Design of Dissimilarity Measures: a New Dissimilarity between Species Distribution Areas

Christian Hennig[1] and Bernhard Hausdorf[2]

[1] Department of Statistical Science, University College London
Gower St, London WC1E 6BT, United Kingdom
[2] Zoologisches Museum der Universität Hamburg
Martin-Luther-King-Platz 3, 20146 Hamburg, Germany

Abstract. We give some guidelines for the choice and design of dissimilarity measures and illustrate some of them by the construction of a new dissimilarity measure between species distribution areas in biogeography. Species distribution data can be digitized as presences and absences in certain geographic units. As opposed to all measures already present in the literature, the geco coefficient introduced in the present paper takes the geographic distance between the units into account. The advantages of the new measure are illustrated by a study of the sensitivity against incomplete sampling and changes in the definition of the geographic units in two real data sets.

1 Introduction

We give some guidelines for the choice and design of dissimilarity measures (in Section 2) and illustrate some of them by the construction of a new dissimilarity measure between species distribution areas in biogeography.

Species distribution data can be digitized as presences and absences in certain geographic units, e.g., squares defined by a grid over a map. In the so-called R-mode analysis in biogeography, the species are the objects between which the dissimilarity is to be analyzed, and they are characterized by the sets of units in which they are present.

More than 40 similarity and dissimilarity measures between distribution areas have already been proposed in the literature (see Shi, 1993, for 39 of them). The choice among these is discussed in Section 3.

Somewhat surprisingly, none of these measures take the geographic distance between the units into account, which can provide useful information, especially in the case of incomplete sampling of species presences. Section 4 is devoted to the construction of a new dissimilarity coefficient which incorporates a distance matrix between units. In most applications this will be the geographic distance.

Some experiments have been performed on two data sets of species distribution areas, which explore the high stability of the new measure under incomplete sampling and change of the grid defining the units. They are explained in detail in Hennig and Hausdorf (2005). An overview of their results is given in Section 5.

2 Some thoughts on the design of dissimilarity measures

In many situations, dissimilarities between objects cannot be measured directly, but have to be constructed from some known characteristics of the objects of interest, e.g. some values on certain variables.

From a philosophical point of view, the assumption of the objective existence of a "true" but not directly observable dissimilarity value between two objects is highly questionable. Therefore we treat the dissimilarity construction problem as a problem of the choice or design of such a measure and not as an estimation problem of some existing but unknown quantities.

Therefore, subjective judgment is necessarily involved, and the main aim of the design of a dissimilarity measure is the proper representation of a subjective or intersubjective concept (usually of subject-matter experts) of similarity or dissimilarity between the objects. Such a subjective concept may change during the process of the construction - the decisions involved in such a design could help the experts to re-think their conceptions. Often the initial expert's conceptions cannot even be assumed to be adequately representable by formulae and numbers, but the then somewhat creative act of defining such a representation may still have merits. It enables the application of automatic data analysis methods and can support the scientific discussion by making the scientist's ideas more explicit ("objectivating" them in a way).

Note that Gordon (1990) discussed the problem of finding variable weights in a situation where the researchers are able to provide a dissimilarity matrix between the objects but not a function to compute these values from the variables characterizing the objects, in which case the design problem can be formalized as a mathematical optimization problem. Here we assume that the researchers cannot (or do not want) to specify all dissimilarity values directly, but rather are interested in formalizing their general assessment principle, which we think supports the scientific discourse better than to start from subjectively assigned numbers.

The most obvious subjective component is the dependence of a dissimilarity measure on the research aim. For example, different similarity values may be assigned to a pair of poems depending on whether the aim is to find poems from the same author in a set of poems with unknown author or to assess poems so that somebody who likes a particular poem will presumably also like poems classified as similar. This example also illustrates a less subjective aspect of dissimilarity design: the quality of the measure with respect to the research aim can often be assessed by observations (such as the analysis of dissimilarities between poems which are known to be written by the same author). Such analyses, as well as the connection of the measure to scientific knowledge and common sense considerations, improve the scientific acceptability of a measure.

A starting point for dissimilarity design is the question: "how can the researcher's (or the research group's) idea of the similarity between objects given a certain research aim be translated into a formally defined function of the observed object characteristics?"

