The ecodist Package for Dissimilarity-based Analysis of Ecological Data

Sarah C. Goslee
USDA-ARS

Dean L. Urban
Duke University

Abstract

Ecologists are concerned with the relationships between species composition and environmental factors, and with spatial structure within those relationships. A dissimilarity-based framework incorporating space explicitly is an extremely flexible tool for answering these questions. The R package **ecodist** brings together methods for working with dissimilarities, including some not available in other R packages. We present some of the features of **ecodist**, particularly simple and partial Mantel tests, and make recommendations for their effective use. Although the partial Mantel test is often used to account for the effects of space, the assumption of linearity greatly reduces its effectiveness for complex spatial patterns. We introduce a modification of the Mantel correlogram designed to overcome this restriction and allow consideration of complex nonlinear structures. This extension of the method allows the use of partial multivariate correlograms and tests of relationship between variables at different spatial scales. Some of the possibilities are demonstrated using both artificial data and data from an ongoing study of plant community composition in grazinglands of the northeastern United States.

Keywords: autocorrelation, dissimilarity, **ecodist**, Mantel test, R, spatial pattern.

1. Introduction

Ecologists frequently collect data on species composition and environmental variables at a number of sites, and also record the location of each sample. Increasing awareness of the ecological importance of spatial structure within a community or landscape, and of the statistical problems caused by autocorrelated data, has led to interest in spatially-related questions: Is there an underlying spatial structure in the data? If the species composition data are spatially structured, does that structure correspond to spatial structure in the environmental data, or are there spatial factors or processes that have not been measured? In an ecological context, distance apart may be more useful than geographic location. Vegetation composition and

environmental variables can also be expressed in the form of dissimilarity matrices ("distance apart" for sample composition), leading to a consistent analytic framework that will allow us to answer these questions without requiring the data to conform to particular distributions or assumptions.

Urban, Goslee, Pierce, and Lookingbill (2002) described a series of dissimilarity-based analyses intended to address these questions in a organized way. The core methods covered in that paper, including nonmetric multidimensional scaling, hierarchical clustering, and Mantel tests, have been implemented for R (R Development Core Team 2007) in the **ecodist** package so that ecologists have convenient access. We expect that community and landscape ecologists will find these statistical techniques useful, as will any researcher dealing with complex multidimensional or spatially-structured data. Our objectives are to review these analyses of nonlinear and multivariate spatial structure and make recommendations for their effective use, and to present a novel extension to the Mantel correlogram. Specifically, we consider (1) simple Mantel tests, especially tests of significance; (2) variations on partial Mantel tests; and (3) spatial structure in ecological data as revealed in ordinary and partial Mantel correlograms. The methods contained in **ecodist** are illustrated using both artificial data and data from grazinglands of the northeastern United States.

The **ecodist** package for R started as a collection of Mantel test methods, including both simple and partial Mantel tests. To the best of our knowledge, partial Mantel tests with multiple partials are not otherwise available in R. Over the past ten years, it has grown to include additional functions for dissimilarity-based methods. Some of these methods, such as metric and nonmetric multidimensional scaling ordination, are available in base packages or in additional packages such as **vegan** (Oksanen, Kindt, Legendre, and O'Hara 2006) and **ade4** (Chessel, Dufour, and Thioulouse 2004) and have only been included within **ecodist** for convenience, but many of the functions in **ecodist** are not otherwise available in R, or are only available elsewhere in a simpler form.

Dissimilarity-based methods in the form of ordination and cluster analysis have a long tradition in ecology. Most multivariate methods implicitly require a particular dissimilarity metric and make assumptions about the underlying distributions. As a simple example, principal components analysis uses Euclidean distances, and assumes a linear relationship between the data and the underlying gradient. These innate requirements are often not clearly understood. The group of methods derived from the Mantel test allow explict choice of appropriate dissimilarity metrics, and can be used with both quantitative and categorical data (Legendre and Fortin 1989; Smouse and Long 1992; Fortin and Gurevitch 1993; Diniz-Filho and Bini 1996; Luo and Fox 1996; Rossi 1996; Legendre and Legendre 1998; Koenig 1999; Urban *et al.* 2002). These dissimilarity-based methods can be applied to any number of species, from one to a community to an entire landscape, and can incorporate any assortment of ancillary variables.

The methods available in **ecodist** require the user to select a dissimilarity metric. Euclidean distances are generally used for environmental data and spatial coordinates, cases where this direct line representation is appropriate. For vegetation abundance data, where presences are meaningful but absences may not be, a dissimilarity measure such as Bray-Curtis is usually preferred. The similar Jaccard metric is used with presence data. The choice of dissimilarity metric is considered in more detail in Legendre and Legendre (1998). The **ecodist** function `distance()` currently includes five indices: Euclidean, Bray-Curtis, Manhattan, Mahalanobis, and Jaccard (Legendre and Legendre 1998). While this duplicates some of the functionality

of `dist()` in base R, `distance()` was written so as to be easily extensible, with functions implemented in C that perform the basic calculations needed for the majority of dissimilarity metrics: paired difference, paired sum, joint presence and absence, and to count the number of times that an item appears in only the first or only the second sample. These functions can be combined as needed to easily calculate the final metric. A simplified version showing only the Bray-Curtis dissimilarity metric illustrates the approach used

```
distance <- function (x, method = "bray-curtis")
{
    A <- paireddiff(x)
    A <- apply(A, 1:2, function(x) sum(abs(x)))
    B <- pairedsum(x)
    B <- apply(B, 1:2, sum)
    D <- A/B
    D
}
```

A second function, `bcdist()` duplicates the `"bray-curtis"` option of `distance()`, but remains in the package to support legacy code developed by our research group.

