Distance-based parametric bootstrap tests for clustering of species ranges

by

Christian Hennig [1] and Bernhard Hausdorf [2]

Research Report No. 110
November 2002

Seminar für Statistik

Eidgenössische Technische Hochschule (ETH)

CH-8092 Zürich

Switzerland

[1] ETH Zürich (LEO), Seminar für Statistik, CH-8092 Zürich, Switzerland, and Universität Hamburg, Fachbereich Mathematik-SPST, 20146 Hamburg, Germany, hennig@math.uni-hamburg.de

[2] Zoologisches Museum der Universität Hamburg, Martin-Luther-King-Platz 3, 20146 Hamburg, Germany, hausdorf@zoologie.uni-hamburg.de

Distance-based parametric bootstrap tests for clustering of species ranges

Christian Hennig [‡] and Bernhard Hausdorf [§]

Seminar für Statistik
ETH Zentrum
CH-8092 Zürich, Switzerland

November 2002

Abstract

This paper deals with species range data, i.e., n species (taxa) are characterized by their presence or absence on c units into which a map is subdivided. Such data occur often in biogeography. We propose some tests for the existence of clusters of species according to their ranges. We define some distance-based test statistics for the presence of clustering, we propose a null model for the generation of a species and an alternative model for clustering. The models include a parameter governing the spatial autocorrelation of its occurrence in the cells and they account for the species richness of the individual cells. The distribution of the test statistics can be estimated by a parametric bootstrap simulation (Monte Carlo with estimated parameters) from the null model. The validity of the p-values and the power of the tests are considered by exemplary simulations. We discuss also, but do not focus on, the determination of the clusters.

Keywords: spatial autocorrelation, presence-absence data, biogeography, clustering under noise, Monte Carlo, double bootstrap

1 Introduction

In this paper we consider data of the following form: let $R = \{1, \ldots, c\}$ be a set of geographic units. An example are the 306 cells into which north-west Europe is subdivided in Figure 1. The data consists of n species ranges, where the range of a species (taxon) A is a subset of R, namely the set of cells where the taxon is known to be present. The taxon could equivalently be represented by a 0-1 vector of length c, which is often referred to as "presence/absence data". In biogeography and especially in research concerning the evolution and differentiation of species, it is of interest whether the taxa form "biotic elements" (Hausdorf, 2002), i.e., clusters of taxa sharing very similar ranges. We focus on the derivation of tests of a null hypothesis of homogeneity against the presence of clustering.

[‡]ETH Zürich (LEO), Seminar für Statistik, CH-8092 Zürich, Switzerland, and Universität Hamburg, Fachbereich Mathematik-SPST, 20146 Hamburg, Germany, hennig@math.uni-hamburg.de

[§]Zoologisches Museum der Universität Hamburg, Martin-Luther-King-Platz 3, 20146 Hamburg, Germany, hausdorf@zoologie.uni-hamburg.de

There are two important characteristics of taxon range data. Firstly, the occurrences of a single species are spatially autocorrelated: the occurrence of a taxon in a unit enlarges the probability of its occurrence in neighboring units. Secondly, in most situations, the regions differ with respect to their potential for taxa richness. These features will often lead to the rejection of homogeneity modeled by standard null models of cluster analysis (Bock (1994, 1996) and also Section 3.2), which therefore should not be applied in the present setup. Instead, we are interested in the detection of a clustering that cannot be explained by spatial autocorrelation of the single taxa and the different potentials of the cells for taxa richness alone. Null models for taxon ranges discussed previously in the ecological literature ignore the modeling either of the spatial autocorrelation (Cook and Quinn, 1998) or of the different taxa richness of the units (Roxburgh and Matsuki, 1999).

We start by the definition of a null model for the range of a population of taxa in Section 2.1, taking into account the distribution of the sizes of the taxon ranges, the taxa richness of the units, and the spatial autocorrelation. The null model may be of interest not only for the test of clustering, but also in other research areas where presence/absence data are involved, such as the investigation of nestedness between taxon ranges (cf. Cook and Quinn (1998); Hausdorf and Hennig (2003); Patterson and Atmar (1986); Wright and Reeves (1992); Wright et al. (1998) and also Section 3.1) or the analysis of association between pairs of taxa (Palmer and van der Maarel, 1995). In Section 2.2, the estimation of the involved parameters for autocorrelation and the units potentials for taxa richness is discussed.

In Section 2.4, we propose some test statistics, which formalize the presence of clustering. All test statistics are based on distances. The choice of an adequate distance measure between taxon ranges is discussed in Section 2.3. While our tests are attempted to distinguish the null model from various alternatives leading to more clustered taxon ranges, we propose a particular alternative model for the presence of taxa clusters in Section 2.5, which is related to the vicariance biogeography (Nelson and Platnick, 1981; Wiley, 1988; Humphries and Parenti, 1999).

We present the application of the methodology to two datasets. One of them consists of 366 ranges of land snail species in 306 cells of north-west Europe as shown in Figure 1, the other consists of 55 land snails species ranges in 251 cells in Israel and Palestine in Section 3.1.

The p-value of the test statistics have to be computed by a Monte Carlo simulation. Since the autocorrelation and the taxa richness parameters are to be estimated from the data, the simulation is a so-called "parametric bootstrap" (Davison and Hinkley, 1997), and the validity of the p-value is not guaranteed. This problem is tackled by an exemplary application of the double bootstrap (Beran, 1988). The distribution of p-values is compared to the simulated distribution under the alternative model in Section 3.2.

In Section 3.3, the application of cluster analysis methods to populations of taxon ranges is discussed briefly. A concluding discussion is given in Section 4.

2 The tests

2.1 Null model

The non-occurrence of taxa clusters is modeled so that all taxa $A_1, \ldots, A_n \subseteq R$ are generated independently according to the same probabilistic routine. Therefore, we define a model for a single taxon, and the null model for the whole population is that n taxa are generated independently according to the single taxon model.

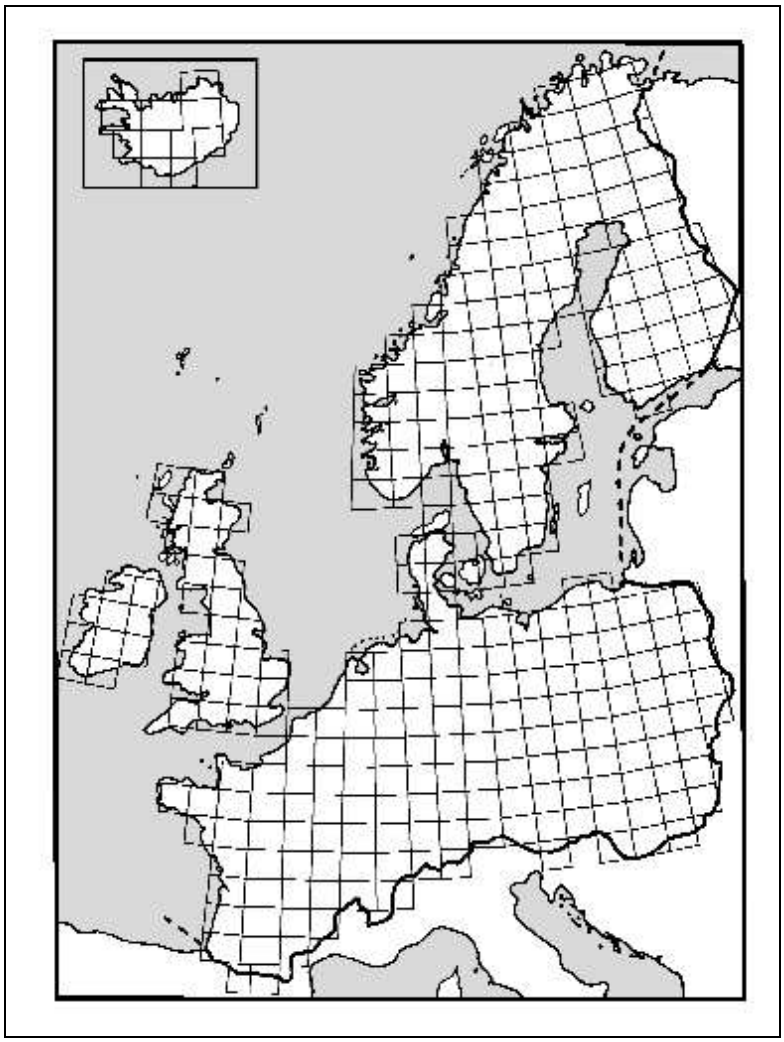

Figure 1: Geographic cells of the north-west European snails data.