This requires at first a basic identification of how the observed characteristics are related to the researcher's concept. For species distribution areas, we start with the idea that similarity of two distribution areas is the result of the origin of the species in the same "area of endemism", and therefore distribution areas should be treated as similar if this seems to be plausible. Eventually, the dissimilarity analysis (using techniques like ordination and cluster analysis) could provide us with information concerning the historic process of the spciation (Hausdorf and Hennig, 2003).

It is clear that the dissimilarity measure should become smaller if (given constant sizes of the areas) the number of units in which both species are present becomes larger. Further, two very small but disjunct distribution areas should not be judged as similar just because the number of units in which both species are not present is large, while we would judge species present at almost all units as similar even if their few non-occurrences don't overlap. This suggests that the number of common absences is much less important (if it has any importance at all) for dissimilarity judgments than the number of common presences. The species distribution area problem is discussed further in the next section.

Here are some further guidelines for the design and "fine-tuning" of dissimilarity measures.

- After having specified the basic behaviour of the dissimilarity with respect to certain data characteristics, think about the importance weights of these characteristics. (Note that variable weights can only be interpreted as importance weights if the variables are suitably standardized.)
- Construct exemplary (especially extreme) pairs of objects in which it is clear what value the dissimilarity should have, or at least how it should compare with some other exemplary pairs.
- Construct sequences of pairs of objects in which one characteristic changes while others are held constant, so that it is clear how the dissimilarity should change.
- Think about whether and how the dissimilarity measure could be disturbed by small changes in the characteristics, what behaviour in these situations would be adequate and how a measure could be designed to show this behaviour.
- Think about suitable invariance properties. Which transformations of the characteristics should leave the dissimilarities unchanged (or only changed in a way that doesn't effect subsequent analyses, e.g. multiplied

by a constant)? There may be transformations under which the dissimilarities can only be expected to be approximately unchanged, e.g. the change of the grid defining the geographic units for species areas.
- Are there reasons that the dissimilarity measure should be a metric (or have some other particular mathematical properties)?
- The influence of monotone characteristics on the dissimilarity should not necessarily be linear, but can be convex or concave (see the discussion of the function u below).
- If the measure should be applied to a range of different situations, it may be good to introduce tuning constants, which should have a clear interpretation in terms of the subject matter.

3 Jaccard or Kulczynski coefficient?

We denote species areas as sets A of geographic units, which are subsets of the total region under study $R = \{r_1, \dots, r_k\}$ with k geographic units. $|A|$ denotes the number of elements in A (size of A).

The presumably most widely used dissimilarity measure in biogeography is the Jaccard coefficient (Jaccard, 1901)

$$d_J(A_1, A_2) = 1 - \frac{|A_1 \cap A_2|}{|A_1 \cup A_2|}.$$

This distance has a clear direct interpretation as the proportion of units present in A_1 or A_2, but not in both of them. It does not depend on the number of common absences, which is in accord with the above discussion.

However, there is an important problem with the Jaccard distance. If a smaller area is a subset of a much larger area, the Jaccard distance tends to be quite large, but this is often inadequate. For example, if there are $k = 306$ units (as in an example given below), $A \subset B,\ |A| = 4, |B| = 20$, we have $d_J(A, B) = 0.8$, though both species may have originated in the same area of endemism. A may only have a worse ability for dispersal than B. We would judge A as more similar (in terms of our research aims) to B than for example a species C with $|C| = 20,\ |B \cap C| = 10$, but $d_J(B, C) = 0.67$. The reason is that the Jaccard denominator $|A_1 \cup A_2|$ is dominated by the more dispersed species which therefore has a higher influence on the computation of d_J.