2. Simple Mantel test

The simple Mantel statistic is effectively the correlation between two dissimilarity matrices. This is a normalized version of the original Mantel statistic (Mantel 1967). The hypothesis of a Mantel test is that the degree of dissimilarity in one dataset corresponds to the degree of dissimilarity in another independently-derived dataset. For vegetation and environmental data, a simple Mantel test determines whether environmentally similar sites also have similar species compositions, and different sites have dissimilar species compositions.

The simple test is implemented using formula syntax in the `mantel()` function. An example with artificial data will illustrate the general approach. Twenthy samples were located at uniform distances along a transect, and abundances for five species were drawn from a normal distribution. Euclidean distances are used to represent space, and Bray-Curtis dissimilarities are used for the normally-distributed species data. A simple Mantel test examines whether closer-together samples (smaller distance between sample locations) are more similar than farther-apart samples.

```
R> library("ecodist")
R> set.seed(9876)
R> sampleloc <- 1:20
R> species <- matrix(rnorm(100), nrow = 20, ncol = 5)
R> sampleloc.edist <- distance(sampleloc, "euclidean")
R> species.bcdist <- distance(species, "bray-curtis")
R> mantel(species.bcdist ~ sampleloc.edist)

    mantelr        pval1        pval2        pval3   llim.2.5%  ulim.97.5%
 0.01770804   0.45900000   0.54200000   0.84100000  -0.08224247  0.03649103
```

The `mantel()` function returns the Mantel r statistic, and three p values from a randomization procedure described below. For most circumstances, `pval1`, assessing the significance of the null hypothesis that `r <= 0`, is the appropriate choice. For this artificial example, `pval1` provides the desired hypothesis test.The test results are non-significant, leading to the conclusion that closer samples are not more similar. The null hypotheses for the other two tests, that `r >= 0` (`pval2`) and that `r = 0` (`pval3`) are only relevant in special cases, most commonly when model matrices or transformations of the dissimilarities are used. The confidence limits around Mantel r will be discussed later. In the special case where the dissimilarity matrix is actual distance between samples, the simple Mantel test assesses linear spatial structure in species composition or environmental variability. The standard Mantel test only indicates that a linear relationship exists, not the direction of the relationship, which might be indicated by a positive or negative correlation in the original data. A signed Mantel test may provide this additional information when needed (Oberrath and Bohning-Gaese 2001). Signed tests are not currently implemented in **ecodist**, but are under development.

In cases where the distributions of dissimilarity matrices are skewed, ranking the data prior to the Mantel test may help to compensate (Faust and Romney 1985). Simulation studies have shown that the ranked test provides similar or higher power than the unranked Mantel test, although this has not been proven analytically (Dietz 1983). This test is equivalent to the Spearman correlation coefficient. The simple Mantel test is based on a linear model; ranking the data may linearize otherwise nonlinear relationships between dissimilarity matrices (Legendre and Legendre 1998). Monotonicity (correlation of ranks) is a looser requirement than linearity. The ranked version of the previous test returns a slightly higher Mantel r, but is still not significant (as expected for randomly-generated data).

```
R> mantel(species.bcdist ~ sampleloc.edist, mrank = TRUE)

    mantelr       pval1       pval2       pval3   llim.2.5%  ulim.97.5%
 0.03808221  0.33300000  0.66800000  0.64300000 -0.01192652  0.08266478
```

Care must be taken when interpreting Mantel results, since a relationship between dissimilarities does not necessarily correspond to a similar relationship in the original data (Dutilleul, Stockwell, Frigon, and Legendre 2000). For a given sample size, a significant Mantel statistic is much lower than would be expected from familiarity with correlation analysis (Legendre 2000; Urban *et al.* 2002). Despite this, the test on dissimilarity matrices has higher statistical power than the equivalent ANOVA (Somerfield, Clarke, and Olsgard 2002).

Although Mantel (1967) proposed a normal approximation for assessing the significance of the Mantel test, this method is rarely appropriate for ecological data. Instead, permutation testing should be used. Permutation tests are distribution free, but are only effective if the various dissimilarity matrices are independent (Legendre and Legendre 1998; Manly 1997). The permutation test for a simple Mantel test is straightforward: the rows and columns of one matrix are jointly randomized. This simple restricted randomization procedure that preserves the structure among pairs of sampling locations (Fortin and Payette 2002).

The appropriate number of permutations to use for testing significance has been debated in the literature. Jackson and Somers (1989) believe that at least 10,000 permutations should be used, while other researchers feel that 1,000 permutations is sufficient (Manly 1986; Luo and Fox 1996). To evaluate these conflicting claims, we generated pairs of dissimilarity matrices

from data with sample sizes of 10, 25, 50 and 100, and Mantel r statistics of 0.01, 0.05, 0.10, 0.25 and 0.50, using the `corgen` function from **ecodist** with the `epsilon = 0.0001` option to produce data with a specified correlation. The `corgen` function will generate data such that the differences between the desired and actual correlation is no greater than `epsilon`. Smaller values of `epsilon` produce more accurate results, but it can take longer to generate the data. For larger sample sizes, the variance at the higher Mantel r levels was very small, so we omitted these tests, instead adding `r = 0.025` for `n = 100`. For each pair of matrices we calculated p using a range of permutation numbers (500; 1,000; 10,000; 100,000; 1,000,000), and repeated the test of significance 1,000 times to provide an estimate of distribution. Because the calculations are extremely time-consuming, especially for higher-numbers of permutations, the results have been included in the **ecodist** package. All the combinations described above are included in the data, but for the sake of brevity only the `n = 50` results are included here (Figure 1), but simulation results for all sample sizes can be examined in the included `permresults` data. Set up a simulation example for sample size 10 and correlation of 0.01:

```
R> permresults <- data.frame(matrix(0, nrow = 1000, ncol = 17))
R> colnames(permresults) <- c("n", "r", "p500", "l100", "u100", "p1000",
+     "l500", "u500", "p10000", "l1000", "u1000", "p100000", "l10000",
+     "u10000", "p1000000", "l100000", "u100000")
```

