00000100000
00001010000
00010001000
00100000100
01000000010
10000000001
Software notes

Ecography 36: 1070–1075, 2013
doi: 10.1111/j.1600-0587.2013.00444.x

Subject Editor: Thiago Rangel. Accepted 31 May 2013

recluster: an unbiased clustering procedure for beta-diversity turnover

Leonardo Dapporto, Matteo Ramazzotti, Simone Fattorini, Gerard Talavera, Roger Vila and Roger L. H. Dennis

L. Dapporto (leondap@gmail.com) and R. L. H. Dennis, Dept of Biological and Medical Sciences, Oxford Brookes Univ., Headington, Oxford, OX3 0BP, UK. RLHD also at: Inst. for Environment, Sustainability and Regeneration, Staffordshire Univ., Mellor Building, College Road, Stoke-on-Trent, ST4 2DE, UK. – M. Ramazzotti, Dipto di Scienze Biomediche Sperimentali e Cliniche – viale Morgagni 50, IT-50134 Firenze, Italy. – S. Fattorini, Azorean Biodiversity Group (CITA-A) and Platform for Enhancing Ecological Research and Sustainability (PEERS), Univ. of Azores, Angra do Heroísmo, Terceira, Azores, Portugal, and Water Ecology Team, Dept of Biotechnology and Biosciences, Univ. of Milano Bicocca, Piazza della Scienza 2, IT-20126 Milan, Italy. – G. Talavera and R. Vila, Inst. de Biologia Evolutiva (CSIC-Univ. Pompeu Fabra), Passeig Marítim de la Barceloneta 37, ES-08003 Barcelona, Spain.

When dissimilarity matrices of faunistic and phylogenetic beta-diversity turnover indices are projected in dendrograms, a high frequency of ties and zero values produces trees whose topology and bootstrap support are affected by the order of areas in the original presence–absence matrix. We tested the magnitude of this bias and developed R functions to obtain consensus trees after shuffling of matrix row order and applied this algorithm to a multiscale bootstrap procedure. Our functions not only solve the bias of row order but, owing to varying support for different bootstrap scales, reveal fundamental characteristics about the structure of species assemblages.

Recently, there has been a renewed interest in the study of beta-diversity metrics, mostly due to developments in partitioning widely used indices into nestedness and turnover components (Baselga 2010, Leprieur et al. 2012). Nestedness is determined by differences among areas in the ordination of species loss, whereas turnover accounts for species replacement (Baselga 2010). Recent worldwide assessments revealed that turnover indices can disclose faunistic and phylogenetic biogeographic structures (Kreft and Jetz 2010, Holt et al. 2013).

Hierarchical clustering facilitates regionalization of communities by converting dissimilarity matrices into bifurcated dendrograms. Bifurcations also occur when an area shows intermediate dissimilarities between others (Legendre and Legendre 1998), as expected when different sources contribute elements to biotas. Support for nodes can be tested by bootstrap methods that re-sample random sets of the original variables (species) to construct a series of trees, and which ultimately search for concordance among subsampled trees and the original tree. Together with classical bootstrap (BP) values, approximately unbiased p-values (AU) can be obtained by multiscale bootstrap. Multiscale bootstrap alters the species number of the re-sampled datasets among different scales so as to change the probability of each species being included in the matrix. The frequency of the sites falling into their original cluster is counted at different scales, and then p-values are obtained by analyzing frequency trends. BP and AU supports can be calculated by the 'pvclust' R package (Suzuki and Shimodaira 2006).

Here, we 1) illustrate how a flaw in clustering methods has important consequences for turnover analysis, 2) present R functions contributing unbiased dendrograms and bootstrap values, and 3) show that multiscale bootstrap can provide important information in biogeography.

We delineated a hypothetical archipelago with two main islands (**A** and **B**), one large and one small island closely related to each main one (**lA**, **sA**, **lB** and **sB**, respectively) and three other completely intermediate islands: one large (**lint**), one small (**sint**) and one extra-small (**xint**). These islands hosted 31 species: 12 species occurring on **A** or **B** and a single species endemic to both **lint** and **sint** represented the basis for turnover, while 18 were widespread (Fig. 1). To test the phylogenetic beta-diversity pattern, we created a phylogenetic hypothesis where species from different areas were randomly related, with the exception of the **sint** and **lint** endemic (s31) which remained closely related to a **B** species (Fig. 1g).

