CORRESPONDENCE

Accurate delineation of biogeographical regions depends on the use of an appropriate distance measure

ABSTRACT

The use of analytical techniques to delineate biogeographical regions is becoming increasingly popular. One recent example, Heikinheimo et al. (*Journal of Biogeography*, 2007, **34**, 1053–1064), applied the *k*-means clustering algorithm to define the biogeography of the European land mammal fauna. However, they used the Euclidean distance measure to cluster grid cells described by species-occurrence data, which is inappropriate. The Euclidian distance yields misleading results when applied to species-occurrence data because of the double-zero problem and the species-abundance paradox. We repeat their analysis using the Hellinger distance, a measure appropriate for species-occurrence data and which has been shown to outperform other such measures. Our results differ substantially from those presented by Heikinheimo *et al.* We argue that the rigorous application of appropriate statistical techniques is of crucial concern within conservation biogeography.

Keywords Clustering, conservation biogeography, double-zero problem, Euclidean distance, Europe, Hellinger distance, *k*-means, mammalian fauna, presence/absence data, species-abundance paradox.

The delineation of biogeographical regions is often a necessary first step in conservation planning. Analytical solutions designed to elucidate the spatial structure in complex biological data are increasingly applied to this problem (Procheş, 2005; Moline & Linder, 2006; Mackey *et al.*, 2008; Patten & Smith-Patten, 2008). In particular, Heikinheimo *et al.* (2007) used the *k*-means clustering algorithm to define the biogeography of land mammals in Europe. However, Heikinheimo *et al.* (2007) used the Euclidean distance measure to cluster grid cells described by species-occurrence data, which in our view is inappropriate. We show here that Heikinheimo *et al.*'s (2007) results, namely the spatial delineation of biogeographical regions, are considerably altered when a more appropriate distance measure is used to cluster grid cells.

The *k*-means clustering algorithm is a descriptive multivariate technique and as such does not require that objects be normally distributed (Legendre & Legendre, 1998). However, the solution produced by *k*-means is highly dependent on the use of a distance measure appropriate to the data at hand. By default, *k*-means calculates the within-cluster sums of squares using the Euclidean distance between objects and their centroids. The Euclidean distance is derived from the classic Pythagorean formula:

$$D(x_1, x_2) = \sqrt{\sum_{j=1}^{p} (y_{1j} - y_{2j})^2},$$

where x_1 and x_2 are objects (typically geographical sites) described by $j = \{1 \ldots p\}$ descriptors (e.g. a species list) and y_1 and y_2 are values (e.g. abundance or occurrence) of descriptor j for the objects in question. When the Euclidean distance is used as a measure of dissimilarity among sites described by species occurrences, as is the case in Heikinheimo *et al.* (2007), one encounters what is typically called the 'double-zero problem': two sites that have no species in common are at a distance equal to that of two sites that share species. Using the formula above, it is clear that the Euclidean distance yields a value of 0 for two sites at which a species does not occur ($y_1 = 0$ and $y_2 = 0$) and at which a species does occur ($y_1 = 1$ and $y_2 = 1$). A species is likely to occur at two sites because of the presence of some similarity in environment between the sites (similar climatic conditions, similar habitat, absence of a competitor, etc.). However, the absence of a species at a site can occur for two main reasons: the site is located outside the species' distribution range (true negative), or the species is not detected at a site that is located within its distribution range (false negative). Thus, when comparing two sites at which a species is absent, one should not assume the comparison to be a true negative.

In addition to the double-zero problem, the use of the Euclidean distance with species-occurrence data may also lead to the 'species-abundance paradox' (Legendre & Legendre, 1998). The paradox can be illustrated by a hypothetical example comparing three sites at which the occurrence of four species has been recorded (Table 1). The paradox arises because the Euclidean distance between sites x_1 and x_2 ($D = 1$), which share one species in common, is smaller than the Euclidean distance between sites x_2 and x_3 ($D = 2$), which share two species in common. The species-abundance paradox occurs most frequently when two sites share only a fraction of the species pool in common. Thus, the paradox is expected to be a particular problem at the margins of biogeographical regions where sites may be quite different from one another, rather than in the centre of a region where sites are likely to be very similar in their species assemblages.

Table 1 A hypothetical example to illustrate the species-abundance paradox (see text): occurrence (1 = presence, 0 = absence) of four species (*j*, *k*, *l*, *m*) recorded at three sites (x_1, x_2, x_3).

	Species			
Sites	*j*	*k*	*l*	*m*
x_1	0	0	1	0
x_2	1	0	1	0
x_3	1	1	1	1

www.blackwellpublishing.com/jbi
doi:10.1111/j.1365-2699.2008.01990.x

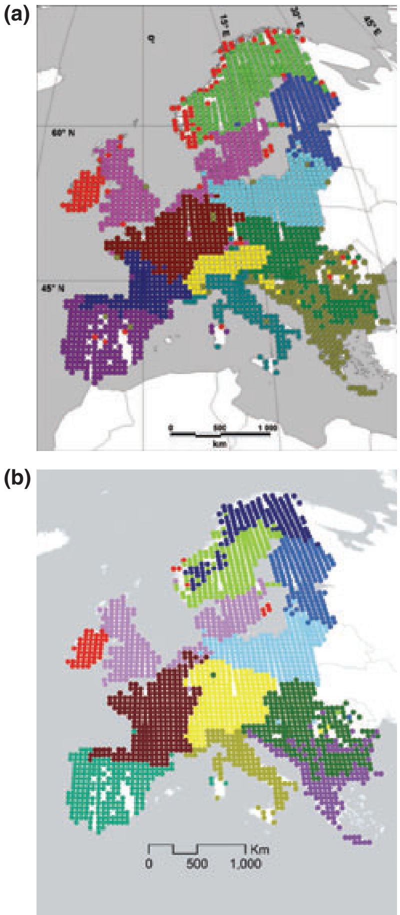

Figure 1 The *k*-means clustering of the 'all-species' set of European land mammal occurrence data in Heikinheimo *et al.* (2007) into 12 clusters using the Euclidean distance (a) and the Hellinger distance (b). Grid cells are plotted using the Mollweide (equal-area) NAD27 projection.