Generate correlated data:

```
R> corxy.len010.r010 <- corgen(len = 10, r = 0.01, epsilon = 0.0001)
```

Then record p value, lower and upper confidence limits:

```
R> all.nperm <- c(500, 1000, 10000, 100000, 1000000)
R> all.nboot <- c(100, 500, 1000, 10000, 100000)
R> for(i in 1:1000) {
+     permresults[i, 1] <- 10
+     permresults[i, 2] <- 0.01
+     for(j in 1:5) {
+       thismantel <- mantel(corxy.len010.r010$y ~ corxy.len010.r010$x,
+       nperm = all.nperm[j], nboot = all.nboot[j])
+       permresults[i, (3*j):(3*j + 2)] <- c(thismantel[c(2,5,6)])
+       }
+     }
```

With a sample size of 50 (1,225 dissimilarities), the 0.05 correlation is on the edge of significance at the 95% level. With 500 permutations, the p value from the same pair of dissimilarity matrices ranged from 0.010 to 0.074, and 21% of the tests had a p greater than 0.05. With 10,000 permutations, the range was reduced to 0.035-0.046, and with 1,000,000 permutations the range in p was 0.040-0.041. The tests using a higher Mantel statistic ($r = 0.10$) showed little variation in p with number of permutations. The tests of a weak relationship ($r = 0.01$) showed a wide variation, but always below the usual 0.05 threshold. The tendency is toward less variability at higher sample sizes. A survey of ecological studies using the Mantel test found that most authors either used 1,000 permutations or did not state how many were used.
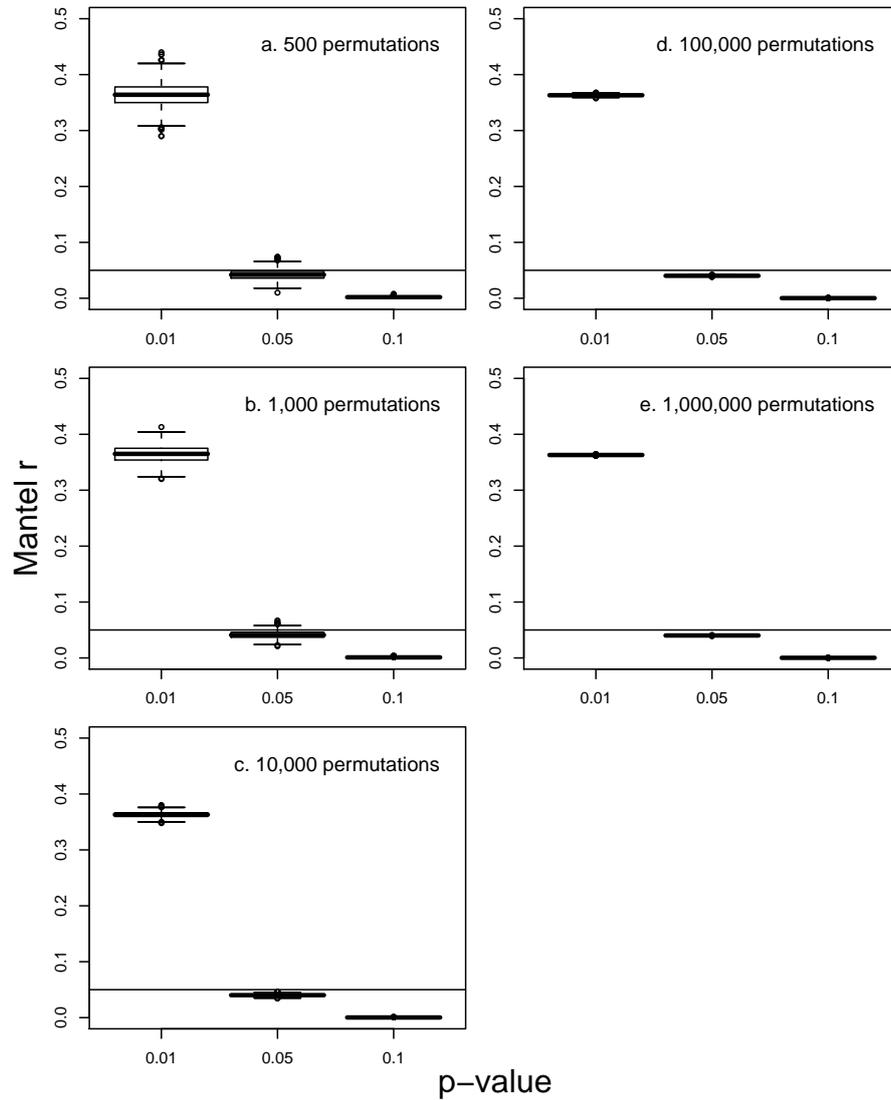
Figure 1: The distribution of p values resulting from repeated permutation testing of data with specified correlation coefficients. Sample size was 50 (1,225 pairs of dissimilarities), and each test of significance was repeated 1,000 times. The horizontal line marks the conventional 0.05 level of significance.