Giving both species the same influence improves the situation, because $|A \cap B|$ is small related to $|B|$, but large related to $|A|$. This takes into account differences in the sizes of the species areas to some extent (which is desirable because very small species areas should not be judged as very similar to species occupying almost all units), but it is not dominated by them as strongly as the Jaccard distance. This leads to the Kulczynski coefficient (Kulczynski, 1927)

$$d_K(A_1, A_2) = 1 - \frac{1}{2}\left(\frac{|A_1 \cap A_2|}{|A_1|} + \frac{|A_1 \cap A_2|}{|A_2|}\right),$$

for which $d_K(A,B) = 0.4$ and $d_K(B,C) = 0.5$ while the good properties of the Jaccard coefficient mentioned above are preserved. However, the Jaccard coefficient is a metric (Gower and Legendre, 1986) while the triangle inequality is not fulfilled for the Kulczynski coefficient. This can be seen as follows. Consider $D \subset B,\ |D| = 4,\ |D \cap A| = 0$. Then $d_K(D,B) + d_K(B,A) = 0.8 < d_K(A,D) = 1$. But this makes some sense. Using only set relations and ignoring further geographical information, the dissimilarity between A and D should be the maximal value of 1 because they are disjunct. On the other hand, for the reasons given above, it is adequate to assign a small dissimilarity to both pairs A, B and B, D, which illustrates that our subject matter concept of dissimilarity is essentially non-metric. Therefore, as long as we do not require the triangle inequality for any of the subsequent analyses, it is more adequate to formalize our idea of dissimilarity by a non-metric measure. Actually, if we apply a multidimensional scaling algorithm to embed the resulting dissimilarity matrix in the Euclidean space, such an algorithm will essentially reduce the distance between A and D in the situation above, which is satisfactory as well, because now the fact that the common superset B exists can be taken into account to find out that A and D may have more in common than it seems from just looking at $A \cap D$. For example, they may be competitors and therefore not share the same units, but occur in the same larger area of endemism.

Note that the argument given above is based on the fact that $|B| = 20$ is much smaller than the whole number of units. This suggests that a more sophisticated approach may further downweight the relation of $|A_1 \cap A_2|$ to the size of the larger area, dependent on the number of common absences (an extreme and for our aims certainly exaggerated suggestion is the consideration of $1 - \frac{|A_1 \cap A_2|}{|A_1|}$ where A_1 is the smaller area, see Simpson, 1960).

4 Incorporating geographic distances

Assume now that there is a distance d_R defined on R, which usually will be the geographic distance. Obviously this distance adds some useful information. For example, though A and D above are disjunct, the units of their occurrence could be neighboring, which should be judged as a certain amount of similarity in the sense of our conception.

Furthermore, small intersections (and therefore large values of both d_J and d_K) between seemingly similar species areas may result from incomplete sampling or very fine grids.

The motivation for the definition of our new geco coefficient (the name comes from "geographic distance and congruence") was that we wanted to maintain the equal weighting of the species of the Kulczynski coefficient while incorporating the information given by d_R.

The general definition is

$$d_G(A_1, A_2) = \frac{1}{2}\left(\frac{\sum_{a \in A_1} \min_{b \in A_2} u(d_R(a,b))}{|A_1|} + \frac{\sum_{b \in A_2} \min_{a \in A_1} u(d_R(a,b))}{|A_2|}\right),$$

where u is a monotone increasing transformation with $u(0) = 0$. To motivate the geco coefficient, consider for a moment u as the identity function. Then, d_G is the mean of the average geographic distance of all units of A_1 to the respective closest unit in A_2 and the average geographic distance of all units of A_2 to the respective closest unit in A_1. Thus, obviously, $d_G(A_1, A_1) = 0$, $d_G(A_1, A_2) \geq 0$, $d_G(A_1, A_2) = d_G(A_2, A_1)$ and $d_G(A_1, A_2) \leq \max u(d_R)$. If $u(d_R(a, b)) > 0$ for $a \neq b$, then $d_G(A_1, A_2) > 0$ for $A_1 \neq A_2$. d_G reduces to the Kulczynski coefficient by taking $d_R = \delta$ with $\delta(a, a) = 0$, $\delta(a, b) = 1$ if $a \neq b$ and u as the identity function, because

$$|A_1 \cap A_2] = \sum_{a \in A_1} \min_{b \in A_2} (1 - \delta(a, b)) = \sum_{b \in A_2} \min_{a \in A_1} (1 - \delta(a, b)).$$

It follows that d_G is not generally a metric, though it may become a metric under certain choices of u and d_R (δ is a metric, which shows that demanding d_R to be a metric does not suffice). Given that A and D from the example of the previous Section are far away from each other and B is present at both places, the violation of the triangle inequality may still be justified.

