Numerical Ecology with R, second edition, 2018

This file provides corrections of small mistakes found in the printed edition (Errata), as well as improved explanations and updated, additional information (Addenda and updates).

Last update: 05 March 2025

Errata

P. 56, Table 3.1

Bottom right, last entry. One should read:

Binary variables:
Simple matching coefficient
`dist.binary(., method=2)`

Explanation: function **`dist.binary()`** accepts 0 and nonnegative values only (i.e., no standardized variables). All strictly positive values are converted to 1, and the coefficient `method = 2` computes the (square root of the one-complement of the) ratio between matching pairs (i.e. double zeros and double 1s) and the total number of variables.

P. 78, code

In function **`tanglegram()`**, set `sort` to `FALSE` instead of `TRUE` :

```
tanglegram(
  untangle(dend12),
  sort = FALSE,
```

P. 79, L. 1-2

The second sentence of the paragraph should read: *Colours highlight common clusters, whereas sites **connected by black lines** have different positions in the two trees.*

P. 133, last line of code

The line should read:

```
summary.MRT(spe.ch.mvpart.wrap)
```

Explanation: in this case, the usual **`summary()`** command displays the list of items contained in the result object of an **`MRT()`** run. In contrast, **`summary.MRT()`** has been tailored to display the most interesting results in an accessible format with supplementary results such as discriminant species and indicator values (IndVal).

P.134

The first paragraph should begin as follows:

The **`MRT()`** results are resented in a form close to canonical ordination results […]

The second paragraph should be amended as follows:

The **`summary.MRT()`** display of the result object provides information about…

The two last lines of the second paragraph should be amended as follows:

[…] for the right. The result object itself lists the complete results from which the **`summary.MRT()`** display is drawn.

See also the "Addenda" section below, for p. 134.

P.157

In the fourth paragraph ("A compromise scaling…", Line 4 missing word (in bold hereunder):

representation of site and **species** scores without […]

P. 158, last line of code

A closing bracket is missing at the end of the line, which should read:

```
screeplot(env.pca, bstick = TRUE, npcs = length(ev))
```

P. 173, Caption of Fig. 5.5

The caption should read: "Imputation of missing data in PCA. Procrustes rotation of the original PCA of environmental variables and the one performed on data where ~~three~~ missing values have been imputed. Scaling 1, only the sites are represented. Original sites: red; sites in imputed PCA: blue. Left: 3 missing values, or 1%; right: 32 missing values, or 10%." [delete "three"]

P. 200, code, L.7

A minus sign is missing : (-0.5). The line should read:

```
G <- F %*% diag(Y.eig$values ^(-0.5))
```

P. 220, eq. 6.2

The equation should read (see correction in red):

$$F = \frac{\mathrm{SS}(\hat{\mathbf{Y}})/m}{\mathrm{RSS}/(n-m-1)}$$

P. 226, **R** code block, second comment

The second comment should read (modification underscored):

```
# Partial RDA - effect of water chemistry, controlling for physiography
```

P. 265, **R** code block, last comment

The last comment is ambiguous because it uses the word "object" in two different meanings. It should read:

```
# Classification of two new objects (identification)
# A new data frame is created with two sites:
```

P. 308 bottom and p. 309 top

The description of the Holm correction is incorrect. Here is the correct procedure, as in Legendre & Legendre (2012, p.23) :

(1) Order the p-values from left to right, so that $p_1 \leq p_2 \leq \ldots \leq p_i \ldots \leq p_k$
(2) Compute adjusted $p_i' = (k - i + 1) \times p_i$; adjusted probabilities may be larger than 1
(3) Proceeding from left to right, if an adjusted p-value in the ordered series is smaller than the one occurring at its left, make the smallest equal to the largest one
(4) Replace the adjusted p (i.e., p') in the order of the initial vector
(5) If any $p > 1$, make it equal to 1
(6) Compare the adjusted p_i' to the unadjusted α significance level and make the statistical decision.

P. 345 line 2

Replace the value 1.0011188 by 1.011188 .