Some notation: let P_A be the distribution of the number of units per taxon, i.e., the distribution of $|A|$ for a taxon range A. We will estimate P_A by the empirical distribution $\hat{P}_A$ of taxon ranges in the dataset. Let P_C be a distribution over the units, which determines the probability of a unit to be drawn as a new part of the taxon range we are about to generate. As described below, P_C is used conditionally on a subset of candidate cells, which changes in every generation step. We interpret P_C as describing the potential of the environmental conditions of the units to harbour the species. We will estimate P_C by

$$\hat{P}_C\{i\} = \frac{|\{A_j,\ j = 1, \ldots, n\ :\ i \in A_j\}|}{\sum_{j=1}^{n} |A_j|} \text{ for } i \in R = \{1, \ldots, c\}, \tag{1}$$

i.e., the ratio of the number of taxa in unit i to the sum of all occupations of units by all taxa. Histograms of $\hat{P}_A$ and the distribution of the number of taxa per cell for the example datasets are given in Figure 4.

For each unit i define $N(i)$ to be the set of its neighbors, i.e., all units which have a common border with i.

The null hypothesis should have the interpretation that all clustering of taxa can be attributed to the spatial autocorrelation of the taxon ranges, the structure of the richness

potentials of the units and the variation in range sizes. Therefore we define a null model dependent on the distributions P_A and P_C, and on an autocorrelation parameter p_{disj} defined below, which has also to be estimated from the data.

To generate a single range A, we proceed as follows: in every step the range grows by one unit. Usually the new unit is a neighbor of the previously occupied cells, but with a probability of p_{disj}, the new unit occurs in the non-neighborhood. In reality, the areas occupied by a single taxon consist of few connected groups of cells. Sometimes the whole area is connected. The model should produce taxon distribution patterns with realistic autocorrelation by utilizing the concept of neighborhood. Note that it is not intended to formulate a dynamic model for the real evolution of the taxa. In detail:

Step 0. Generate $r = |A|$, the number of units, randomly from P_A.

Step 1. The first unit i is generated from the richness distribution P_C, i.e., $B_1 = \{i\}$ with probability $P_C\{i\},\ i \in R$. B_j always denotes the set of units occupied by the taxon after Step j. If $r = 1$, set $A = B_1$ and end. Else for $s = 2, \ldots, r$:

Step s.1. Let N_{s-1} be the set of all neighbors of the units of B_{s-1}, and let R_{s-1} be the set of all non-neighboring units that are not occupied by the taxon until step $s-1$, i.e.,

$$N_{s-1} = \left(\bigcup_{i \in B_{s-1}} N(i) \right) \setminus B_{s-1}, \quad R_{s-1} = R \setminus (N_{s-1} \cup B_{s-1}) .$$

We use a parameter p_{disj} as the probability that the new unit for the taxon causes a disjunction, which means that it occurs in the non-neighborhood of the previous area. In Section 2.2 we discuss how this parameter can be estimated from the data. Given that N_{s-1} and R_{s-1} are non-empty, Step s.2 is performed with probability p_{disj} and Step s.3 is performed else. If $N_{s-1} = \emptyset$, perform Step s.3, if $R_{s-1} = \emptyset$, perform Step s.3.

Step s.2 (new non-neighbor). $B_s = B_{s-1} \cup \{i\}$, where i is chosen randomly according to P_C restricted to the points of R_{s-1}. For $i \in R_{s-1}$:

$$P\{i\} = \frac{P_C\{i\}}{\sum_{j \in R_{s-1}} P_C\{j\}}.$$

Unless $s = r$, let $s = s + 1$ and go to Step s.1.

Step s.3 (new neighbor). $B_s = B_{s-1} \cup \{i\}$, where i is chosen randomly according to P_C restricted to the points of N_{s-1}. For $i \in N_{s-1}$:

$$P\{i\} = \frac{P_C\{i\}}{\sum_{j \in N_{s-1}} P_C\{j\}}.$$

Unless $s = r$, let $s = s + 1$ and go to Step s.1.

Step $r + 1$. Set $A = B_r$, end.

2.2 Estimation of the involved parameters

The parameter p_{disj} used for the generation of a range under null model is the probability that a new unit, which is assigned to a range in Step s, is not a neighbor of the previous area B_{s-1}.

This probability has a counterpart in the real dataset. For $i = 1, \ldots, n$, let a_i be the number of connected areas of the distribution pattern of the taxon A_i. a_i can be determined by the "depth-first search" algorithm explained in Cormen, Leiserson and Rivest (1990, p. 477) applied to the neighborhood graph of the units of the taxon. We call $a_i - 1$ the "number of disjunctions" of a taxon, because if the taxon would be generated by our null model, this would be the number of necessary new units that do not occur in the neighborhood of the previous pattern (the first unit is not counted as a disjunction). The probability of disjunction can then be estimated by

$$\hat{q}_{disj} = \frac{\sum_{i=1}^{n}(a_i - 1)}{\sum_{i=1}^{n}(|A_i| - 1)}. \tag{2}$$

$\hat{q}_{disj}$ is not a very good estimator for p_{disj}, because $\hat{q}_{disj}$ does not account for some situations that may occur during the generation of a taxon according to the null model:

- It may happen that $N_s = \emptyset$ or $R_s = \emptyset$. In these cases, p_{disj} does not apply and a jump is forced or impossible, respectively.
- It may happen that a new connected area emerges if areas, which started from non-neighboring units, meet later in the taxons growing process.

However, the relation between $\hat{q}_{disj}$ and p_{disj} under the null model can be simulated: several populations can be generated for various values of p_{disj} ($\hat{P}_A$ and $\hat{P}_C$ held constant), and the corresponding value of $\hat{q}_{disj}$ can be calculated by (2). In our experience the relation can very well be fitted by a linear regression, as can be seen e.g. in the Figure 2 for the example datasets:

$$\hat{q}_{disj} = b_0 + b_1 p_{disj} + \text{Error}.$$

$\hat{b}_0, \hat{b}_1$ denote the least squares estimators of b_0 and b_1, and an estimator $\hat{p}_{disj}$ can be calculated by computing the p_{disj}-value giving raise to the real data value $\hat{q}_{disj,real\ data}$ (dashed line in Figure 2) according to the estimated regression line:

$$\hat{p}_{disj} = \left(\hat{q}_{disj,real\ data} - \hat{b}_0\right) / \hat{b}_1. \tag{3}$$

The values of the covariate p_{disj} used in the simulations shown in Figure 2 have been chosen somewhat experimentally. As a default, we suggest to estimate a preliminary $\tilde{p}_{disj}$ by simulating four datasets for each of $p_{disj} = 0, 0.1, \ldots, 0.9, 1$. To estimate $\hat{p}_{disj}$, we choose $p_{disj} = \tilde{p}_{disj} - 0.1, \tilde{p}_{disj} - 0.09, \tilde{p}_{disj} - 0.08, \ldots, \tilde{p}_{disj} + 0.09, \tilde{p}_{disj} + 0.1$ four times each.

While the distribution of the size of the taxon ranges in the null model is governed directly by the empirical distribution $\hat{P}_A$, the approximation of P_C is more complicated, because the numbers of taxa per unit are not generated directly. There are two problems:

1. Units at the border of the whole area and units with few neighboring cells are more difficult to reach by Step $s.3$, and therefore they may occur too seldom in setups where p_{disj} is low. This problem could be overcome to some extent by adjusting $\hat{P}_C$ to the number of neighbors:

$$\hat{P}_{CN}\{i\} = \frac{\hat{P}_C\{i\}/\min(1, |N(i)|)}{\sum_{j=1}^{c}\hat{P}_C\{j\}/\min(1, |N(j)|)} \text{ for } i \in R = \{1, \ldots, c\}.$$