Because of the highly nested pattern, the dissimilarity matrix contained many pairs having zero dissimilarity, in particular all those involving **xint**. Moreover, several dissimilarities were frequently repeated (Fig. 1c). A similar pattern occurred for phylogenetic beta-diversity (Supplementary material Appendix 1). Despite the intermediate characteristics of **xint**, **sint** and **lint**, the UPGMA tree for faunistic

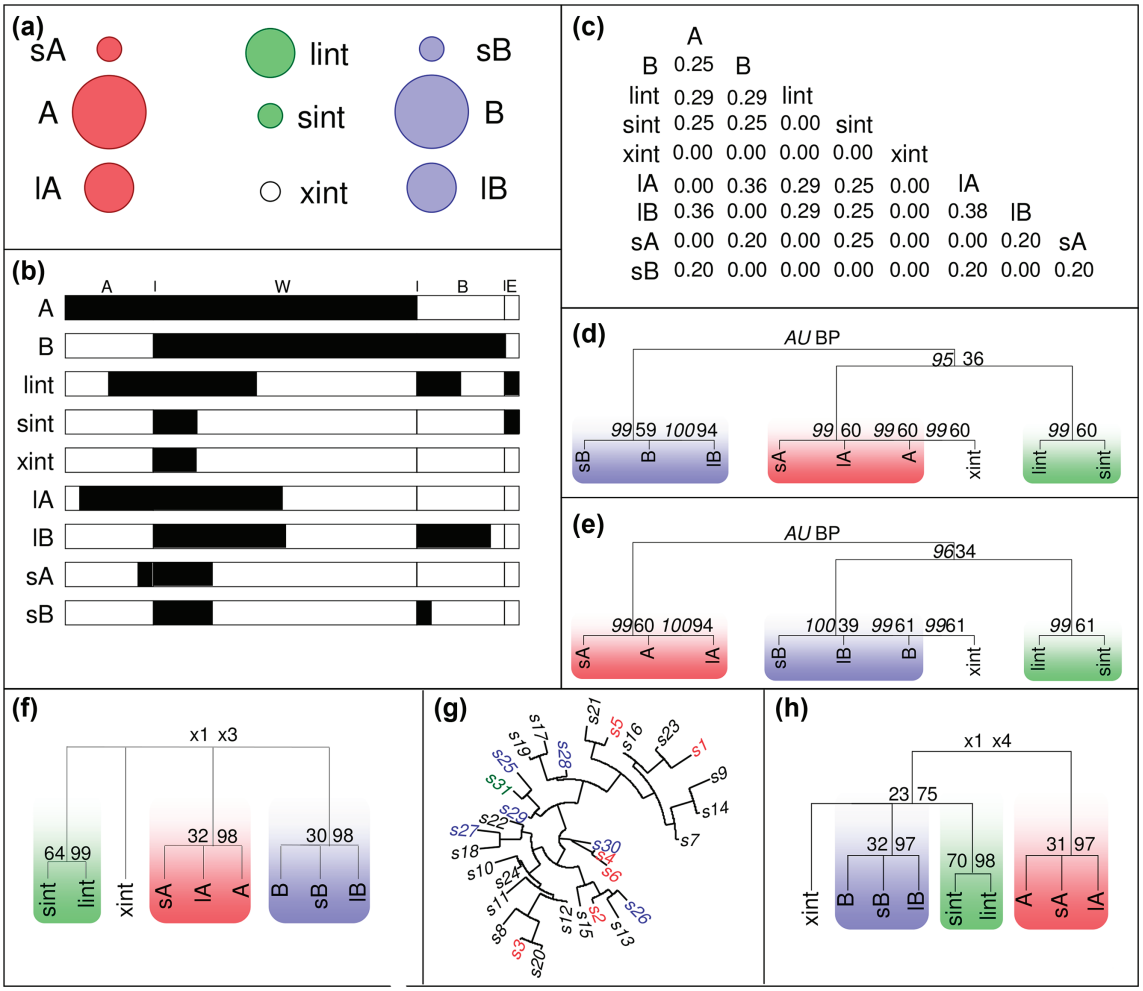

Figure 1. The virtual Archipelago composed of 9 areas with two main ones (A and B) surrounded by small (s) and large (l) islands; **lint** and **sint** and **xint** are a large and a small and an extra small intermediate islands (a). These islands host a hypothetical community of 31 species (rectangles); 18 widespread species (W), 12 species exclusive of A and B and one species (E) endemic to **sint** and **lint** (b). Species exclusive of A and B and the endemic (E) were responsible for turnover resulting in the Simpson dissimilarity matrix shown in (c). Different UPMGA trees are obtained if the order of the first two lines in the matrix is inverted (d) A–B; (e) B–A. AU (left) and BP (right) bootstrap as obtained by 'pvclust' is indicated on nodes. Consensus trees are obtained by resampling (100×) the row order for faunistic beta-diversity (f), and for phylogenetic beta-diversity (h) according to a hypothetical phylogenetic tree (g). Exclusive and endemic species in (g) are coloured according to island group. Bootstrap (BP) values for two different scales of bootstrap: × 1 (left) and × 5 (right) are shown.

turnover placed these islands in the **A** group. In particular, **xint** was very closely linked to **A**, while **sint** and **lint** were linked at a higher dissimilarity (Fig. 1d). There was no theoretical basis to link these islands to the **A** group; nevertheless, bootstrap values obtained by pvclust revealed strong BP and/or AU supports for these unsound nodes. If only the row order of **A** and **B** is reversed in the presence/absence matrix, the linkage of intermediate islands is transferred to **B** with similar supports, despite the dissimilarity matrix being identical (Fig. 1e). Actually, when tied values occur, several solutions are possible at each agglomeration step and many different trees may be generated (Rohlf 1993). Statistical packages typically do not cope with this flaw but use arbitrary linking rules (Legendre and Legendre 1998). Bootstrapping systematically re-samples species but maintains the original order of sites. Consequently, the pairs linked in the first reference tree are more likely to be linked during the entire bootstrap procedure resulting in false strong supports obtained by pvclust.

Because of the tendency of turnover indices to produce tied and zero values, there is great potential for the order of sites to influence tree topology. The number of zero values for a given dataset is the same for faunistic and phylogenetic turnover metrics and, actually, row order affects topology of both indices (Supplementary material Appendix 1). Another artefact occurring in packages relying on the 'hclust' R function like pvclust is the recognition of polytomies as series of bifurcating branches; bootstrap support is also computed for a reduced set of links without any foundation for this selection (Fig. 1d–e).