Note that for general d_R, $\sum_{a \in A} \min_{b \in B}(1-d_R(a,b)) = \sum_{b \in B} \min_{a \in A}(1-d_R(a,b))$ does not hold, and therefore it is favorable for the aim of generalization that $|A \cap B|$ appears in the definition of the Kulczynski coefficient related to $|A|$ *and* $|B|$. A corresponding generalization of the Jaccard distance would be less intuitive.

The identity function may be reasonable as a choice for u in particular situations, but often it is not adequate. Consider as an example d_R as geographic distance, and consider distribution areas A, B, C and D all occupying only a single geographic unit, where the unit of A is 10 km distant from B, 5000 km distant from C and 10000 km distant from D. Then, if u is the identity function, the geco distances from A to B, C and D are 10, 5000 and 10000, thus distribution area D is judged as twice as different from A than C. But while in many circumstances a small geographic distance is meaningful in terms of the similarity of distribution areas (because species may easily get from one unit to another close unit and there may be similar ecological conditions in close units, so that species B is in fact similar to A), the differences between large distances are not important for the similarity between species areas and units which are 5000 and 10000 km away from A may both simply not be in any way related to the unit of A. Thus, we suggest for geographical distances a transformation u that weights down the differences between large distances. A simple choice of such a transformation is the following:

$$u(d) = u_f(d) = \begin{cases} \frac{d}{f * \max d_R} & : d \leq f * \max d_R \\ 1 & : d > f * \max d_R \end{cases}, 0 \leq f \leq 1.$$

That is, u_f is linear for distances smaller than f times the diameter (maximum geographical distance) of the considered region R, while larger geographical distances are treated as "very far away", encoded by $u_f = 1$. This yields $\max d_G = \max u(d_R) = 1$, makes the geco coefficient independent of the scaling of the geographical distances (kilometers, miles etc.) and directly comparable to the Kulczynski distance. In fact, $f = 0$ (or f chosen so that $f * \max d_R$ is smaller than the minimum nonzero distance in R) yields the Kulczynski distance, and $f = 1$ is equivalent to u chosen as the identity function scaled to a maximum of 1. f should generally be chosen so that $f * \max d_R$ can be interpreted as the minimal distance above which differences are no longer meaningful with respect to the judgment of similarity of species. We suggest $f = 0.1$ as a default choice, assuming that the total region under study is chosen so that clustering of species may occur in much smaller subregions, and that relevant information about a particular unit (e.g., about possible incomplete sampling) can be drawn from a unit which is in a somewhat close neighborhood compared to the whole area of the region. $f = 0.1$ has been used in both experiments below. A larger f may be adequate if the region under study is small, a smaller f may be used for a very fine grid.

There are alternatives to the choice of u that have a similar effect, e.g., $u(d) = \log(f * d + 1)$. However, with this transformation, f would be more difficult to choose and to interpret.

The geco coefficient may be used together with more sophisticated measures d_R quantifying for example dissimilarities with respect to ecological conditions between units or "effective distances" taking into account geographical barriers such as mountains.

5 Experiments with the geco coefficient

We carried out two experiments to explore the properties of the geco coefficient and to compare it with the Kulczynski coefficient. Full descriptions and results can be found in Hennig and Hausdorf (2005).

The first experiment considers the sensitivity against incomplete sampling. The data set for this experiment includes the distribution of 366 land snail species on 306 grid squares in north-west Europe. The data set has been compiled from the distribution maps of Kerney et al. (1983). These maps are interpolated, i.e., presences of a species have been indicated also for grid squares in which it might have not been recorded so far, but where it is probably present, because it is known from the surrounding units. Therefore this data set is artificially "complete" and especially suitable to test the effect of incomplete sampling on biogeographical analyses.

To simulate incomplete sampling, every presence of a species in a geographic unit given in the original data set has been deleted with a probability P (which we chose as 0.1, 0.2 and 0.3 in different simulations; 100 replications have been performed for all setups) under the side condition that every species is still present in the resulting simulated data. To compare the Kulczynski distance and the geco coefficient, we computed the Pearson correlation between the vector of dissimilarities between species in the original data set and the vector of dissimilarities between species in the simulated data set. We also carried out a non-metric MDS and a cluster analysis based on normal mixtures (see Hennig and Hausdorf, 2005, for the whole methodology) and compared the solutions from the contaminated data sets with the original solutions by means of a Procrustes-based coefficient (Peres-Neto and Jackson, 2001) and the adjusted Rand index (Hubert and Arabie, 1985).