Permutation tests do take a long time with large datasets, so for preliminary analyses, 1,000 permutations is probably adequate. With the speed of computers now available, there is no reason that 10,000 or more permutations cannot routinely be used in analyses intended for publication. The number of permutations used should always be specified in publications. If not explicitly stated, the examples in this paper used 1,000 permutations to reduce the time needed.

We have added a simple modified bootstrapped confidence interval to the Mantel statistic

using the method described by (Manly 1997). Bootstrapped confidence intervals provide a means to compare Mantel correlations computed for different study areas, or for the same study area sampled at different times. The user can specify the number of subsamples, the level of resampling, and the size of the desired confidence interval. Subsamples are taken from the elements of the original dissimilarity matrix, and used to calculate the Mantel statistic. Bootstrapping with replacement is not effective with dissimilarities, since the dissimilarity between a sample and its replicate is by definition zero, so sampling without replacement is used. The process is repeated for the desired number of times, and used to calculate the upper and lower bounds of the confidence interval. Stable confidence limits are provided by using at least 1,000 permutations. Note that the confidence limits give an estimate of population-level variation, while the significance test compares the r value from the given data to that obtained from other orderings of the complete set of samples, so they do not always agree.

3. Partial Mantel test

The simple Mantel test only allows the use of two dissimilarity matrices. Many methods have been proposed for the inclusion of additional matrices. The most commonly used is the partial correlation approach, so that the Mantel statistic is the partial correlation of A and B given all other explanatory variables. The partial Mantel test has frequently been used to examine the relationship between vegetation composition and environmental similarity once geographic distance is taken into account. The desired partials are specified using formula notation. Additional partials can be included by adding additional terms to the formula. Returning to the artificial data generated above, assume that a third variable, perhaps soil type, may also be contributing to patterns in species composition. An ecologist might be interested in the relationship between species composition and sample location once the effects of soil have been removed. In the `mantel()` function, one or more partial variables can be added to the formula separated by + symbols. In the first example below, the test is on species vs. location, given the effects of soil, while the second test is on species vs. soil, given the effects of space.

```
R> soil <- runif(20)
R> soil.edist <- distance(soil, "euclidean")
R> mantel(species.bcdist ~ sampleloc.edist + soil.edist)

    mantelr       pval1       pval2       pval3   llim.2.5%  ulim.97.5%
 0.01836871  0.46800000  0.53300000  0.84600000  -0.07480468  0.03826385


R> mantel(species.bcdist ~ soil.edist + sampleloc.edist)

    mantelr       pval1       pval2       pval3   llim.2.5%  ulim.97.5%
-0.08875399  0.86400000  0.13700000  0.23000000  -0.12215541  -0.04361347
```

The appropriate permutation for a partial Mantel test has been the subject of much discussion (Manly 1986; Smouse, Long, and Sokal 1986; Anderson and Legendre 1999; Legendre 2000). It is important to be aware that all of these methods produce the same value for the Mantel coefficient r, but can affect the corresponding p and the underlying assumptions about data structure. For the partial Mantel test between A and B given C, it is very common to take the

residuals of A on C and of B on C and use these residuals in what becomes effectively a simple Mantel test. The correlation between residuals is equivalent to a partial correlation coefficient, and the permutation test randomizes row and column order within the residuals. Detailed analysis of this residual permutation method has demonstrated that it leads to inflated Type I error, and should never be used (Legendre 2000).

While other statistically correct methods have been proposed, we prefer to randomize the rows and columns of one of the original dissimilarity matrices, while leaving the others unpermuted. This is the same permutation described in the earlier section on simple Mantel tests. For both simple and partial tests, the permutation is applied only to one dissimilarity matrix. For the example above, A would be permuted, and B and C left alone, and the test statistic recalculated. We consider this to be the most conceptually straightforward of the potential methods. Structural relationships between B and C are preserved, and permutation of the original data makes it possible to design tests for very complex structures. This permutation test is statistically acceptable unless extreme outliers are present (Anderson and Legendre 1999; Legendre 2000). Whichever method is chosen, it is important to describe the method used in publications so that others can evaluate the results properly.

Mantel tests are rarely performed in isolation. Instead, many tests are performed using the same data to investigate sets of hypotheses. Multiple comparison procedures are needed to correct the critical p value at which to accept the null hypothesis (Legendre and Legendre 1998). The Bonferroni correction is conservative, but simple to calculate. For a set of n tests, the modified critical p value to reject the null hypothesis is $\frac{\alpha}{n}$ for each test. With a large number of tests, as is usual for Mantel analysis, the critical p quickly becomes impractically small. At minimum, a global test of significance should be used for each set of tests. The Bonferroni test of global significance requires that for n tests, at least one p is less than $\frac{\alpha}{n}$. If this criterion is not met, the entire set of null hypotheses cannot be rejected (Legendre and Legendre 1998). If the criterion is met, then the individual tests within the table can be evaluated.

The response of partial Mantel tests to spatial autocorrelation is a concern (Oden and Sokal 1992; Raufaste and Rousset 2001; Castellano and Balletto 2002). The partial Mantel test is fairly robust to autocorrelation, but the permutation test will fail in extreme examples because not all orderings of the data are equally likely (Oden and Sokal 1992). One simple solution is to choose a conservative critical p value for assessing significance (Oden and Sokal 1992). A more sophisticated answer is to use restricted randomizations, so that all randomizations are not equally likely (Fortin and Payette 2002). The third approach, and the one we advocate, is to identify and remove spatial structure from the data as part of the analysis through the use of partial Mantel tests or non-linear spatial detrending (see below).