P. 372, bottom

The text should read:

where q is the number of species. When n is large, n becomes close to $(n - 1)$, n_i becomes close to $(n_i - 1)$ and the equation simplifies to:

P. 385, just below eq. 8.15

The sentence and formula should read:

Note that $\mathrm{SS}_{\mathrm{Total}} = \sum_{i=1}^{n} \mathrm{SS}_i$

P. 386, just below eq. 8.16

The sentence and formula should read:

Again, $\mathrm{SS}_{\mathrm{Total}} = \sum_{i=1}^{n} \mathrm{SS}_i$

P. 394, end of paragraph "Minimum value"

Complete the lase sentence as follows:

and Podani's nestedness sum to 1 when $b = c$ (cases 2 and 9).

P. 394 last paragraph and p. 395 first paragraph

The text in the manual is not wrong. However: the following replacement text better conveys the authors' intent:

Maximum value – The Podani and Baselga nestedness indices are not based on the same definition of nestedness. The framed Podani index can reach a value of 1, in particular when the two sites are totally similar, i.e. if $b = c = 0$ and $a > 0$. On the other hand, since its index takes the value 0 when $b = c = 0$, Baselga considers that at least one unique species must be present in either site, in addition to the common species (condition: $b > 0$ or $c > 0$). Furthermore, the two sites cannot host the same number of unique species; this is why the numerator of the left-hand side of the equation has been designed so that the value of Nes_{BJ} is 0 when $b = c$ (cases 2 and 9). The Baselga index peaks at the Jaccard D_J dissimilarity when b or c is 0 (case 8). This highlights an important difference between the two definitions of nestedness: Podani and Schmera (2011)

propose that *a* is a major component of nestedness if $a > 0$, to the point of considering that two totally similar sites are completely nested. Baselga, on the other hand, places more emphasis on the difference in unique species, $|b - c|$, with a modest contribution from *a*. Furthermore, for fixed *b* and *c*, Podani's nestedness increases monotonically with *a*, while Baselga's nestedness increases up to a certain maximum and then decreases (cases 3, 4 and 5).

It is up to the Reader to decide which of the two definitions best suits his or her research objectives. If one looks for an index that focuses primarily on the number of common species rather than unique species, Podani and Schmera's index (2011) is the appropriate choice. If one prefers an index that focuses more on the $|b - c|$ difference in unique species, one will choose Baselga's index (2010, 2012).

P. 397, L. -9

Correct site numbers as follows:

[…] then drops at sites 7, 6 and 5, […]

P. 398, R code

Add the red code line below (present in the script but inadvertently dropped from the book version):

```
fish.jac.neigh <- diag(fish.jac[-1, ]) # Jaccard Dⱼ index
absc <- c(2:7, 9:30) # Abscissa
```

Addenda and updates

The following entries are proposed to follow changes in recent versions of **R** and packages, to improve explanations or to add some recent pieces of information.

P. 27, text and code at bottom of page

Replace **`vegtrans()`** by **`abundtrans()`**.

Explanation: in package **`labdsv{}`** the name of function **`vegtrans()`** has been replaced by **`abundtrans()`**. The function is the same.

P. 42, second paragraph, second sentence

Complete the sentence as follows (addendum underscored):

The exercise consists in computing several dissimilarity matrices based on appropriate similarity coefficients: the Jaccard (S_7), Sørensen (S_8) and Ochiai (S_{14}) similarities.

P. 134, above the title of Sect. 4.12.4

Add the following text:

This is how the main results contained in the `spe.ch.mvpart.wrap` object we have just produced are provided. Users can extract the data most relevant to their scientific questions.

- `$pourct` : node × species matrix giving the contribution of each species to the R^2 of the analysis, expressed as a percentage for each node; the sum of each row (i.e., each node) is equal to 100.
- `$R2` : node × species matrix giving the contribution to the R^2 of each species, at each node; the sum of each row is equal to the contribution of each node to the global R^2; the sum of each column is equal to the total contribution of each species to the tree's explanation. The sum of this matrix is equal to the global R^2.
- `$MOYs` : branch × species matrix giving the average abundance of each species in the sites of the branch considered. In our example, there are 6 branches, 4 of which are terminal (leaves), not counting the tree root. Branches numbered 1 and 2 are those of the first node. 3 and 4 are the branches (and leaves) of the left-hand subdivision. 5 and 6 are those of the right-hand subdivision.
- `$RWhere` and `$LWhere` : line numbers of the sites carried by the right (R) and left (L) branches of each node.
- `$TABLE1` : please note that the last row of this table contains the column totals. For the rest:
    - columns 1 to 3: contribution of each species to the R^2 of the analysis for each explanatory variable at the thresholds used in the analysis. The sum of the totals of these three columns is the R2 of the analysis expressed as a percentage;
    - column 4: sum of previous columns; total contribution of each species to the explanation of the tree. The total of this column is the R^2 of the analysis expressed as a percentage; in our example, this value is 62.95%. The R^2 is therefore equal to 0.63, i.e. the 1-complement of the tree's RE, which is equal to 0.37 (Fig. 4.27);
    - column 5: contribution of each species to the total variance of the data, expressed as a percentage.