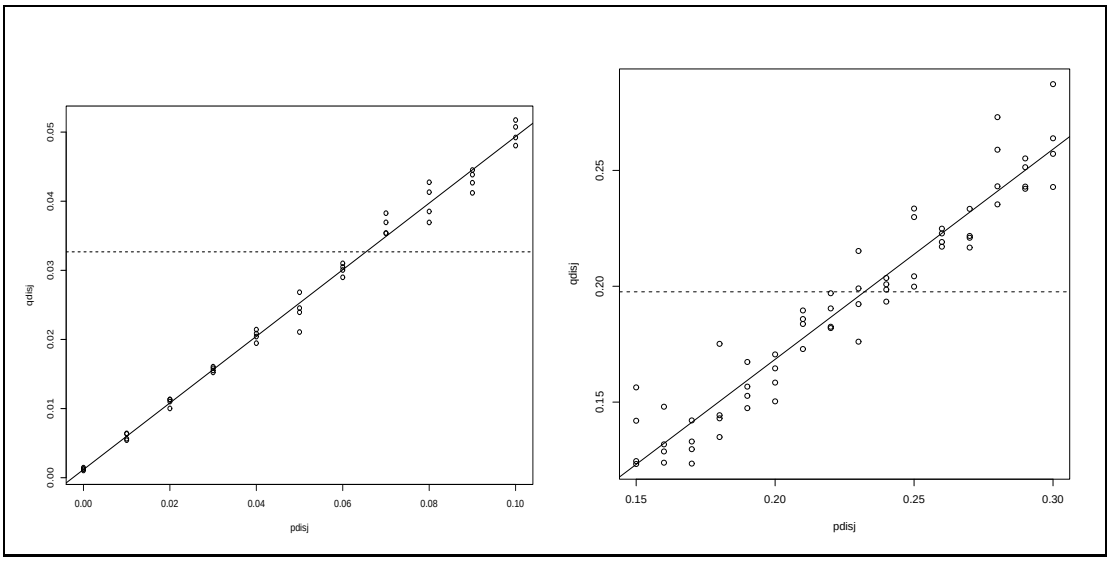

Figure 2: Relation of p_{disj} and $\hat{q}_{disj}$ for simulated data with the parameters of the north-west European data (left side) and the Israeli data (right side). The dashed line corresponds to the observed $\hat{q}_{disj,real\ data}$.

2. A unit cannot be drawn more than once during the generation of a taxon, so that units with a large richness value will necessarily be under-represented in large taxon ranges. Figure 3 shows the approximation of the number of taxa per unit in the north-west European data (x-axis) by the corresponding numbers of five datasets generated from the null model (y-axis; i.e., there are five points for each of the $n = 306$ units). Ideally, the values should be distributed symmetrically about the solid line with unit slope. This is fairly well fulfilled, but the largest numbers of taxa per unit (about 120 and above) lead to somewhat too low values under the null model. Our experience indicates that such a pattern is typical. This problem could be solved by performing some nonlinear regression on simulated data such as those shown in Figure 3, but we leave this for future research. An analogous phenomenon is reported even for established null models that ignore the spatial autocorrelation (Cook and Quinn, 1998).

2.3 Distances

Our test statistics will be based on a distance measure between species ranges. We think that it is inadequate to treat the taxon ranges, characterized as 0-1 vectors, as metric data, as would be necessary to apply most of the test statistics given in Bock (1996). Recent work on testing homogeneity in coincidence graphs, as which our data can be represented, is motivated by null models that do not account for spatial autocorrelation (Godehardt and Jaworski, 2003).

A variety of (dis-)similarity measures between species ranges has been proposed (Cheetham and Hazel, 1969). The most popular dissimilarity measure in biology is the Jaccard distance:

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

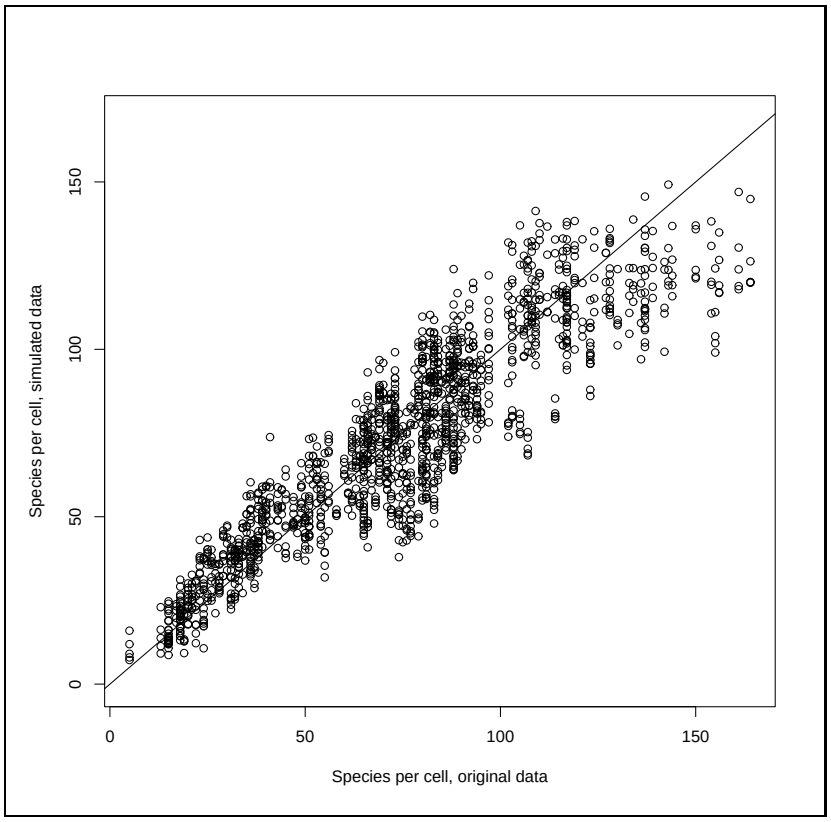

Figure 3: Number of species per cell for north-west European data (x-axis) vs. five simulated populations from the null model.

This measure has a serious drawback for our purposes. It yields a very large distance value if one of the species ranges, A_1, say, is much smaller than the other, even if the smaller one is completely included in the larger one. In this case, the denominator equals $|A_1|$, and the size of the common occurrence of the species is not related to $|A_2|$.

One of our aims is to test the validity of the vicariance biogeography (Nelson and Platnick, 1981; Wiley, 1988; Humphries and Parenti, 1999), which claims that new taxa have emerged by the fragmentation of ancestral biota by the appearance of a barrier. Thus, biotic elements, i.e., clusters of ranges, should be found on both sides of the barrier, but there is no necessity to expect that the ranges of the taxa belonging to the same biotic elements should have more or less the same size. Therefore we work with the Kulczynski distance, which equals to one minus the "2nd Kulczynski coefficient" given in Cheetham and Hazel (1969), which balances the relation between the area of common occurrence $|A_1 \cap A_2|$ and the two range sizes in an intuitive way:

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

d_K equals 0 iff the taxa are identical and 1 iff the taxa are disjunct. Note that d_K is not a metric. The triangle inequality holds very often, but can be violated: consider A_1 and A_2, both with only one cell, having no cell in common but $|A_1 \cap A_3| = |A_2 \cap A_3| = 1$, $|A_3| = 2$. Then, $d_K(A_1, A_2) = 1$, $d_K(A_1, A_3) = d_K(A_2, A_3) = \frac{1}{4}$.

More sophisticated distances may be constructed by taking also the spatial distances between the non-common cells of the taxa into account.

2.4 Test statistics

Testing a homogeneity hypothesis against a clustering alternative based on distance measures is a difficult task. The literature on this topic concentrates mainly on test statistics for which exact or approximative distributions can be derived under the assumption of simple homogeneity models. Surveys are given in Bock (1994), Bock (1996), Ripley (1981, Chapter 8).

Since such homogeneity models are not realistic in autocorrelated spatial structures, and the distribution of our test statistics will be derived by means of a Monte Carlo simulation, this kind of distribution theory is not useful in our setup, and we are not restricted to the use of statistics that are motivated by such a theory. Instead, we aim at formalizing directly what "clustering" means in terms of distances. The resulting tests may be applied not only in our setup, but generally where distances are used to detect clusters in the data and no simple homogeneity model is applied.

We distinguish two approaches to define test statistics based on distances. The first approach is to investigate the properties of a graph, where the species are interpreted as vertices and the smallest m distances, m being a tuning constant, define the edges.

S_1: **isolated vertices.** A classical test statistic is the number of isolated vertices S_1 in the resulting graph, for which a distribution theory under some homogeneity models for distances has been derived by Ling (1973) and Godehardt and Horsch (1995). The drawback of this test statistic is that a clustered dataset may lead to more, fewer or about the same number of isolated vertices as homogeneous data: clusters in the data may cause edges that are highly concentrated on few vertices and thus many isolated vertices, near neighbors for all points and thus few isolated vertices, or a compromise between both. Godehardt (personal communication) argues that the resulting test should be two-sided.