We repeated Heikinheimo *et al.*'s (2007) analysis with a distance measure more appropriate for species-occurrence data. Although several such measures are already in use by biogeographers [e.g. the Kulczynski coefficient (Moline & Linder, 2006) and the Bray–Curtis coefficient (Procheş, 2005)], we chose the Hellinger distance measure (Rao, 1995). When compared with other distance measures appropriate for species-abundance data (chord distance, chi-squared distance, Bray–Curtis distance), the Hellinger distance has been shown to be the most representative of the true geographical distance among sites (Legendre & Gallagher, 2001). As in Heikinheimo *et al.* (2007), we clustered 2183 grid cells characterized by the presence/absence records of 124 mammal species collected by the Societas Europaea Mammalogica (http://www.european-mammals.org) to prepare the *Atlas of European mammals* (Mitchell-Jones *et al.*, 1999). All analyses were performed with R ver. 2.6.0 (R Development Core Team, 2007) and ArcGIS ver. 9.1 (ESRI, 2005).

Our cluster analysis produced biogeographical regions considerably different from those presented by Heikinheimo *et al.* (2007) (Fig. 1). Differences were most apparent in central Europe, characterized by three regions in Heikinheimo *et al.* (2007) and two in this paper, and Scandinavia, where the delineation of regions was much altered. Comparisons between our results and those of Heikinheimo *et al.* (2007) for 10 additional species sets are provided in Appendix S1 in Supporting Information.

The differences in the delineation of European land mammal biogeographical regions between Heikinheimo *et al.*'s (2007) study and this paper highlight the importance of using an appropriate distance measure in multivariate analyses of complex biological data. The issue is compounded in this case because Heikinheimo *et al.*'s (2007) results are so readily applicable to conservation planning. Meaningful conservation plans require information on the spatial distribution of organisms, ideally a true representation of the spatial structure inherent in empirical species composition data. In future, we hope that a greater emphasis on the application of rigorous multivariate techniques within conservation biogeography will lead to an improved understanding of the spatial distribution of organisms, and ultimately their conservation.

Sara A. Gagné* and Raphaël Proulx

Geomatics and Landscape Ecology Research Laboratory, Department of Biology, Carleton University, 1125 Colonel By Drive, Ottawa, ON, Canada K1S 5B6

**E-mail: saraanne.gagne@gmail.com*

REFERENCES

ESRI (2005) *ArcGIS, version 9.1.* Environmental Systems Research Institute, Redlands, CA, USA.

Heikinheimo, H., Fortelius, M., Eronen, J. & Mannila, H. (2007) Biogeography of European land mammals shows environmentally distinct and spatially coherent clusters. *Journal of Biogeography*, **34**, 1053–1064.

Legendre, P. & Gallagher, E.D. (2001) Ecologically meaningful transformations for ordination of species data. *Oecologia*, **129**, 271–280.

Legendre, P. & Legendre, L. (1998) *Numerical ecology*, 2nd edn. Elsevier, Amsterdam.

Mackey, B.G., Berry, S.L. & Brown, T. (2008) Reconciling approaches to biogeographic regionalization: a systematic and generic framework examined with a case study of the Australian continent. *Journal of Biogeography*, **35**, 213–229.

Mitchell-Jones, A.J., Amori, G., Bogdanowicz, W., Krystufek, B., Reijnders, P.J.H., Spitzenberger, F., Stubbe, M., Thissen, J.B.M., Vohralik, V. & Zima, J. (1999) *The atlas of European mammals.* Academic Press, London.

Moline, P.M. & Linder, H.P. (2006) Input data, analytical methods and biogeography of *Elegia* (Restionaceae). *Journal of Biogeography*, **33**, 47–62.

Patten, M.A. & Smith-Patten, B.D. (2008) Biogeographical boundaries and Monmonier's algorithm: a case study in the northern Neotropics. *Journal of Biogeography*, **35**, 407–416.

Procheş, S. (2005) The world's biogeographical regions: cluster analyses based on bat distributions. *Journal of Biogeography*, **32**, 607–614.

R Development Core Team (2007) *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna.

Rao, C.R. (1995) A review of canonical coordinates and an alternative to correspondence analysis using Hellinger distance. *Quaderns d'Estadística i Investigació Operativa*, **19**, 23–63.

SUPPORTING INFORMATION

Additional Supporting Information may be found in the online version this article:

Appendix S1 Biogeographical maps created using the Euclidean distance and the Hellinger distance for the 10 species sets defined by Heikinheimo *et al.* (2007).

Please note: Wiley-Blackwell is not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.

Editor: Bradford Hawkins

doi:10.1111/j.1365-2699.2008.01990.x