We have detected this bias in a highly simplified and ad hoc constructed dataset. In real settings, the existence of ties can be less common. We thus tested these problems in a real case study, the thoroughly assessed butterfly fauna of the

west Mediterranean islands (Fig. 2e). We also performed a maximum likelihood reconstruction based on COI barcoding sequences (640 bp) for all west-Mediterranean island species available in public databases (GenBank and BOLD, Supplementary material Appendix 1). We randomly ordered the sites in the matrix several times and the bias recurs (Fig. 2a), since a high fraction of zero values also appears in this dissimilarity matrix (Supplementary material Appendix 1). In the cases shown in Fig. 2a–b, great differences occur in the tree topology and most of the nodes involved in such discrepancies received high AU p-values. BP values were much lower and only in a few cases higher than 50%. Any biogeographical interpretations based on these spurious relationships (e.g. a higher similarity of Sicily to Balearics with respect to circum-Italian islands or vice-versa) might be inconsistent. We used only faunistic beta-diversity for this preliminary assessment because 'pvclust' does not allow analysis of phylogenetic beta-diversity.

To solve this problem, we propose that a series of trees are produced after randomly re-ordering the row order, then consensus trees and bootstrap analysis among the consensus trees are computed. We have written a series of R functions included in the, 'recluster' package (Fig. 3) available at CRAN <http://cran.r-project.org/web/packages/recluster/index.html>. The 'recluster.dist' function computes faunistic and phylogenetic dissimilarity matrices. Faunistic beta-diversity only requires a presence–absence matrix while phylogenetic beta-diversity also needs a phylogenetic tree. Computation of phylogenetic beta-diversity is based on R functions recently provided by Leprieur et al. (2012). The diagnostic function 'recluster.hist' plots the distribution of values in any dissimilarity matrix together with the number of zero values and percentage of tied cells (Supplementary material Appendix 1). The function 'recluster.cons' creates a 'phylo' object representing a consensus tree from a set of trees based on dissimilarity matrices with different order of areas. The consensus rule for accepting nodes can range between 50 and 100%. The first decision for a recluster analysis concerns this threshold. Another diagnostic function 'recluster.node.strength' helps in this decision by evaluating the magnitude of the row-order-bias for different thresholds. This function computes the non-consensus tree with the original order of areas; then, by default, it creates six trees by 'recluster.cons' based on increasing consensus rules from 50 to 100% in 10% steps. Subsequently, it scores the percentage of times each node is recovered for the different consensus runs. Nodes with low percentages are highly affected by row ordering since they are collapsed in most consensus trees.