The dissimilarity matrix used need not come from data; instead it can be be generated from a research hypothesis. A common use is to test group membership; the model matrix would have 0 dissimilarity for pairs in the same group and 1 for those in a different group (Legendre and Fortin 1989). The Mantel framework provides additional flexibility, since other variables, including space, can be incorporated into the test of group membership. Very complex models can be tested using sophisticated model matrices (Sokal, Oden, Walker, and Waddle 1997). Caution should be used when interpreting results, as differences in group size within a model matrix can bias the results (Luo and Fox 1996). A larger sample will have less variance in the permutation results, all else being equal.

4. Spatial structure

The ability to incorporate geographic distance directly into the analysis is a useful feature of dissimilarity-based analysis. Note that this inferential framework incorporates space as "distance apart" rather than location, which are two very different concepts (Urban *et al.* 2002). From an ecological perspective, relative location - how samples are arranged in space with respect to each other - may well be more meaningful than the absolute geographic location.

Ecologists often use the presence of a significant correlation between compositional or environmental dissimilarity and geographic distance ("space") as an indication of the presence of spatial autocorrelation. Geographic distance is incorporated into partial Mantel tests, where the partial correlation of compositional and environmental dissimilarity given geographic distance is assumed to correct for any spatial autocorrelation present. The assumption of linearity implicit in the Mantel test invalidates this approach for all but the simplest spatial structures. The linear assumption can sometimes be satisfied by using a ranked test, or by transforming the geographic distances (Legendre and Fortin 1989; Legendre and Trousselier 1998). The general hypothesis of the Mantel test, that close samples are similar and far ones different, may not be appropriate for examining spatial structures either; often close samples are similar, and far ones are unrelated (Lefkovitch 1984). Leduc, Drapeau, Bergeron, and Legendre (1992) dealt with this by truncating the geographic distance matrix, and discarding all the far-apart pairs. A more flexible test would eliminate the linearity assumption and make these work-arounds unnecessary.

As an example, we will use a simulated landscape with nine equally-spaced peaks arrayed in a square, and sampled on a 25 x 25 grid as seen in Figure 2. To the eye, there is a definite spatial pattern, perhaps corresponding to a series of mountain peaks, yet the Mantel $r = -0.002$ with $p = 0.576$. The simulated landscape is contained in the `bump` dataset in **ecodist**. A partial Mantel test on space would only remove this nonexistent linear component of the relationship. Mantel analysis incorporating geographic distance would satisfy the researcher that spatial autocorrelation is properly taken care of, when in fact it is being completely neglected because the relationships are strongly nonlinear.

The Mantel correlogram explicitly considers the spatial structure at different scales within a dissimilarity matrix (Oden and Sokal 1986; Legendre and Fortin 1989). To investigate the spatial structure within vegetation, compositional dissimilarity is compared to a series of binary model matrices specifying membership in a particular distance class. Each comparison is tested by permutation, so a multiple comparison correction is needed before interpreting the significance within each distance class. The global Bonferroni test discussed above can be used as a starting point. Legendre and Legendre (1998) also discuss a sequential correction intended for the ordered tests in a correlogram; this approach essentially biases the interpretation in terms of "near" distance classes, which are more likely to be of interest ecologically. Because the permutation test is repeated for each distance class, calculation of the correlogram can be time-consuming for large datasets and high numbers of permutations.

The hypothesis of a Mantel correlogram is that the mean compositional dissimilarity within a distance class differs from the mean of all the other distance classes combined (Dutilleul *et al.* 2000). This is not a particularly interesting hypothesis; a more satisfying alternative would be to look at the dissimilarities within a distance class on its own, without specifying its relationship to other distance classes. If the dissimilarities are first standardized to z-scores,
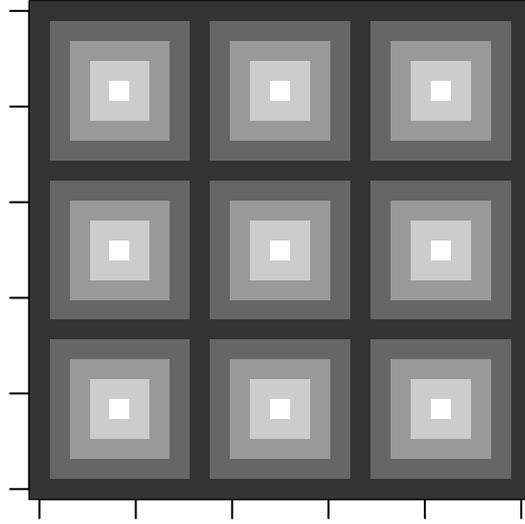
Figure 2: A simulated landscape with obvious but nonlinear spatial pattern.

the appropriate statistic is:

$$r(d) = \frac{-1 \times \sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij} * z_{ij}}{\sum_{i=1}^{n} \sum_{j=1}^{n} w_{ij}} \tag{1}$$

where z_{ij} is the standardized dissimilarity between each pair i and j from the n samples, and w_{ij} is the weight for that pair (typically 1 if it is within the distance class and 0 if it is not). The significance of this simple piecewise autocorrelation statistic is tested by repeatedly permuting the standardized dissimilarity matrix by rows and columns. The piecewise correlogram for the simulated "bump" landscape is shown in Figure 3. The correlogram shows definite non-linear spatial pattern. The value of nearby grid cells are significantly positively related (filled first circle), while cells at 5 distance units apart are significantly negatively related. This pattern of alternating more- and less-similar grid cells continues across the entire range of distances. In practice, the last few distance classes would probably be discarded, since there are few pairs greater than 20 or 25 units apart on this grid.

```
R> X <- col(bump)
R> Y <- row(bump)
R> geo.dist <- distance(cbind(as.vector(X), as.vector(Y)), "euclidean")
R> value.dist <- distance(as.vector(bump), "euclidean")
R> bump.pmgram <- pmgram(value.dist, geo.dist)
R> plot(bump.pmgram)
```