In terms of Pearson correlations to the original data set, the geco coefficient yielded mean values larger than 0.975 for all values of P and outperformed the Kulczynski coefficient on all 300 simulated data sets. The results with respect to the MDS and the clustering pointed into the same direction. The tightest advantage for the geco coefficient was that its clusterings obtained a better Rand index than Kulczynski "only" in 78 out of 100 simulations for $P = 0.1$.

The second experiment explores the sensitivity against a change of the grid. The data set for this experiment includes the distribution of 47 weevil species in southern Africa. We used a presence/absence matrix for 2 degree latitude x 2 degree longitude grid cells as well as a presence/absence matrix for 1 degree latitude x 1 degree longitude grid cells, both given by Mast and Nyffeler (2003).

Hausdorf and Hennig (2003) analyzed the biotic element (species area cluster) composition of the weevil genus *Scobius* in southern Africa using Kulczynski distances. The results obtained with a 1 degree grid differed considerably from those obtained with a 2 degree grid. On the coarser 2 degree grid, a more clear clustering and more seemingly meaningful biotic elements have been found, though the finer grid in principle provides more precise information. Hausdorf and Hennig (2003) suggested that "If the grid used is too fine and the distribution data are not interpolated, insufficient sampling may introduce artificial noise in the data set".

If the 1 degree grid is analysed with the geco coefficient, the structures found on the 2 degree grid by geco and Kulczynski coefficients can be reproduced and even a further biotic element is found. The geco analyses on both grids are much more similar to each other (in terms of Pearson correlation, Procrustes and adjusted Rand index) than the two Kulczynski analyses.

6 Conclusion

We discussed and introduced dissimilarity measures between species distribution areas. We used some techniques that are generally applicable to the design of dissimilarity measures, namely the construction of archetypical extreme examples, the analysis of the behaviour under realistic transformations or perturbations of the data and the introduction of nonlinear monotone functions and clearly interpretable tuning constants to reflect the effective influence of some characteristics of the data.

References

GORDON, A. D. (1990): Constructing Dissimilarity Measures. *Journal of Classification, 7/2, 257-270.*

GOWER, J. C. and LEGENDRE, P. (1986): Metric and Euclidean Properties of Dissimilarity Coefficients. *Journal of Classification, 3/1, 5-48.*

HAUSDORF, B., and HENNIG, C. (2003): Biotic Element Analysis in Biogeography. *Systematic Biology, 52, 717-723.*

HENNIG, C. and HAUSDORF, B. (2005): A Robust Distance Coefficient between Distribution Areas Incorporating Geographic Distances. To appear in *Systematic Biology.*

HUBERT, L. and ARABIE, P. (1985): Comparing Partitions. *Journal of Classification, 2/2, 193-218.*

JACCARD, P. (1901): Distribution de la florine alpine dans la Bassin de Dranses et dans quelques regiones voisines. *Bulletin de la Societe Vaudoise des Sciences Naturelles, 37, 241-272.*

KERNEY, M. P., CAMERON, R. A. D., and JUNGBLUTH, J. H. (1983): *Die Landschnecken Nord- und Mitteleuropas.* Parey, Hamburg and Berlin.

KULCZYNSKI, S. (1927): Die Pflanzenassoziationen der Pieninen. *Bulletin International de l'Academie Polonaise des Sciences et des Lettres, Classe des Sciences Mathematiques et Naturelles, B, 57-203.*

MAST, A. R. and NYFELLER, R. (2003): Using a null model to recognize significant co-occurrence prior to identifying candidate areas of endemism. *Systematic Biology, 52, 271-280.*

PERES-NETO, P. R. and JACKSON, D. A. (2001): How well do multivariate data sets match? The advantages of a Procrustean superimposition approach over the Mantel test. *Oecologia, 129, 169-178.*

SHI, G. R. (1993): Multivariate data analysis in palaeoecology and palaeobiogeography-a review. *Palaeogeography, Palaeoclimatology, Palaeoecology, 105, 199-234.*

SIMPSON, G. G. (1960): Notes on the measurement of faunal resemblance. *American Journal of Science, 258-A, 300-311.*