As a corollary, a large number of tied and zero values can thwart the outcome of obtaining a 100% consensus for any cluster, as in our dataset (Supplementary material Appendix 1). In these cases use of a 100% consensus is not recommended, because most of the pattern would be lost. By using a 50% consensus, we obtained consistent representations for both virtual and real datasets with the former clearly showing the expected pattern in both faunistic and phylogenetic diversity (Fig. 1f–h, 2c–d).

The function 'recluster.boot' allows bootstrapping of nodes in the original consensus tree by applying a user-defined number of consensus trees with user-defined numbers of sampled species. Species sampled twice (or more) have a double effect in determining the faunistic dissimilarity matrix but have no effect on phylogenetic dissimilarity because extra branches are not incorporated in the phylogenetic sub-trees. In a different way to 'hclust', this function considers nodes collapsed by consensus as true polytomies and only provides support for the inner nodes (Fig. 1, 2).

Most nodes received weak support when the number of species randomly sampled with replacement was the same as in the original dataset. That occurred for both databases and for both faunistic and phylogenetic beta-diversity (Fig. 1f–h, 2c–d). In a set of highly nested assemblages, as displayed by most island assemblages, turnover is encompassed by a substantially reduced percentage of species (e.g. just one species for **sA**, **sB**). All bootstrap iterations excluding these species resulted in these islands being identical to **xint** (where only widespread species occur) instead of their **A** or **B** sources, and support for **sA**-**A** and **sB**-**B** links decreased. By applying a multiscale bootstrap, the increase in the number of species randomly selected with repetition can provide greater opportunities for these special taxa to enter the bootstrap matrices, thus increasing the support for these nodes. Accordingly, when numbers of columns greater than the original were used, the bootstrap values linking **sA-lA-A** and **sB-lB-B** increased considerably in both faunistic and phylogenetic beta-diversity (Fig. 1). On the other hand, when a node has a weak (× 1) support because it equally links-up intermediate areas, the increase in the number of species is expected to produce a slower increase in support. We thus designed the 'recluster.multi' function to perform multiscale bootstrap analysis. This function requires the same inputs as 'recluster.boot' and a number of different scales to be applied as a multiplier for the species sampled at each step. The results are stored in a matrix providing bootstrap values for each node (rows) for each bootstrap scale (columns). A first type of node may accrue a rapid increase of support in a multiscale bootstrap. A second kind of node may eventually receive gradual increase in support. It must be noted that by indefinitely multiplying the number of species, all nodes would attain 100% support at some point. Identifying the two kinds of nodes permits recognition as to which links among areas are actually supported by data, even on the basis of a restricted set of species, and which links are uncertain. The 'recluster.identify.nodes' function helps in a proper selection of the parameters to ascertain which nodes belong to each class. This function is useful for inspecting the trend of support at different scales and, by using the Partitioning Around Medoids algorithm, it is possible to find the level where two classes of supports are best distinguished. This function confirmed that both types of nodes are represented in the butterfly dataset, the ×5 scale being that with the best separation between them (Fig. 2e). The tree can be plotted with the function 'recluster.plot' showing bootstrap values from selected scales (×1–×5), indicating the different types of nodes with different colours (Fig. 2c, d). Consensus trees do not have branch lengths; however recluster.cons optionally computes them by a least squares regression. The resulting ultrametric tree can be compared to the initial dissimilarity matrix using different