For one variable, using Euclidean distance, this metric converges on the familiar Moran autocorrelation. Like the Moran autocorrelation function, this statistic usually falls between -1 and 1, but is not bounded by those limits. Unlike the Moran function, this correlogram can be used for multivariate data.
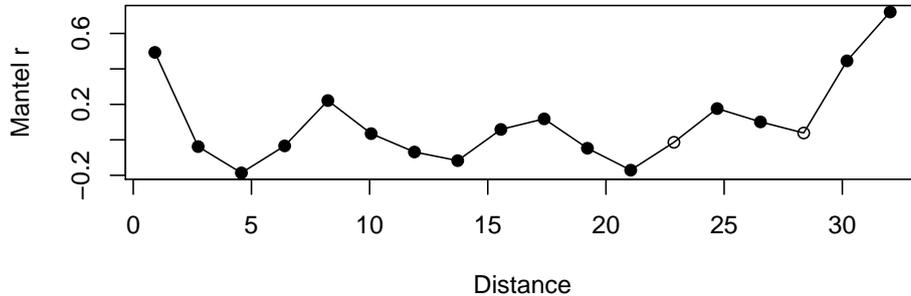
Figure 3: Multivariate correlogram for the simulated landscape shown in Figure 2.

This autocorrelation function tends to resemble the standard Mantel correlogram for one variable despite the difference in hypotheses, but it provides some additional features that the Mantel correlogram does not have. It can be used as the basis for a piecewise removal of the non-linear component of spatial variation. Instead of removing the linear component of spatial variation across the entire range of distances, as done in a partial Mantel test on space, spatial variation is removed separately from each distance class, which approximates the nonlinear structure within the data. If the structure is linear, this process will give the same results, since the sequence of segments will reconstruct that line. Selection of number of classes is the result of a trade-off between resolution and adequate sample size within each class. Some trial and error may be needed to find a set of classes that partitions the data into similarly-sized groups.

The `pgram()` function requires two dissimilarity matrices, one for the data of interest and one for the spatial location with which to compare it. It is also possible to specify an additional dissimilarity matrix for a partial correlations (currently only a single such matrix), the number of classes or step size to use for the lags, whether to return the piecewise residuals, and the number of permutations to employ. It returns a matrix containing the lag distance, the group size, the Mantel statistic and the p value.

Urban *et al.* (2002) discussed the need for a partial Mantel correlogram, which would investigate the relationship between two variables at different spatial scales while controlling for the effects of others. This would make it possible to identify the spatial scale of influence of a particular environmental variable. The multivariate correlogram described here can easily be extended to partial tests, unlike the standard Mantel correlogram. The relationship between compositional dissimilarity and environmental dissimilarity is computed for each distance class. A modified permutation procedure is used, where one dissimilarity matrix is permuted by row and column as usual, but dissimilarities outside the distance class are ignored when calculating the test statistic. This procedure retains the internal structure of the dissimilarity matrices.

5. Illustration: Plant community composition

This example is drawn from a larger study of plant community composition and biodiversity in grazinglands of the northeastern United States (Tracy and Sanderson 2000). A clustered sampling design was chosen for this study, with two to five vegetation samples taken on 12 farms located from New York north to Maine. The vegetation data used here are the mean values of canopy cover estimates from 10 quadrats (1 m^2 each) located within a single pasture. Within these samples, we identified 115 plant species. Dominant species were mostly common forage species, and 31 species were encountered only once. Location as used here is a function of latitude and longitude for each farm since the sites roughly fall along a southwest to northeast line, and is in arbitrary units. The question of interest here is whether there is a relationship between plant community composition and the percentage forest cover within a circle of 1 km radius surrounding the farm (USGS National Land Cover Data, 1992). There is a change in species identity from north to south, so the effects of location must be accounted for when considering any relationship between community composition and other factors. Bray-Curtis distance was used for plant cover data, while Euclidean distance was used for forest cover and for location.

```
R> data("graze")
R> graze.bcdist <- distance(graze[, 3:ncol(graze)], "bray-curtis")
R> sitelocation.edist <- distance(graze$sitelocation, "euclidean")
R> forestpct.edist <- distance(graze$forestpct, "euclidean")
```

Data visualization and plots are the best place to start with a novel dataset, although it can be difficult to visualize a multidimensional dataset. For the purposes of this paper we will begin by using Mantel tests and piecewise correlograms to explore the structure of the data. Simple Mantel tests reveal that there is a significant relationship between composition and forest cover, and between composition and location. There is no relationship between forest cover and location, as demonstrated by the third test below. Since the question of interest is whether a relationship exists between community composition and surrounding forest cover once spatial effects have been accounted for, the usual approach would be to take a partial Mantel test on composition vs forest cover given the effects of location, the fourth test below, which gives a significant result.

```
R> mantel(graze.bcdist ~ forestpct.edist, nperm = 10000)[c("mantelr",
+     "pval1")]

  mantelr      pval1
0.3873425 0.0001000

R> mantel(graze.bcdist ~ sitelocation.edist, nperm = 10000)[c("mantelr",
+     "pval1")]

  mantelr      pval1
0.2169834 0.0032000

R> mantel(forestpct.edist ~ sitelocation.edist, nperm = 10000)[c("mantelr",
+     "pval1")]
```

```
   mantelr       pval1
0.04725453 0.15370000


R> mantel(graze.bcdist ~ forestpct.edist + sitelocation.edist,
+     nperm = 10000)[c("mantelr", "pval1")]


  mantelr       pval1
0.3867244 0.0001000
```