S_2: **largest connectivity component.** We propose as an alternative S_2, the number of vertices of the largest connectivity component. A small value of S_2 indicates always that the edges are concentrated and that groups of vertices are not linked. This can be interpreted as indicating a cluster structure, while homogeneous data will lead to a large connectivity component for m not much smaller than n.

In some homogeneity models, S_1 and S_2 are related. Erdös and Renyi (1960) have shown that under equiprobable occurrence of edges and
$m = (n-1)(\log n + r + O(1))/2$ for some constant r the number of isolated vertices and the number of connectivity components minus one converge to the same distribution. Thus, the dataset can asymptotically be expected to consist of one very large connectivity component plus isolated vertices. r should be negative to define a reasonable two-sided test based on S_1. Godehardt and Horsch (1995) show that similar asymptotic results hold under a model for edges stemming from metric distances, but the result for the number of connectivity components needs twice as much edges as the result for the isolated vertices.

The tuning constant should be chosen in order to enable a large connectivity component for homogeneous data and, simultaneously, to prevent that the occurrence of such a large component is enforced for clustered data. This depends on the structure of the expected clustering. The above cited asymptotic theory suggests to choose $m = d(n-1)(\log n + r)$ for some negative r and $d = 0.5$ for S_1, $d = 1$ for S_2. We worked with $r = -3.25$. This leads for the smaller of our analyzed datasets to $m \approx 0.75n$ for S_2, which we think to be a reasonable proportion for small datasets. Asymptotically, $r = -3.25$ yields a mean of 25.8 isolated vertices under the theory of Godehardt and Horsch (1995), which, however, does

not account for our particular data structure. Note that S_1 and S_2 are closely related to Single Linkage clustering, where clusters defined by cutting the tree are the connectivity components of the graph defined by the distances below the cut-point.

A second approach is to define a test statistic based on the size of the smallest distances, which should be small inside of clusters.

S_3: **nearest neighbors.** Manly (1997, Chapter 10) suggests the average distance of the points (taxa) to their mth nearest neighbor, m being again the tuning constant, as test statistic for a Monte Carlo test. We denote this test statistic by S_3. m should be chosen as about the size of the smallest group of points which can be accepted as a meaningful cluster.

The drawback of this statistic is that it may obscure an existing cluster structure in the presence of few outlying points which have very large distances to their nearest neighbors. In our setup, this problem may not be too severe, because the Kulczynski distance has a maximum value of 1, while values of about 0.1 occur typically inside of clusters. However, we think that not only the smallest, but also the largest distances should be taken into account, because a dataset may produce smaller distances than a null model for all pairs of points, and this does not indicate clustering.

S_4: **distance ratio.** As an alternative, we define S_4 as the ratio between the sum of the smallest m and the sum of the largest m distances. A strategy for the choice of m is that m would be optimal if the smallest m distances would be the distances inside the clusters (under assumption of their presence) and the largest m distances would all lie between the clusters (or between points not belonging to any cluster). However, we expect the test not to be too sensitive against the choice of m, because under arbitrary cluster structures the m smallest distances will include much more within-cluster distances than the m largest distances. We propose $m = n(n-1)/8$ as a default choice, which means that 50% of the distances are used to compute the statistic. As opposed to S_3, S_4 may be driven to reject homogeneity by extreme outliers even if clusters do not exist, but such a dataset should also not be interpreted as homogeneous.

2.5 Alternative model

To perform power comparisons, we specify an alternative model to formalize clustered taxa ranges. The alternative model is motivated by the vicariance theory (cf. Section 2.3), according to which an ancestral biota was fragmented by the appearance of a barrier so that two new taxa emerged out of one. For comparability, the alternative model should be able to reproduce the distributions P_A and P_C and the autocorrelation parameter p_{disj} from the null model. The idea is as follows: we split the cells in two subregions R_1 and R_2 and define $p_1 = P_C(R_1)$. If available, knowledge about geographical barriers can be implied. We choose $n \geq n_1 > 0$ such that $n = n_1 + n_2$. n_1 is interpreted as the number of taxa resulting from a vicariance event. n_2 is the number of species not affected by the barrier. We choose some constant $f < 1$ to determine the factor by which the probability is multiplied that a taxon with originated on one side of the barrier R_j enters a cell from the other side R_k, $k \neq j$ (which may happen if the vicariance event has been in the very past), $f = 0.01$, say.

The n_2 unaffected taxon ranges are generated as in Section 2.1. The principle for the generation of the n_1 taxon ranges originating from the vicariance event is as follows: we

define a richness distribution for species originated in R_k, $k = 1, 2$:

$$P_{Ck}\{i\} = \frac{f_i P_C\{i\}}{\sum_{j=1}^{c} f_j P_C\{j\}}, \text{ for } i \in R = \{1, \ldots, c\},$$

where $f_i = 1$ for $i \in R_k$, $f_i = f$ for $i \notin R_k$ else.

Step 0.5. With probability p_1 let $k = 1$, else $k = 2$. The taxon range originates in R_k.

Step 1. The first cell is drawn according to P_{Ck}, but with $f_i = 0$ for $i \notin R_k$ (the generation has to start in R_k always).

Step s**.** As in Section 2.1, but with P_C replaced by P_{Ck}.

We focus on comparability of the parameters with the null model, and therefore the model cannot be interpreted as formalizing a real vicariance process. The main difference is that the origin of the taxon (and the whole range, if it is small enough) might be arbitrarily far away from the modeled barrier; only the side with respect to the barrier is determined.

However, the alternative model should produce clustered taxon ranges, and the degree of clustering is governed by the number of taxa n_1 originating from vicariance. Note that even $n_1 \approx n$ does not necessarily guarantee a very strong cluster structure, because very small taxon ranges will usually be restricted to R_1 or R_2 even in the null model, while very large taxon ranges will extend across the barrier in the alternative model as well. Significantly clustered data can be expected in situations where the range size distribution P_A generates ranges of medium size with high probability. The subregions should be chosen so that $p_1 \approx 0.5$ to guarantee enough taxon ranges to belong to each "model cluster". At least under strong spatial autocorrelation, i.e., small p_{disj}, R_1 and R_2 should be (more or less) connected in themselves and connected to each other, because the taxa should be concentrated in one of the subregions under the alternative model with high probability, while not too few taxa should be present in both subregions under the null model.

3 Application

3.1 Data examples

In this section, we apply the tests to two datasets. One of them consists of 366 ranges of land snail species in 306 cells of north-west Europe as shown in Figure 1, the other consists of 55 land snail species ranges in 251 cells in Israel and Palestine. For the north-west European dataset, almost all species present in the study area have been included. The Israeli dataset is restricted to the species with a range boundary in Israel. The north-west European data set has been compiled from range maps provided by Kerney et al. (1983). The Israeli data set is is derived from the database of the Israel National Mollusc Collection in the Hebrew University of Jerusalem (see also Kadmon and Heller (1998)).

Some characteristics of the data, namely the histogram of the range size distribution $\hat{P}_A$ and the histogram of the species richness of the cells are shown in Figure 4. The parameter p_{disj} has been estimated according to (3) as shown in Figure 2. For the north-west European data we get $\hat{p}_{disj} = 0.065$, which indicates a very strong spatial autocorrelation where most species ranges are connected. For the Israeli data we get $\hat{p}_{disj} = 0.231$. The difference between the disjunction parameters for the north-west European data and the Israeli data is due to the different quality of the data. The European data are derived