P. 141, before the title of Sect. 4.15

Add the following text:

A new clustering method with spatial or temporal contiguity constraint, based on the general agglomerative clustering algorithm of Lance and Williams (1967), is described in a paper by Guénard and Legendre (2022). The method is implemented in function **`constr.hclust()`** of the **`adespatial`** package in **R**. As graphical output, the function produces a series of maps, each one corresponding to a clustering level; a

dendrogram can also be produced using the **plot.hclust()** function of the **stats** package. The method is summarized in a teaching document by Legendre (2021).

The dendrograms produced after clustering by functions **constr.hclust()** of package **adespatial** and **chclust()** of package **rioja** differ in several respects, described here.

• For Ward clustering, functions **hclust()** of **stats** and **constr.hclust()** of **adespatial** implement by default the ward.D2 algorithm, which follows the Ward (1963) minimum-variance clustering criterion, whereas function **chclust()** applies the ward.D algorithm, which does not implement that criterion. The difference between these two algorithms was described by Murtagh & Legendre (2014). It is also briefly described in the documentation files of functions **hclust()** and **constr.hclust()**. All hierarchical clustering strategies compatible with Lance & Williams (1967) general clustering algorithm are implemented in **constr.hclust()**, as they are in **hclust()**. For comparison with **chclust()** output, one can produce ward.D results obtained with **constr.hclust()**, although that is not the recommended option, and these results should be comparable to those of **chclust()**, except for the presentation of the dendrograms.

• Another difference is that the **chclust()** function can only handle one-dimensional study designs (spatial transect or time series), whereas **constr.hclust()** can handle spatial transects, time series, and sites observed on a two-dimensional map. It could also analyse survey designs in three spatial dimensions, although no plotting function is currently available to represent the clustering results of 3D data.

• The author of the **rioja** package used a special plotting function, **plot.chclust()**, for representing constrained clustering results along spatial transects or time series in the form of dendrograms. A dendrogram is the usual way of representing hierarchical clustering results. That function uses the ordered sequence of observations along the spatial transect or the time series to position the sites or time points along the abscissa of the plot. This improves readability and facilitates interpretation of the constrained clustering result. The reversals, which are an inherent part of constrained clustering results, are not, however, represented in **plot.chclust()** dendrograms. To achieve this, the function uses a modified dendrogram algorithm described by Grimm (1987).

• One can also produce a dendrogram from a constrained clustering result computed by function **constr.hclust()**, which has classes {"constr.hclust", "hclust"}, although that is not the preferred representation. For that purpose, one must use the **plot.hclust()** function of stats as follows : `stats:::plot.hclust(constr.hclust.output.file, …)`. That function correctly represents the reversals in the hierarchical clustering result but it does not necessarily align the objects in their natural sequence along the spatial transect or time series of the study that has generated the data.

• No effort has been made by the **constr.hclust()** authors to write a special dendrogram-plotting function because the normal plotting output of that function is the **plot.constr.hclust()** function, which produces a series of maps, selected by specifying the number of clusters to represent on the map. These maps represent the different groups graphically using symbols, colours, and other graphical arguments. When the data have been collected along a spatial transect or a time series, the sites are positioned in their natural order on the map. When the data come from a 2-dimensional geographic area, the constrained clustering maps are two-dimensional.

• Constrained clustering results often produce reversals in dendrograms because the nestedness (or ultrametric property) of the hierarchical clustering result, which tries to implement Ward's minimum-variance criterion, is often violated by the priority given by the algorithm to the spatial or temporal contiguity constraint. The ultrametric property of nested hierarchical clustering results is described in Legendre & Legendre (2012), Section 8.3; reversals are explained in Section 8.6. The **plot.chclust()** function available in package **rioja** to plot the results of constrained clustering does not represent these reversals. Instead, it represents the dendrogram as if the clustering was ultrametric, to facilitate interpretation of the results.