But is this sufficient? These Mantel tests show only the linear components of variation, not more complex spatial patterns. In particular, note that the Mantel r of the vegetation vs forest test is very close to that from the same test with location "removed", suggesting that the inclusion of the partial had no effect. The correlogram used in conjunction with these tests is considerably more informative. The piecewise correlogram for plant community composition (Figure 4a) shows a roughly linear relationship at closer distances, with space unimportant at farther distances (> 4 units). A Mantel test probably gives adequate results in this case. The correlogram for forest cover, however, shows a clear nonlinear pattern with a strong "dip" in the middle of the correlogram (Figure 4b). Note that the Mantel test was nonsignificant because there is little or no linear component to the relationship; using Mantel tests alone would have given a false impression of the structure of the data.

```
R> vegetation.pmgram <- pmgram(graze.bcdist, sitelocation.edist,
+     nperm = 10000, nclass = 6)
R> forest.pmgram <- pmgram(forestpct.edist, sitelocation.edist,
+     nperm = 10000, nclass = 6)
R> par(mfrow = c(2, 1))
R> plot(vegetation.pmgram, main = "a. Plant community composition")
R> plot(forest.pmgram, main = "b. Surrounding forest cover")
```

Frequently, a partial Mantel test on space would be used to remove the effects of spatial structure. The preferred method of calculating the partials is to use regression residuals. This cannot account for non-linear structures; a correlogram calculated on the residuals of forest cover vs location shows the same pattern as in the original data (Figure 5a as compared to Figure 4b).

```
R> forest.resids <- residuals(lm(as.vector(forestpct.edist) ~
+     as.vector(sitelocation.edist)))
R> forest.resids.pmgram <- pmgram(forest.resids, sitelocation.edist,
+     nperm = 10000, nclass = 6)
R> par(mfrow = c(2, 1))
R> plot(forest.resids.pmgram, main = "a. Forest cover: Residuals")
R> forest.presids <- residuals(pmgram(forestpct.edist, sitelocation.edist,
+     resids = TRUE, nclass = 6))
R> forest.presids.pmgram <- pmgram(forest.presids, sitelocation.edist,
+     nperm = 10000, nclass = 6)
R> plot(forest.presids.pmgram, main = "b. Forest cover: Piecewise residuals",
+     ylim = range(forest.resids.pmgram$mgram[, "piecer"]))
```

a. Plant community composition
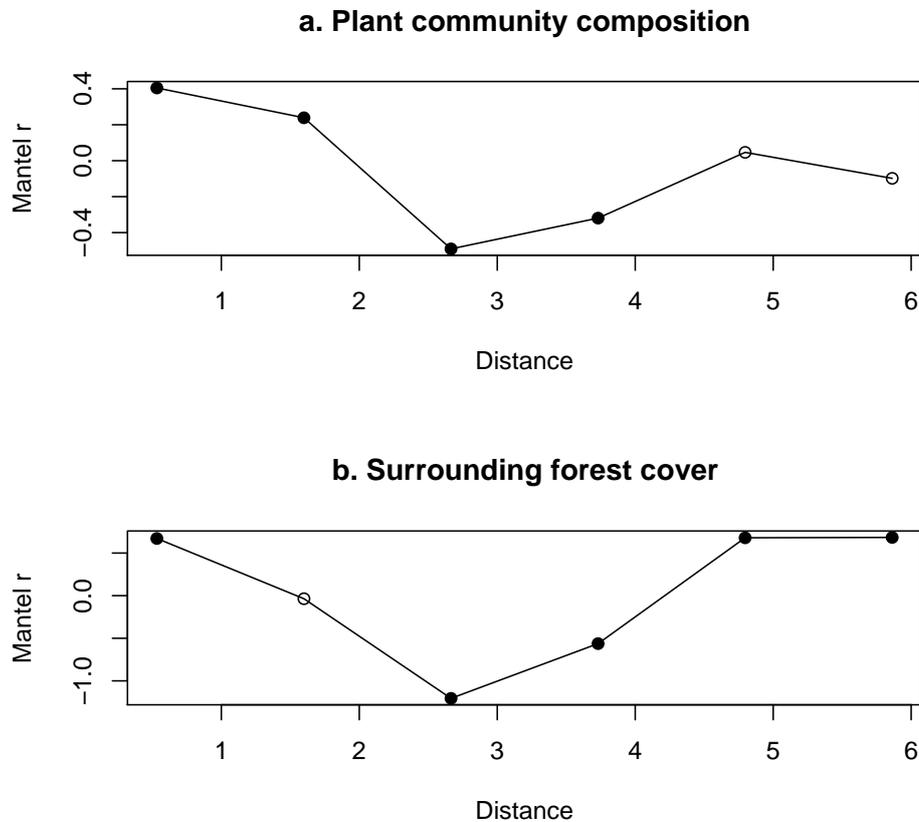

b. Surrounding forest cover

Figure 4: Piecewise Mantel correlograms for plant community composition against location (a) and forest cover in the area surrounding the sample site (b).