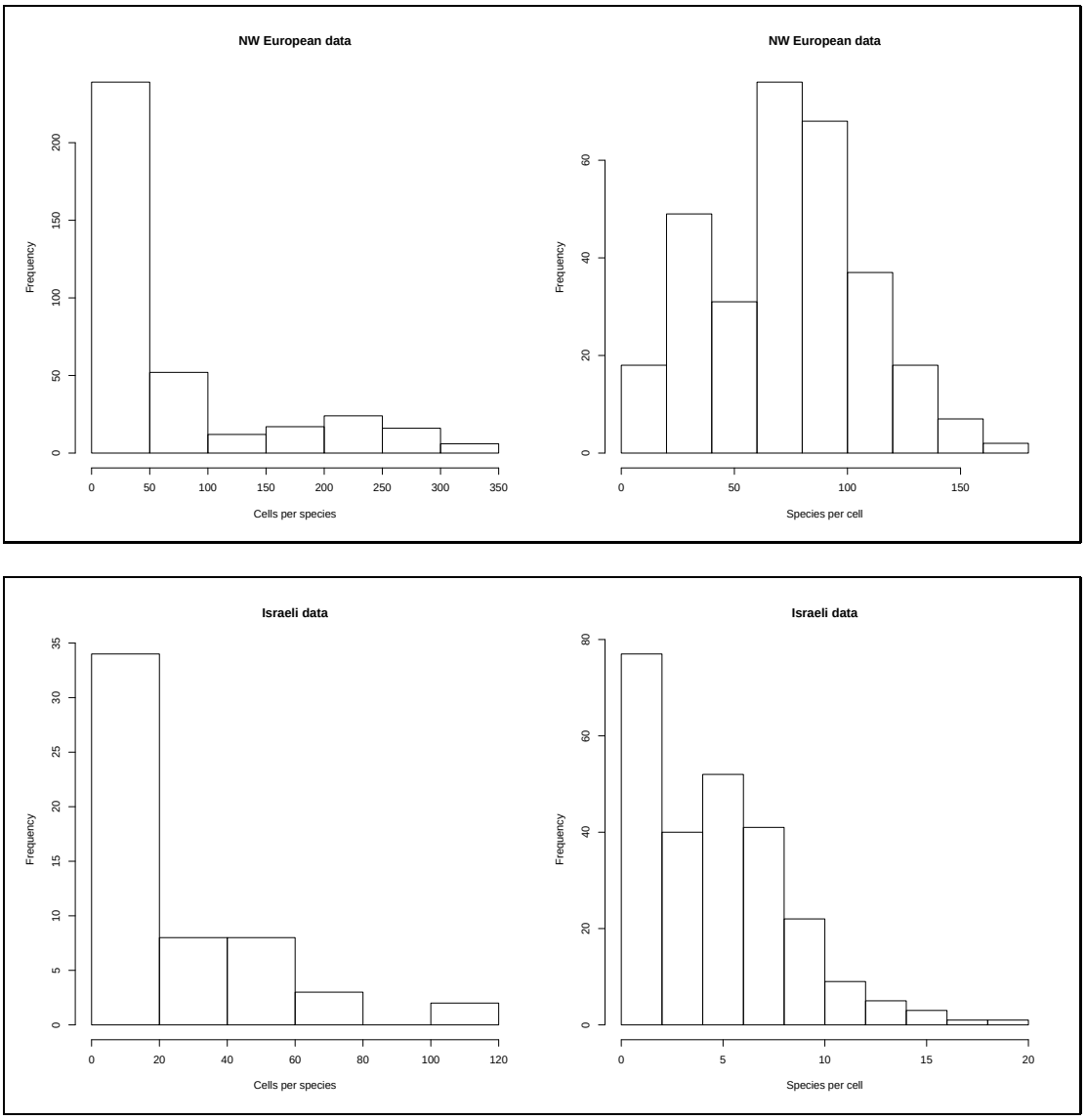

Figure 4: Histograms of cells per species (P_A, left) and species per cell (right) for north-west European data (above) and Israeli data (below).

from interpolated range maps, whereas the Israeli data are true records in the respective cells.

Figure 5 shows an MDS representation of the Kulczynski distances of the datasets. The non-metric MDS of Kruskal (1964) has been used. Both datasets appear neither clearly clustered, nor clearly homogeneous.

The p-values of the tests are computed with 200 simulation runs for the north-west European data and 1000 simulation runs for the smaller Israeli dataset. For the choice of the tuning constants see Section 2.4.

The north-west European dataset yielded for the statistics S_1 ($m = 484$), S_3 ($m = 3$) and S_4 a value smaller than the smallest value obtained during the simulations, i.e., $p = 1/201$ for S_3 and S_4 and $p = 2/201$ for S_1 (two-sided). The tests indicate clearly that the dataset deviates from the null model.

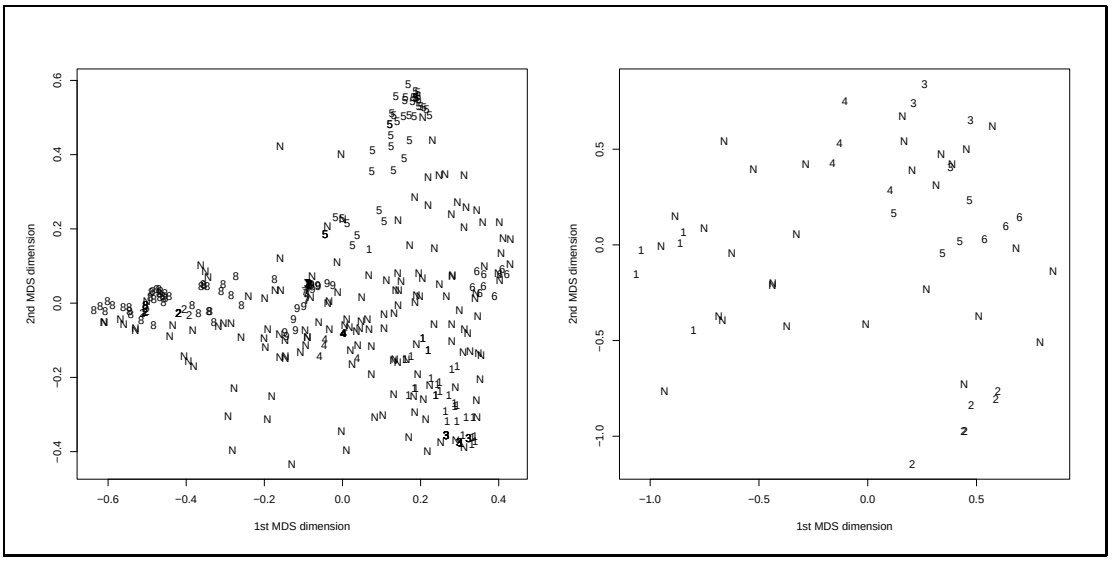

Figure 5: First two dimensions of non-metric MDS. Characters indicate model based clustering with noise ("N"). Left side: north-west European data, right side: Israeli data.

An interesting result is that the dataset generates the second largest value of S_2 (largest connectivity component, $p = 200/201$, $m = 967$), i.e. this test indicates the opposite of clustering. The result has been confirmed with other choices of the tuning constants. A closer look at the data indicated that many of the edges of the graph, on which the test has been performed, stemmed from situations where a much larger taxon range included a much smaller range, leading to a Kulczynski distance of about 0.25-0.45. Such relations between species are referred to as "nestedness" in biogeography (Patterson and Atmar, 1986; Wright and Reeves, 1992; Cook and Quinn, 1998; Wright et al., 1998). A hierarchy of such inclusions can be constructed, which links most of the taxa of the dataset. The large size of the largest connectivity component can be explained by this phenomenon. The nested species build chains between the small clusters which are apparently also present in the dataset. Such effects are known in Single Linkage clustering, to which the test statistic S_2 is related. We confirmed the extraordinary nestedness of the dataset by constructing a nestedness test using the same null model (Hausdorf and Hennig, 2003).

The test results of the Israeli dataset indicate also that the apparent non-homogeneity cannot only be attributed to spatial autocorrelation and variation in the species richness of the cells. The dataset resulted in the smallest test statistic values compared to the simulations ($p = 1/1001$) for S_2 ($m = 41$), S_3 ($m = 3$) and S_4. The number of isolated vertices S_1 ($m = 21$) is significantly large (two-sided $p = 18/1001$) for these data.