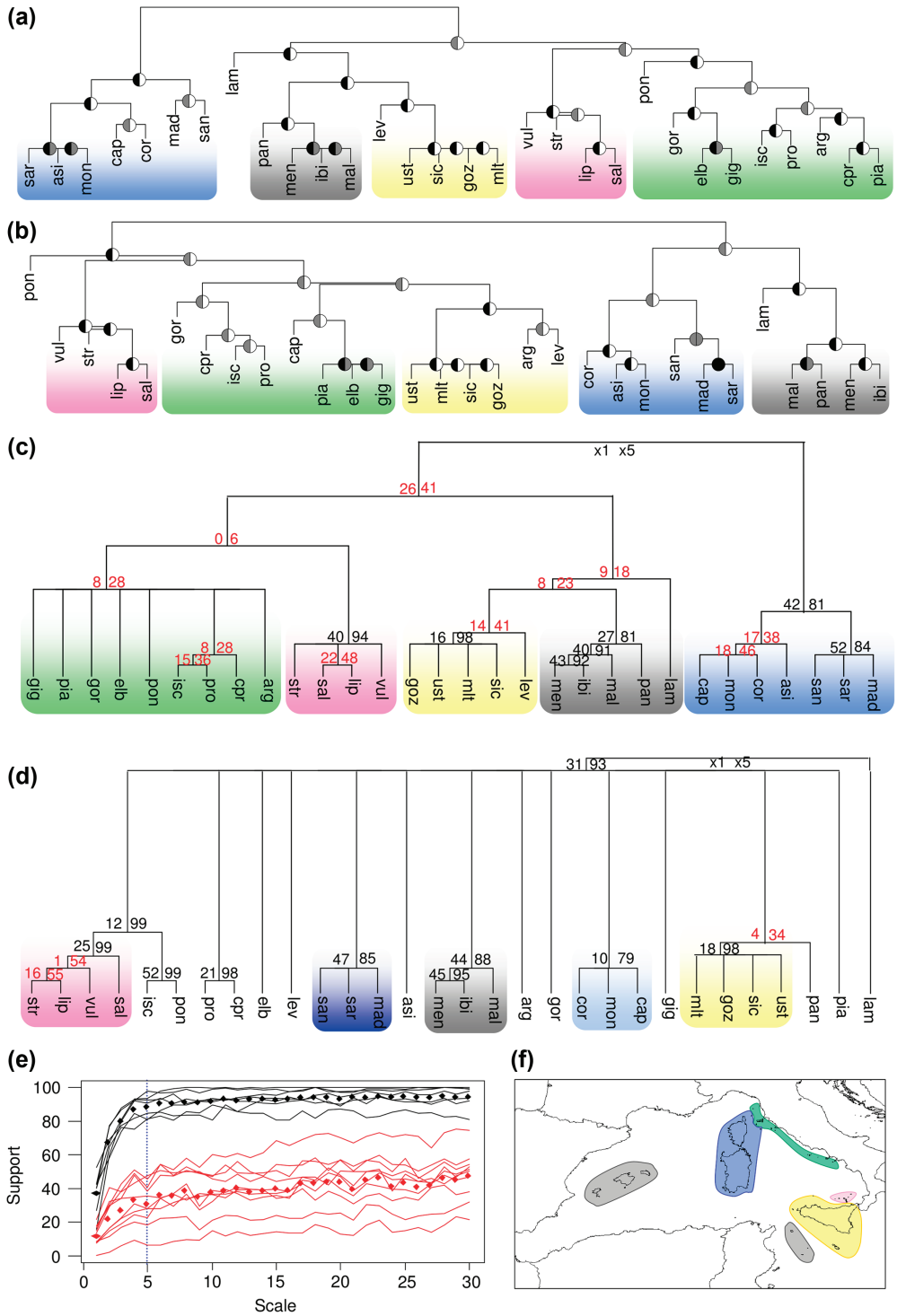

Figure 2. Two examples of UPGMA trees obtained with 'pvclust' by re-sampling the order of butterfly faunas from west Mediterranean islands (a, b). For clarity bootstrap values have been represented by circle sectors representing AU (left) and BP (right) values (white 0–50, grey 50–90, black 90–100). The consensus tree obtained by resampling (100×) the row order for faunistic beta-diversity is represented in (C) together with bootstrap (BP) values for two different scales of bootstrap: ×1 (left) and ×5 (right). The ×5 scale has been chosen according to recluster.identify.nodes function on the basis of a multiscale bootstrap of 30 scales (d), which also serves to identify nodes with considerable increase in support (black) and nodes with no consistent increase (red). The consensus tree obtained by resampling (100×) the row order for phylogenetic beta-diversity is represented in (d) together with bootstrap (BP) values for two different scales of bootstrap: 1× (left) and 5× (right) (Supplementary material Appendix 1). Island abbreviations: san, Sant'Antioco; sar, Sardinia; mad, La