Ciencias, Universidad Nacional de General Sarmiento, AR

Mónica Díaz Porres orcid.org/0000-0003-4926-5184
Departamento de Ciencias Básicas, Universidad Nacional de Luján, Argentina; Instituto de Ecología y Desarrollo Sustentable, (UNLu – CONICET), AR

Leonardo Saravia orcid.org/0000-0002-7911-4398
Centro Austral de Investigaciones Científicas, (CADIC – CONICET), AR

REFERENCES

1. **Bedano JC, Domínguez A, Arolfo, R.** Assessment of soil biological degradation using mesofauna. *Soil and Tillage Research*. 2011; 117: 55–60. DOI: https://doi.org/10.1016/j.still.2011.08.007
2. **Chytrý M, Tichý L, Holt J, Botta-Dukát Z.** Determination of diagnostic species with statistical fidelity measures. *Journal of Vegetation Science*. 2002; 13: 79–90. DOI: https://doi.org/10.1111/j.1654-1103.2002.tb02025.x
3. **De Cáceres M, Legendre P.** Associations between species and groups of sites: indices and statistical inference. *Ecology*. 2009; 90: 3566–3574. DOI: https://doi.org/10.1890/08-1823.1
4. **De Cáceres M, Legendre P, Wiser SK, Brotons L.** Using species combinations in indicator value analyses. *Methods in Ecology and Evolution*. 2012; 973–982. DOI: https://doi.org/10.1111/j.2041-210X.2012.00246.x
5. **de la Vega H, Falco L, Saravia L, Sandler R, Duhour A, Coviella C.** EcoIndicators; 2019. DOI: https://doi.org/10.5281/zenodo.5772829
6. **de la Vega H, Falco L, Saravia L, Sandler R, Duhour A, Velazco VN, Coviella C.** An algorithm for the identification of indicator taxonomic units and their use in analyses of ecosystem state. *Revista de Modelamiento Matemático de Sistemas Biológicos*. 2022; 2: 1–10. URL: https://revistammsb.utem.cl/?p=437. DOI: https://doi.org/10.1101/2022.05.16.492087
7. **Dufrêne M, Legendre P.** Species assemblages and indicator species: the need for a flexible asymmetrical approach. *Ecological Monographs*. 1997; 67: 345–366. DOI: https://doi.org/10.1890/0012-9615(1997)067[0345:SAAIST]2.0.CO;2
8. **George PB, Keith AM, Creer S, Barrett GL, Lebron I, Emmett BA, Robinson DA, Jones DL.** Evaluation of mesofauna communities as soil quality indicators in a national-level monitoring programme. *Soil Biology and Biochemistry*. 2017; 115: 537–546. DOI: https://doi.org/10.1016/j.soilbio.2017.09.022
9. **Guerra CA, Bardgett RD, Caon L, Crowther TW, Delgado-Baquerizo M, Montanarella L, Navarro LM, Orgiazzi A, Singh, BK, Tedersoo L, Vargas-Rojas, R, Briones MJI, Buscot F, Cameron EK, Cesarz S, Chatzinotas A, Cowan DA, Djukic I, van den Hoogen J, Lehmann A, Maestre FT, Marín C, Reitz T, Rillig MC, Smith LC, de Vries FT, Weigelt A, Wall DH, Eisenhauer N.** Tracking, targeting, and conserving soil biodiversity. *Science*. 2021; 371(6526): 239–241 DOI: https://doi.org/10.1126/science.abd7926
10. **El Mujtar V, Muñoz N, Prack Mc Cormick B, Pulleman M, Tittonell P.** Role and management of soil biodiversity for food security and nutrition; where do we stand? *Global Food Security*. 2019; 20: 132–144. DOI: https://doi.org/10.1016/j.gfs.2019.01.007
11. **Ooms A, Dias A, van Oosten A, Cornelissen J, Ellers J, Berg M.** Species richness and functional diversity of isopod communities vary across an urbanisation gradient, but the direction and strength depend on soil type *Soil Biology and Biochemistry*. 2020; 148: 107851. DOI: https://doi.org/10.1016/j.soilbio.2020.107851
12. **Ricotta C, Carboni M, Acosta AT, Mason NN.** (ed.) Let the concept of indicator species be functional! *Journal of Vegetation Science, Wiley*. 2015; 26: 839–847. DOI: https://doi.org/10.1111/jvs.12291
13. **Ricotta C, Pavoine S, Cerabolini BEL, Pillar VD.** A new method for indicator species analysis in the framework of multivariate analysis of variance. *Methods In Vegetation Science*. 2021; 32(2): e13013. DOI: https://doi.org/10.1111/jvs.13013
14. **Roberts DW.** labdsv: Ordination and Multivariate Analysis for Ecology. R package. version 2.0-1; 2019. https://CRAN.R-project.org/package=labdsv.
15. **Rocha L, Hegoburu C, Torremorell A, Feijoó C, Navarro E, Fernández H.** Use of ecosystem health indicators for assessing anthropogenic impacts on freshwaters in Argentina: a review. *Environmental Monitoring and Assessment*. 2020; 192: 611. DOI: https://doi.org/10.1007/s10661-020-08559-w
16. **Sandler RV.** Indicadores de sustentabilidad del suelo basados en la estructura y funcionamiento de la fauna edáfica. Ph.D. dissertation. Universidad Nacional de General Sarmiento. Argentina; 2019. URL: http://repositorio.ungs.edu.ar:8080/xmlui/handle/UNGS/728.

TO CITE THIS ARTICLE:

Duhour AE, de la Vega H, Falco L, Coviella C, Velazco N, Sandler R, Rionda M, Díaz Porres M, Saravia L 2023 Ecoindicators: An R Package for the Identification of Indicator Taxonomic Units. *Journal of Open Research Software,* 11: 11.. DOI: https://doi.org/10.5334/jors.452

Submitted: 30 December 2022 **Accepted:** 29 August 2023 **Published:** 13 October 2023

Journal of Open Research Software is a peer-reviewed open access journal published by Ubiquity Press.