3.2 Simulations: double bootstrap and power

Since the autocorrelation and the distributions P_A and P_C are to be estimated from the data, our simulation of the p-value is a "parametric bootstrap" (Davison and Hinkley, 1997), and the validity of the p-value is not guaranteed. The accuracy of the p-values can be improved by the adjustment via a double bootstrap algorithm (Beran, 1988; Davison and Hinkley, 1997). The principle of the double bootstrap adjustment is that the original p-value can itself be interpreted as the test statistic. The whole simulation of the p-value (of which the number of simulation runs is denoted by N) has to be repeated for K datasets

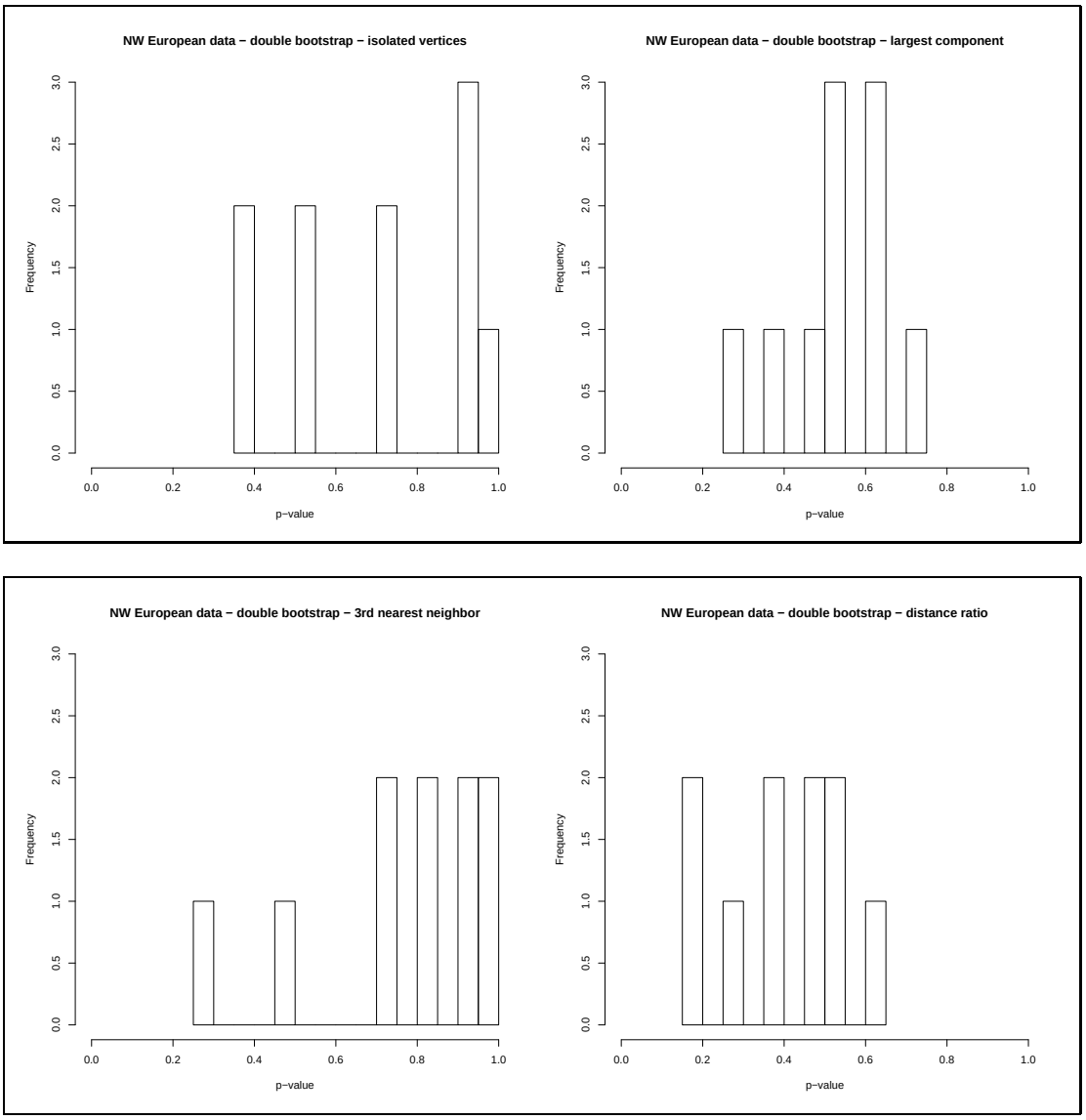

Figure 6: Histograms of p-values from double bootstrap ($K = N = 10$) for NW European data: S_1 (above left), S_2 (above right), S_3 (below left), S_4 (below right).

generated from the null model with the parameters estimated from the data, while the parameters for the KN null model datasets used in the simulation runs are estimated from the K datasets of the first bootstrap stage. The adjusted p-value is then the Monte Carlo p-value of the original p-value compared to those of the K simulations. The double bootstrap simulation is computationally very expensive. Therefore we performed it with moderate numbers of simulation runs ($K = N = 10$ for the north-west European data and $K = N = 100$ for the Israeli data) not to adjust our p-values, but to assess informally their validity. To save computing time, p_{disj} has been estimated based on only 11 covariate values symmetricallky about $\hat{p}_{disj}$ for the original dataset, with four simulated datasets for each covariate value, compare Section 2.2.

The results can be seen in the Figures 6 (NW European data) and 7 (Israeli data). For the north-west European data, the histograms based on 10 simulations can only give

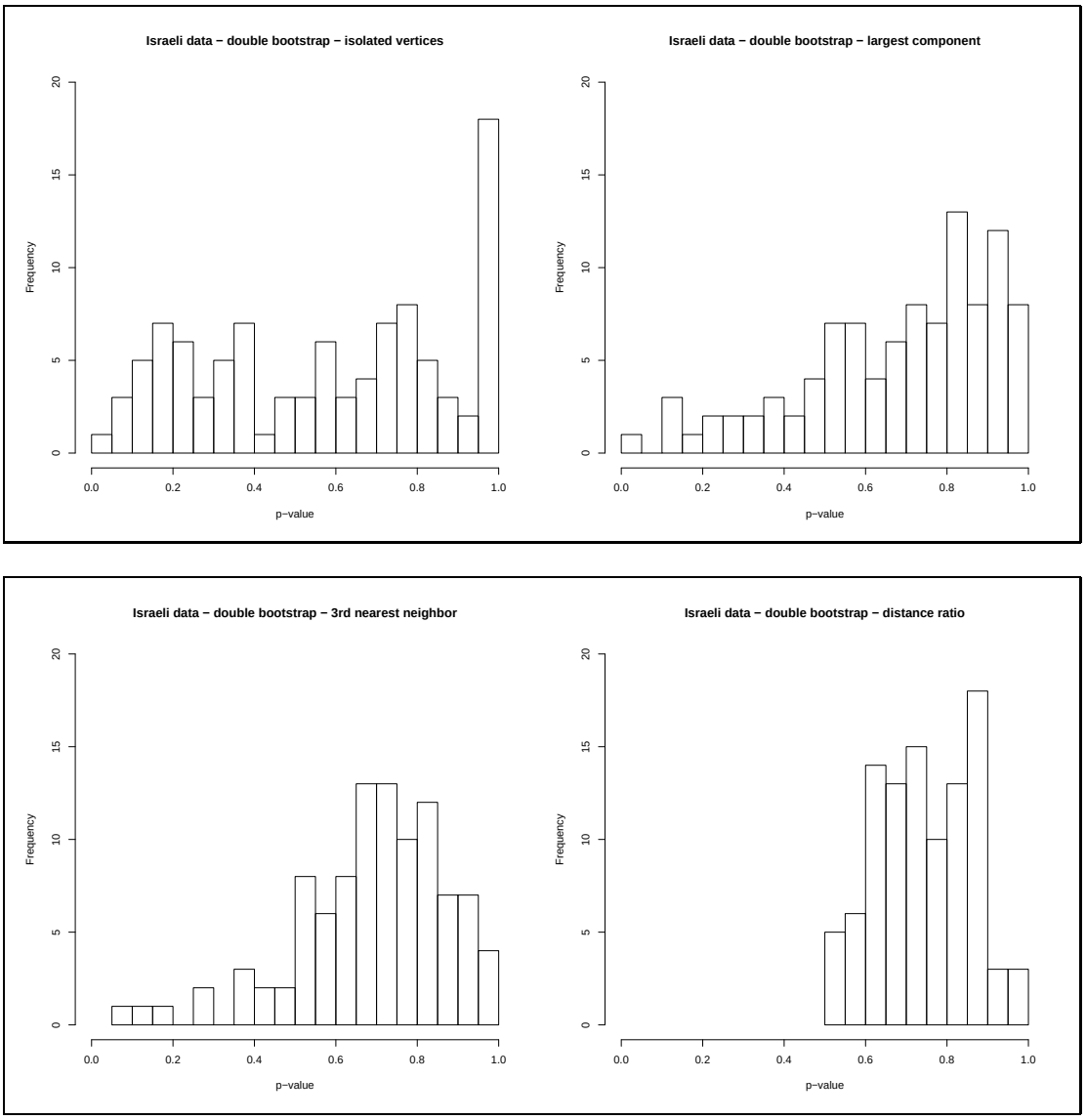

Figure 7: Histograms of p-values from double bootstrap ($K = N = 100$) for Israeli data: S_1 (above left), S_2 (above right), S_3 (below left), S_4 (below right).

a very rough impression. We presume that the variance of the p-values is a bit smaller those of the "ideal" uniform distribution on $[0, 1]$. This can be expected for a parametric bootstrap, because the estimation of the parameters may induce more variability into the simulations, and therefore the p-value of the source data (bootstrap sample of the first stage) may seem more stable compared to simulated data (bootstrap samples of the second stage) than for a Monte Carlo simulation without nuisance parameters. Since all generated p-values are larger than the smallest possible value of $1/11$, the significance of the test statistics computed for the original dataset is confirmed. The histograms for S_1 and S_3 give us the impression that the corresponding tests are conservative.