Maddelena; asi, Asinara; mon, Montecristo; cap, Capraia; cor, Corsica; lam, Lampedusa; mal, Mallorca; ibi, Ibiza; men, Menorca; sic, Sicily; lev, Levanzo; ust, Ustica; mlt, Malta; goz, Gozo; str, Stromboli; vul, Vulcano; sal, Salina; lip, Lipari; pon, Ponza; isc, Ischia; gig, Giglio; elb, Elba; gor, Gorgona; arg, Argentario; pia, Pianosa; cpr, Capri; pro, Procida. Coloured groups of islands represent consistent clusters among analyses: blue, Sardinia-Corsica; grey, Balearics and south-west Sicilian islands; purple, Aeolian islands; yellow, circum-Sicilian islands; green, circum-Italian Peninsula islands; as shown on the map (e).

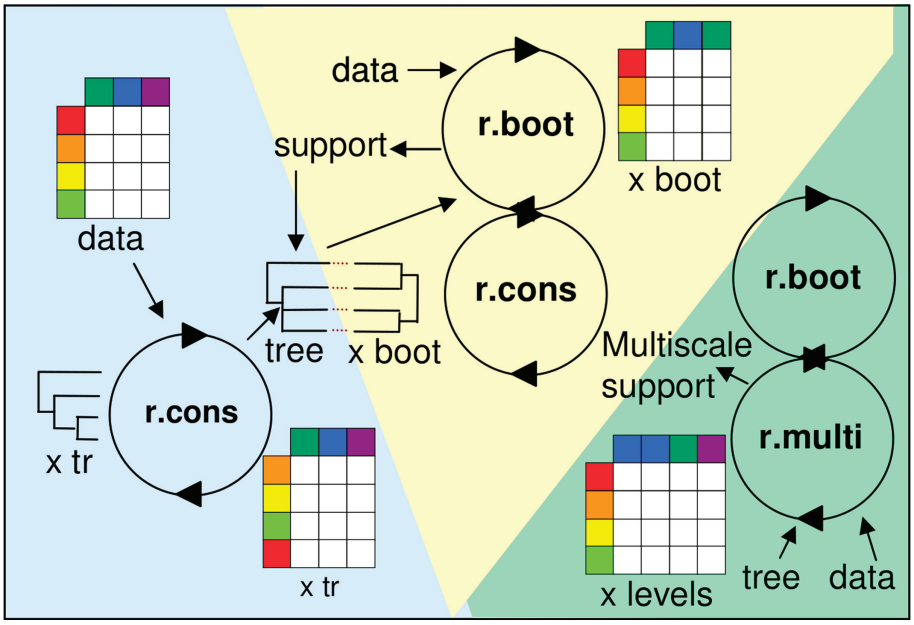

Figure 3. A scheme for the three main functions provided in this study. The matrix with sites in rows and species in columns is resampled tr times for rows by recluster.cons (r.cons) and a first reference consensus tree calculated. Then recluster.boot (r.boot) performs a bootstrap analysis by resampling species boot times and each time recluster.cons creates a consensus tree. This provides support. Finally, recluster.multi (r.multi) repeats recluster.boot using an increasing number of sampled species for each requested level and providing multiscale support for the starting tree.

goodness-of-fit measures (cophenetic distances, 2-norm; Mérigot et al. 2010) available in several R packages.