When comparing a Ward constrained clustering dendrogram computed with function **plot.chclust()** to one produced by function **constr.hclust()** with the `ward.D` option, one must be aware that the

`plot.chclust()` dendrogram will present the observations in their natural spatial or temporal order, which is good for interpretation, whereas the **`constr.hclust()`** dendrogram will not; and that reversals will be shown in the **`constr.hclust`** dendrogram but not in the **`plot.chclust`** dendrogram. Otherwise, the topology of the two dendrograms, except for reversals, will be the same. On the other hand, only the **`constr.hclust()`** function can produce clustering results implementing Ward's (1963) minimum-variance criterion (`method = "ward.D2"`) in a constrained clustering context, and present the clustering results in the form of one- or two-dimensional maps.

References

Grimm, E. C. CONISS: A FORTRAN 77 program for stratigraphically constrained cluster analysis by the method of incremental sum of squares. *Computers & Geosciences* 13: 13-35 (1987).

Guénard, G. and P. Legendre. Hierarchical clustering with contiguity constraint in R. *Journal of Statistical Software* 103(7): 1–26 (2022). https://doi.org/10.18637/jss.v103.i07.

Lance, G. N. and W. T. Williams. A general theory of classificatory sorting strategies. I. Hierarchical systems. *Computer Journal* 9: 373–380 (1967).

Legendre, P. Space-constrained hierarchical clustering for the iAtlantic workshop. Teaching document, iAtlantic Ocean Time Series Workshop / Université de Montréal. 9 pp. Available on http://numericalecology.com/documents_enseignement/Spaceconstrained_ hierarchical_clustering.pdf (2021).

Legendre, P. and L. Legendre. *Numerical ecology, 3rd English edition*. Elsevier Science BV, Amsterdam (2012).

Murtagh, F. and P. Legendre. Ward's hierarchical agglomerative clustering method: which algorithms implement Ward's criterion? *Journal of Classification* 31: 274-295 (2014). DOI: 10.1007/s00357-014-9161-z.

Ward, J. H. Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association* 58: 236–244 (1963).

P. 146, text

4.15.2 Noise clustering using the `vegclust()` function

Original paragraph in the published book:

A recent package called **`vegclust`,** developed by Miquel De Cáceres, provides a large [……] in the "Noise" cluster.

Replace by the following, more detailed paragraph:

A recent package called **`vegclust`**, developed by Miquel De Cáceres, provides a large range of options to perform non-hierarchical or hierarchical fuzzy clustering of community data under different models (De Cáceres et al. 2010). An interesting one is called "noise clustering" (Davé and Krishnapuram 1997). This method is an attempt to make fuzzy clustering more robust to outliers. Outliers are defined as follows: once "true cluster" centroids have been defined, capture into a fictitious "Noise" cluster the objects that lie farther than a distance δ from the "true cluster" centroids (De Cáceres et al. 2010). The choice of the value of δ is critical: too small a δ value results in an overly large number of outliers, i.e., a large membership in the "Noise" cluster. Note also that if a cluster has a larger intragroup dispersion than the others, this increases the likelihood that some of its legitimate members be considered as outliers.

P. 169, first block of text (below the **R** code)

Original text:

`envfit()` also proposes permutation tests to assess the significance of the R^2 of each explanatory variable regressed on the two axes of the biplot. But this is not, by far, the best way to test the effect of explanatory variables on a table of response variables. We will explore this topic in Chap. 6.

Replace by the expanded text below:

`envfit()` also proposes permutation tests to assess the significance of the R^2 of each explanatory variable regressed on the two axes of the biplot. The R^2 statistics (noted `r2` in the **`envfit()`** output) are produced for quantitative explanatory variables and for factors. They measure the fit of the data to the explanatory variables. With the default option `choices = c(1, 2)`, only the first two axes of the ordination are considered and the R^2 measures the fit of the data ordinated in two dimensions to each explanatory variable. If the calculation is made to involve all dimensions of a PCA ordination (this can be obtained by changing the values in argument `choices`), the R^2 statistic measures the fit of the full-dimensional data to the explanatory variables. If the ordination was produced by PCoA (Sect. 5.5) or NMDS (Sect. 5.6) of a dissimilarity matrix, the fit is between the response variables and the data transformed by the dissimilarity index used in the ordination. If the ordination is a PCA and the envfit analysis involves all PCA axes, the R^2 is identical to that produced by **`adonis2()`** (Chap. 6). Note, however that function **`envfit()`** has not been designed to replace this other function, which was designed for multivariate analysis of variance by RDA (Sect. 6.3.2.9); its role is to draw explanatory variables onto simple ordination plots.