The results for the Israeli dataset are shown in Figure 7. As for the north-west European dataset, they indicate a smaller variance of the p-values as would be expected for a uniform distribution, and they confirm the significant results of the original tests, because

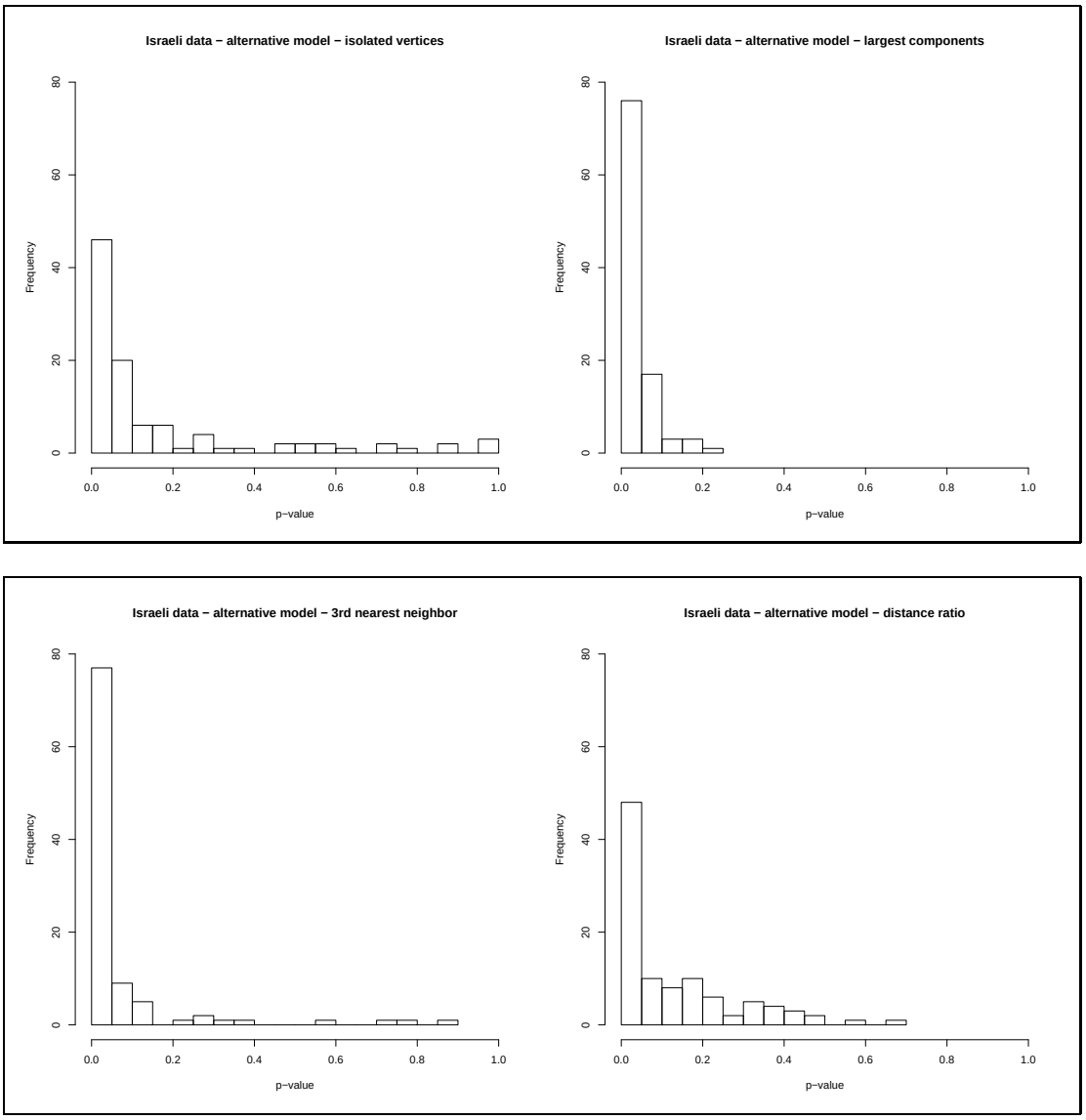

Figure 8: Histograms of p-values of the parametric bootstrap applied to data generated from alternative model with parameters from Israeli data ($K = N = 100$): S_1 (above left), S_2 (above right), S_3 (below left), S_4 (below right).

all averages are clearly above 0.5 and demonstrate all tests to be more or less conservative. As opposed to the north-west European dataset, the most conservative test statistic for this setup seems to be S_4, for which all p-values are above 0.5.

As a comparison, we carried out a simulation of the p-value of the test proposed by Ling (1973), which is a one-sided test based on the statistic S_1. A distance based homogeneity model, under which the distribution of the test statistic can exactly be calculated, is assumed. 100 datasets have been generated from the null model with the parameters from the Israeli data. 38 of these yielded a p-value of below 0.05, indicating that the variation in species richness and the spatial autocorrelation alone lead often to the rejection of such a homogeneity model.

We also carried out a simulation of the power of the tests against the alternative

model introduced in Section 2.5. We used the parameters and the neighborhood structure of the Israeli dataset, i.e., 55 species and 251 cells. R_1 consists of the cells 1 to 142 with $p_1 = 0.5004$. No geographical background knowledge was used, but the demands at the end of Section 2.5 are fairly well fulfilled. We chose $n_1 = 50$ (species belonging to the model clusters), $n_2 = 5$, $f = 0.01$. As in the double bootstrap simulations, there were $K = 100$ generated datasets from the alternative model and $N = 100$ parametric bootstrap samples have been generated each time to compute the p-value. The distributions of the p-values are shown in Figure 8. By comparing the histograms to those of Figure 7, it can be seen that all tests are able to distinguish clearly the alternative model from the null model. The proportion of p-values below 0.05 (estimated power of the 5%-level test) has been 0.46 for S_1, 0.76 for S_2, 0.77 for S_3 and 0.48 for S_4. Note that all significant p-values for the two-sided test with statistic S_1 have been caused by too many isolated vertices in the data from the alternative model. As in the double bootstrap study for the Israeli data, S_4 comes out as the most conservative test. However, Figure 6 suggests that this may depend on the geographical structure of the data.

3.3 Cluster analysis

Since we treated the taxon ranges as distance data in the tests, it may seem natural to use distance based cluster analysis methods such as agglomerative nesting (e.g. Average Linkage or Single Linkage, Kaufman and Rousseeuw (1989, Chapter 5)) or "PAM" (Kaufman and Rousseeuw, 1989, Chapter 2) to perform a cluster analysis on such data. However, we suggest an alternative strategy:

- Map the distance data to $\mathbb{R}^k$ by multidimensional scaling, the nonlinear MDS of Kruskal (1964), say.
- Apply a cluster analysis method which allows for flexible shapes of the clusters and "noise", i.e. points that do not belong to any cluster.

The cluster analysis method could be model-based clustering with noise component as discussed in Fraley and Raftery (2002). The data is modeled by a mixture of Normal distributions and points not belonging to any cluster are modeled by a uniform distribution on the convex hull of the data. The number of clusters and a model for the covariance matrices of the clusters is determined by the Bayesian Information Criterion.

Some properties of the data suggest this approach:

- The MDS plots (cf. Figure 5) suggest, in agreement with biological evidence, that there are lots of taxa ranges which should not be assigned to any biotic element (cluster of taxon ranges), even if clusters exist. The noise component is an elegant way to account for these ranges.
- Dependent on the size and properties of the region where a biotic element is located, its taxon ranges show more or less variation. This can be modeled by varying covariance matrices of Normal mixture components.
- While the Normal assumption for the clusters can hardly be justified, the Normal mixture model together with the various possible restrictions to the covariance structures of the components provides a very rich class of models for the clusters. In terms of distances, the Normal model means that a cluster consists of a core group of objects which are very near to each other ("peak of the bell"). This group may be surrounded by some objects with somewhat larger distances, which, however, can

be separated from noise and other clusters. The assumption that these further objects appear symmetric around the core group may be somewhat questionable, but hierarchical distance based methods rely on implicit assumptions about the shapes of clusters which, from our viewpoint, are similarly difficult to justify.

