Supporting information for the paper

Biogeography of bird and mammal trophic structures

Multidimensional space exploration with 'AMD analysis'

In 'hard' clustering, data are divided into distinct clusters, whereby each element belongs to one cluster. In fuzzy or 'soft' clustering, elements being clustered show degrees of membership to each cluster (a value between 0 and 1 linked to the squared Euclidean distance to the geometric center of the cluster). Most measurements of cluster validity quantify the degree of cluster separation (measuring how distinct clusters are from each other) and the compactness (measuring how internally consistent they are) (e.g., Davies-Bouldin or Dunn Indexes). Clusters whose elements show a high degree of membership can be then considered well-defined, or internally consistent, whereas clusters with many elements with unclear membership can be considered diffuse or fuzzy.

Suppose that a set of elements are distributed in an *n*-dimensional space (defined by *n* variables) forming *c* relatively well-defined clusters. If the number of clusters *c'* selected are smaller than the number of clusters that actually exist in the data (c' < c), two or more clusters, no matter how different and separated they are, will be forced to be included in the same cluster. The average Euclidean distance of the elements of that cluster to the geometric center will be greater than that of clusters that would include them separately. Therefore, both the degree of membership of its elements to that cluster and its average membership degree (AMD) will be lower. If the number of clusters selected is, on the contrary, higher (c' > c), one or more clusters, no matter how well-defined they are, will be divided into different clusters, thus also reducing their AMD value. When clusters are more separated (distinct) and more compact (well-defined), it is when a number of

clusters equal to the number of groups that actually exist in the data is selected (c' = c). This is also when AMD takes its maximum value.

This relationship between the number of clusters that actually exist in the data and the number of clusters selected allowed us to develop an index based on the AMD of the elements composing the clusters, and asking whether a set of elements distributed in *n*-dimensional space form relatively well-defined groups and, if so, how many groups they are:

$$AMD_i = (1/n \sum_{j=1}^{j=n} \mathrm{MD}_j) - 1/c$$

where MD*j* is the degree of membership of sample *j* to the nearest cluster, *n* the number of samples and *c* the number of clusters.

Step 1. Obtaining the AMDi curve

The curve tracking the AMDi values through a series of clustering analysis runs with an increasing number of user-defined clusters (referred from now as 'AMDi curve') allows to test whether a set of samples, located in a multidimensional space, are randomly distributed or form clusters departing from random chance. The existence of clusters is revealed by the existence of a maximum peak in the AMDi curve, when the number of user-defined clusters matches the number of clusters that actually exist in the set of samples.

Step 2. Obtaining a comparison frame

Once the number of clusters is known, it is possible to obtain an average estimate of their degree of definition, based on the maximum value reached by the peak of the

AMDi curve and its sharpness. To do this, it is first necessary to generate comparison patterns with sets of artificial samples, forming clusters with different degrees of definition (user-controlled by their standard deviation). When the samples are generated by randomly distributing them (they do not form clusters), the resulting AMDi curve is flat. When the samples are generated forming clusters (with varying standard deviations), both the maximum AMDi value and the sharpness of the peak indicate how well defined the clusters were generated. The higher the standard deviation, the greater their dispersion and, therefore, the lower their definition. The larger the definition of the groups (lower standard deviation), the higher and sharper is the resulting peak indicating the number of groups. For the clusters in our sample set to be fully comparable with the artificial sets, the number of samples in both sets must be the same or similar and both the number of dimensions of the space in which they are located and the number of clusters they form must be the same.

This analysis, referred as 'AMD analysis', was applied to the 9-dimensional community-level trophic space, in which, according to our first prediction, samples should not be distributed continuously, rather forming discrete groups (see main text).

Identification of trophic guilds

Although our predictions do not refer to the distribution of species in the space defined by the estimated consumption of resources, we applied the first step of the AMD analysis to the species-level trophic space. The AMDi curve obtained does not show a peak (Fig. 2f), which is an indication they do not form well-defined clusters; hence we did not compare this curve with curves generated with artificial samples (second step of the AMD analysis). However, they are also not randomly distributed, like the artificial

samples of figure 2a, since the AMD_i value increases from 2 to 9 user-defined clusters (Fig. 2f). From 9 user-defined clusters, AMD_i does not decrease but it oscillates, which can be interpreted as indicating the existence of 9 groups made up of smaller groups. We lack an explanation for this pattern. Because our analysis is global and the focus is on generality, we selected the nine major trophic guilds given that they are more likely to be shared across biogeographical realms (Fig. 1).

Identification of community trophic structures

Our first prediction is that communities, even if continuously distributed in a geographical space, will not be continuously distributed in the 'community-level trophic space', rather forming well-defined groups separated by empty or sparsely populated regions. The AMDi curve (first step of the AMD analysis) was obtained for the community-level trophic space, which allowed us to test for the presence of clusters, as well as their number. To estimate their degree of definition, four sets of artificial samples were generated (second step of the AMD analysis), using the same number of samples (18418) and dimensions (9) as the community-level trophic space. With the first set, consisting of randomly distributed samples (they do not form clusters), the resulting AMDi curve is flat (Fig. 2a). The other three sets of artificial clusters were generated forming six clusters with increasing values of standard deviation (5, 10 and 15, in a 100-sided space), implying progressively lower degrees of group definition. As expected, the AMDi curve reached the maximum value when the user-defined number of clusters was equal to the number of groups that actually exist in the data (six, in this case). In addition, the larger the definition of the groups (lower standard deviation), the higher and sharper was the resulting peak indicating the number of groups (Figs. 2b, 2c, 2d). The AMDi curves obtained with these four sets of artificial samples were used as

100 standards of comparison with which to estimate the degree of definition of the clusters
101 identified in the community-level trophic space.
102
103 DOI for data and code is doi:10.5061/dryad.nk98sf7st.