P. 234, last text paragraph

Original text:

The three RDAs can be tested as usual, and fractions [a] and [c] can be computed and tested by means of partial RDA. Fraction [b], however, is not an adjusted component of variance and cannot be estimated and tested by regression methods. It has zero degree of freedom. […]

Expanded text:

The three RDAs can be tested as usual, and fractions [a] and [c] can be computed and tested by means of partial RDA. Fraction [b], however, is not an adjusted component of variance and cannot be estimated and tested by regression methods. It has zero degree of freedom. However, an elegant workaround has been devised by Bauman et al. (2018) in the special case of the shared space-environment relationship (i.e., the [b] fraction of two explanatory matrices, one of them modeling spatial structures using methods such as MEM variables [Sect. 7.4]), by means of special permutation procedures based on the spatial layout of the sampling units (torus translations and, in the case of irregular sampling, Moran spectral randomization, Wagner and Dray 2015).

Additional references:

Bauman, D., Vleminckx, J., Hardy, O. J., Drouet, T.: Testing and interpreting the shared space-environment fraction in variation partitioning analyses of ecological data. Oikos **128**, 274-285 (2018)

Wagner, H.H., Dray, S.: Generating spatially constrained null models for irregularly spaced data using Moran spectral randomization methods. Methods in Ecology and Evolution **6**, 1169-1178 (2015)

P. 234, R code

Depending on the version of package {spdep}, due to a reversal in the recording of the two dimensions of the grid (despite the absence of change in the names of the arguments), the line of code involving function **`cell2nb`** must be adapted:

```
if(packageVersion("spdep") < 0.8 {
nb <- cell2nb(4, 10, "queen")
} else {
nb <- cell2nb(nrow = 10, ncol = 4, type = "queen")
# or, equivalent
# nb <- cell2nb(4, 10, "queen", legacy = TRUE)
}
```

If the code above return error messages, use the following:

```
nb <- cell2nb(nrow = 10, ncol = 4, type = "queen")
```

P. 281

The two first paragraphs can be clarified as follows:

The co-inertia analysis represents a compromise between the PCA of **X** (which looks for a combination of variables in **X** with maximum variance), the PCA of **Y** (which looks for a combination of variables in **Y** with maximum variance), and the analysis of the covariance matrix crossing the variables in the two data tables (co-inertia analysis proper). This compromise is obtained by maximising the covariance on each axis, i.e. the product of the standard deviations (sd) of each ordination axis and their correlation: `covar = corr * sdX * sdY`.

The first block of the numerical output can be interpreted as follows. Let's take the example of the first line, which concerns axis 1 of the co-inertia analysis. `eig` (6.78...) is the eigenvalue of the first axis of the CoIA. `covar` (2.60...) is the covariance between axis 1 from the PCA of **X** and axis 1 from the PCA of **Y**. `sdX` (1.99...) and `sdY` (1.63...) are the standard deviation of the PCA axes of **X** and **Y**, respectively. `corr` (0.79...) is the Pearson correlation between axis 1 of the PCAs of **X** and **Y**: `corr = covar / (sdX * sdY)` = 2.60... / (1.99... × 1.63...) = 0.79...

In the second block of the output, `inertia` represents the amount of variance (inertia) of **X** involved in co-inertia, the second column (`max`) shows the maximum possible, i.e. the inertia of the first axis of the PCA of **X**, and `ratio` is the ratio between these two quantities. This ratio is an indication of the link between the structures of the two tables. If it is close to 1, it means that the main structure of **X** is also the one that is most linked to **Y**. A small value for this ratio, on the other hand, would indicate that the structures in **X** linked to **Y** are minor or even random. The third block of the output presents the same results, but from **Y**'s point of view.