The automatic choice of the optimal number of clusters (and covariance matrix model) may be seen as another advantage, but it has to be paid by a necessary decision about the MDS method and the number of MDS-dimensions k. Figure 5 shows the clusterings from model-based cluster analysis with noise applied to 4 MDS-dimensions. The first two dimensions are plotted. An initial noise estimation from the procedure `NNclean` (Byers and Raftery, 1998) has been used as recommended in Fraley and Raftery (1998).

Note that we do not state that MDS followed by Normal mixture fitting should generally be applied to distance data. However, we observed that in our setup it is a valuable alternative. From a biogeographical point of view, the solutions were more useful than those obtained from PAM and hierarchical distance-based methods, as far as we tested them. In either case, we recommend to look at the MDS plots for diagnostic purposes.

Fixed point clustering (Hennig and Christlieb, 2002) and the fit of mixtures of multivariate t-distributions (McLachlan and Peel, 2000, Chapter 7) may also be applied to the outcome of the MDS in our setup.

4 Conclusion

In the present paper we have treated various issues raised by the question if a dataset of taxon ranges characterized by presence/absence information is significantly clustered:

- A null model has been proposed, which accounts for spatial autocorrelation and the variation in species richness potentials of the units.
- Two new distance based test statistics for clustering have been proposed and compared with two already known approaches. The statistics can be applied to any distance data.
- Besides the application of the corresponding parametric bootstrap tests, the results have been validated by a double bootstrap simulation and the power of the tests has been assessed by exemplary simulations based on an alternative model.
- The choice of a distance measure and a clustering method adapted to our setup have been discussed.

We do not conclude with a recommendation of an "optimal" method. We gained valuable information from combining and explaining contradictory results (three out of four tests for clustering have been significant for the NW European data, while the largest connectivity component test pointed in the opposite direction), and from performing MDS and the double bootstrap in a diagnostic spirit. Even the double bootstrap with only ten replications for the NW European data was able to improve the confidence in the result of the parametric bootstrap.

Some words about the proposed test statistics: the sensitivity of the largest connectivity component statistic with respect to "chaining" has to be known. The corresponding test may give valuable insight, though. For the reasons given in Section 2.4, we prefer the distance ratio test to the nearest neighbor test, even if the simulation results of the latter have been flawless for our data. We believe that the conservativeness of the tests depends

on the data structure and has to be examined by double bootstrap for every application. The isolated vertices test will fail to detect some clusterings, because compromises between the situations of our two datasets, of which one led to a significantly small value and the other led to a significantly large value, will occur in practice.

The results of the tests should not be interpreted in a traditionally frequentist manner, because our ensembles of taxa ranges cannot be replied. Instead, we interpret significant test results as "missing features" (Davies, 1995) of the datasets compared with the null model. They may stimulate the discussion about scientific hypotheses, but these hypotheses can never "statistically be proven" by any data.

Software to compute the tests and generate data from null and alternative model will soon be available from the web page

`http://www.math.uni-hamburg.de/home/hennig`

Acknowledgements

We are grateful to J. Heller and O. Steinitz for providing the Israeli distribution data and to Hans-Rudi Roth for helpful comments.

References

R. Beran, Prepivoting Test Statistics: A Bootstrap View of Asymptotic Refinements. *Journal of the American Statistical Association* **83** (1988) 687–697.

H.-H. Bock, Classification and Clustering: Problems for the Future, in: Diday, E., Lechevallier, Y., Schader, M., Bertrand, P., Burtschy, B. (Eds.): *New Approaches in Classification and Data Analysis* (Springer, Berlin, 1994) 3–24.

H.-H. Bock, Probability models and hypotheses testing in partitioning cluster analysis, in: Arabie, P., Hubert, L. J. and De Soete, G. (Eds.): *Clustering and Classification* (World Scientific, Singapore, 1996) 377–453.

S. Byers and A. E. Raftery, Nearest Neighbor Clutter Removal for Estimating Features in Spatial Point Processes. *Journal of the American Statistical Association* **95** (1998) 781–794.

A. H. Cheetham and J. E. Hazel, Binary (presence-ansence) similarity coefficients. *Journal of Paleontology* **43** (1969) 1130–1136.

R. C. Cook and J. F. Quinn, An evaluation of randomization models for nested species subsets analysis. *Oecologia* **113** (1998) 584–592.

T. H. Cormen, C. E. Leiserson, and R. L. Rivest, *Introduction to Algorithms* (MIT Press, Cambridge, 1990).

P. L. Davies, Data Features. *Statistica Neerlandica* **49** (1995) 1–47.

A. C. Davison and D. V. Hinkley, *Bootstrap Methods and their Application* (Cambridge University Press, Cambridge, 1997).

P. Erdös and A. Renyi, On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, **5** (1960), 17–61.

C. Fraley and A. E. Raftery, How Many Clusters? Which Clustering Method? Answers Via Model Based Cluster Analysis. *Computer Journal* **41** (1998) 578–588.

C. Fraley and A. E. Raftery, Model-Based Clustering, Discriminant Analysis, and Density Estimation. *Journal of the American Statistical Association* **97** (2002) 611–631.

E. Godehardt and A. Horsch, Graph-Theoretic Models for Testing the Homogeneity of Data, in: Gaul, W. and Pfeifer, D. (Eds.): *From Data to Knowledge* (Springer, Berlin, 1995) 167–176.

E. Godehardt and J. Jaworski, Two Models of Random Intersection Graphs for Classification. To appear in *Proceedings of GfKl 2002, Mannheim* (Springer, Berlin, 2003).

B. Hausdorf, Units in Biogeography. *Systematic Biology* **51** (2002) 648–652.

B. Hausdorf and C. Hennig, Nestedness of north-west European land snail ranges as a consequence of differential immigration from Pleistocene glacial refuges. Submitted (2003).

C. Hennig and N. Christlieb, Validating Visual Clusters in Large Datasets: Fixed Point Clusters of Spectral Features. *Computational Statistics and Data Analysis*, **40** (2002) 723–739.

C. J. Humphries and L. R. Parenti, *Cladistic biogeography. 2nd ed.* (Oxford University Press, Oxford, 1999).

R. Kadmon and J. Heller, Modelling faunal responses to climatic gradients with GIS: land snails as a case study. *Journal of Biogeography*, **25** (1998) 527–539.

L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis* (Wiley, New York, 1989).

M. P. Kerney, R. A. D. Cameron, and J. H. Jungbluth, *Die Landschnecken Nord- und Mitteleuropas* (Parey, Hamburg, 1983).

J. B. Kruskal, Multidimensional scaling by optimizing goodness of fit to a non-metric hypothesis. *Psychometrika* **29** (1964) 1–27.

R. F. Ling, A probability theory of cluster analysis. *Journal of the American Statistical Association* **68** (1973) 159–164.

B. F. J. Manly, *Randomization, Bootstrap and Monte Carlo Methods in Biology* (2nd Ed.) (Chapman & Hall, London, 1997).

G. McLachlan and D. Peel, *Finite Mixture Models* (Wiley, New York, 2000). Wiley.

G. Nelson and N. Platnick, *Systematics and biogeography: Cladistics and vicariance* (Columbia University Press, New York, 1981).

M. W. Palmer and E. van der Maarel, Variance in species richness, species association, and niche limitation. *Oikos* **73** (1995) 203–213.

B. D. Patterson and W. Atmar, Nested subsets and the structure of insular mammalian faunas and archipelagos. *Biological Journal of the Linnean Society* **28** (1986) 65–82.

B. D. Ripley, *Spatial Statistics* (Wiley, New York, 1981).

S. H. Roxburgh and M. Matsuki, The statistical validation of null models used in spatial association analyses. *Oikos* **85** (1999) 68–78.

E. O. Wiley, Vicariance biogeography. *Annual Review of Ecology and Systematics* **19** (1988) 513–542.

D. H. Wright, B. D. Patterson, G. M. Mikkelson, A. Cutler, W. Atmar, A comparative analysis of nested subset patterns of species composition. *Oecologia* **113** (1998) 1–20.

D. H. Wright and J. H. Reeves, On the meaning and measurement of nestedness of species assemblages. *Oecologia* **92** (1992) 416–428.