Conversely to the uncertain relationships in a single UPGMA clustering (Fig. 2a, b) the consensus trees ensure stronger biogeographic assertion. The final representation for faunistic diversity indicates that Sardinia and Corsica represent a distinct cluster with strong bootstrap support. This is in line with the high incidence of endemics. The Aeolian islands, the other Sicilian islands and the Balearics represent other highly supported groups. Strong support for such clusters only occurs when the number of species used in the bootstrap is much greater than the original number (×5). This provides the important indication that these areas are identified on the basis of a few species responsible for turnover, and that most of the overall diversity is encompassed by nestedness. Phylogenetic beta-diversity creates a less structured tree, but support is much stronger, thus confirming that phylogenetic information can reinforce identification of existing patterns (Holt et al. 2013). The main difference with faunistic data is the absence of an Italian group of islands and the separation of Sardinia and Corsica. Large genetic differences among circum-Italian islands and between Sardinia and Corsica are actually found in several taxa (Kudrna et al. 2011).

In conclusion, we demonstrate how applying cluster analyses to turnover dissimilarity can produce highly biased patterns and bootstrap supports. The peculiar strength of the turnover indices is that they produce differences only when species replacement occurs, while the effect of nested species results in zero dissimilarity. Consequently, the use of turnover in recovering biogeographical patterns has become more popular in recent years. Here, we provide a simple solution for extracting consistent clustering of areas and reveal how, by applying the multiscale approach, nodes grouping sites can be identified based only on a few species. Such few species create meaningful links that cannot be ignored but are undetectable at ×1 level. This phenomenon is expected to occur frequently for small and/or isolated communities. Unlike other approaches (e.g. 'pvclust'), we do not provide computations of p-values for node support after multiscale bootstrap; instead, a measure of the frequency of times that each node is replicated, and primarily the behaviour of this value with multiscale bootstrapping, furnishes a direct and better indication of the re-occurrence of each cluster and the reasons determining such a link. The row order problem highlighted here for virtual and real insular examples is also predicted to affect mainland systems despite highly variable outcomes. Indeed, while islands represent areas defined by clear boundaries, mainland units can vary in grain and size according to subjective decisions which in turn influence the structure and determinants of dissimilarity matrices. However, there is great potential for row bias to occur when small and contiguous units are involved and for regions where nestedness dominates over turnover, as in areas recently re-colonized from refugia (Dobrovolski et al. 2012). The diagnostic functions provided in recluster will be useful in evaluating the importance of the bias in each specific dataset and provide simple solutions.

To cite recluster or acknowledge its use, cite this Software note as follows, substituting the version of the application that you used for 'version 0':

Dapporto, L., Ramazzotti, M., Fattorini, S., Talavera, G., Vila, R. and Dennis, R. L. H. 2013. recluster: an unbiased clustering procedure for beta-diversity turnover. – Ecography 36: 1070–1075 (ver. 0).

Acknowledgements – Financial support was provided by the Spanish Ministerio de Ciencia e Innovación project CGL2010-21226/BOS. MR wish to thank EMBO European Molecular Biology Organization for support (ASTF 0075-2010). We thank Fabien Leprieur for his suggestions.

References

Baselga, A. 2010. Partitioning the turnover and nestedness components of beta diversity. – Global Ecol. Biogeogr. 19: 134–143.

Dobrovolski, R. et al. 2012. Climatic history and dispersal ability explain the relative importance of turnover and nestedness components of beta diversity. – Global Ecol. Biogeogr. 21: 191–197.

Holt, B. et al. 2013. An update of Wallace's zoogeographic regions of the world. – Science 339: 74–78.

Supplementary material (Appendix ECOG-00444 at <www.oikosoffice.lu.se/appendix>). Appendix 1.

Kreft, H. and Jetz, W. 2010. A framework for delineating biogeographical regions based on species distributions. – J. Biogeogr. 37: 2029–2053.

Kudrna, O. et al. 2011. Distribution atlas of butterflies in Europe. – Gesellschaft für Schmetterlingsschutz, Germany.

Legendre, P. and Legendre, L. 1998. Numerical ecology. – Elsevier.

Leprieur, F. et al. 2012. Quantifying phylogenetic beta diversity: distinguishing between 'true' turnover of lineages and phylogenetic diversity gradients. – PLoS One 7: e42760.

Mérigot, B. et al. 2010. On goodness-of-fit measure for dendrogram-based analyses. – Ecology 91: 1850–1859.

Rohlf, F. J. 1993. NTSYSpc, numerical taxonomy and multivariate analysis system. – Applied Biostatistics.

Suzuki, R. and Shimodaira, H. 2006. Pvclust: and R package for assessing the uncertainty in hierarchical clustering. – Bioinformatics 22: 1540–1542.