Taking the piecewise residuals, however, completely removes the relationship between location and forest cover (Figure 5b). The piecewise partial correlogram, which uses the piecewise residuals, shows that there is a linear relationship between composition and forest cover once the effects of space have been removed, and that the relationship is strongest among sites with very similar and very different forest covers (Figure 6). The corresponding Mantel tests, on piecewise residuals for composition vs space, and for forest vs space, are also significant, as expected since the residual relationship is very nearly linear, but this test does not provide any information about the pattern of relationship.

```
R> combined.pmgram <- pmgram(graze.bcdist, sitelocation.edist,
+     partial = forestpct.edist, nperm = 10000)
R> plot(combined.pmgram, main = "a. Composition vs forest given space",
+     xlab = "Difference in Forest Cover")
R> vegetation.presids <- residuals(pmgram(graze.bcdist, sitelocation.edist,
+     resids = TRUE, nclass = 6))

R> mantel(vegetation.presids ~ forest.presids, nperm = 10000,
```

a. Forest cover: Residuals
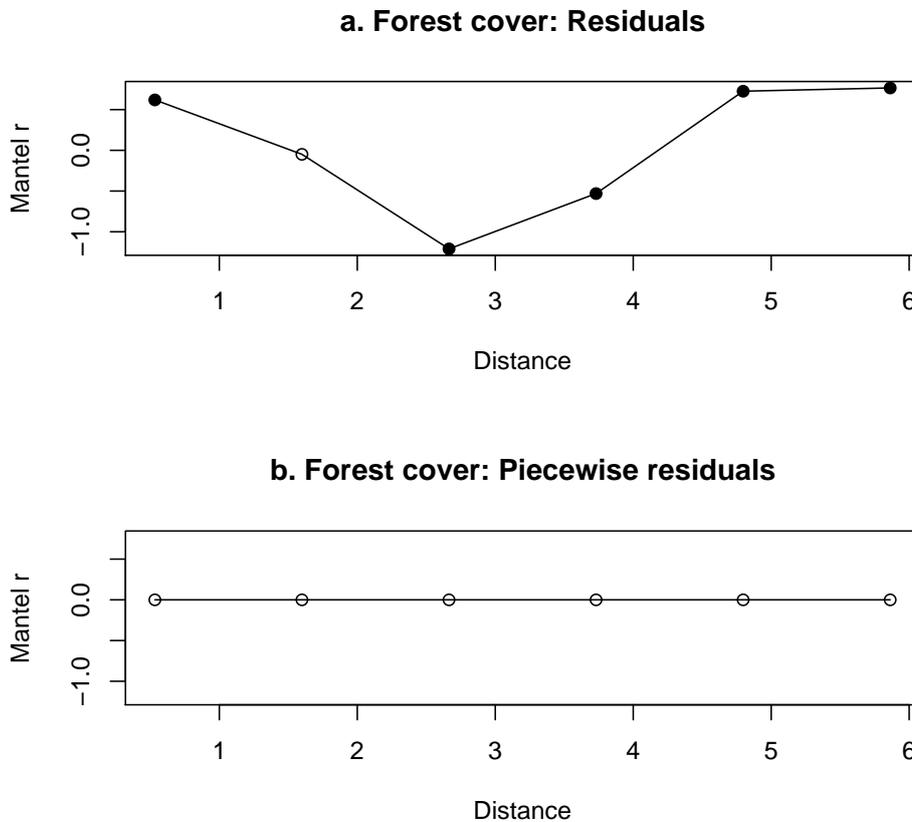

b. Forest cover: Piecewise residuals

Figure 5: Piecewise Mantel correlograms for the linear regression residuals of forest cover against location (a), as in a partial Mantel test; and for piecewise regression residuals (b).

```
+       nboot = 0)[c("mantelr", "pval1")]

  mantelr      pval1
0.2039215 0.0001000
```

In this case, both methods, the conventional and the piecewise, showed a significant relationship between forest cover and plant community composition once spatial effects had been removed, but the magnitude of the relationship was nearly twice as great in the linear analysis than when using nonlinear detrending. The conventional linear analysis would have overestimated the strength of the relationship. The casual assumption of linearity often seen in published Mantel analyses may give rise to misleading conclusions.

6. Conclusions

The **ecodist** package makes a set of dissimilarity-based methods for the analysis of complex ecological data available in R. The package includes standard methods, and also a novel extension that removes the linearity assumption of the Mantel methods. The use of dissimilarity

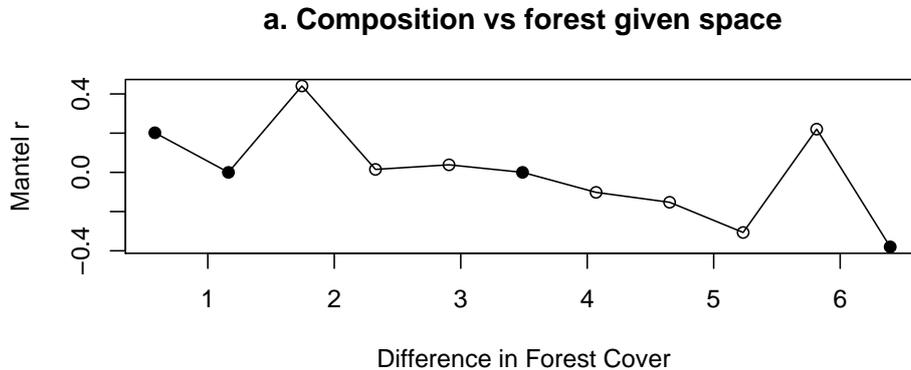**a. Composition vs forest given space**

Figure 6: Partial piecewise Mantel correlogram for plant community composition against forest cover, given space.

methods, which also include standard classification and ordination techniques, allows the explicit inclusion of geographic distance, while requiring only minimal assumptions about the nature of the data. The multivariate correlogram we present here reflects a more appropriate hypothesis than the standard Mantel correlogram. It can be extended to new tests of relationships between variables at different scales, which eliminates the linearity assumption of the usual Mantel tests. The flexibility inherent in this analysis framework requires researchers to consider their choices, including dissimilarity metrics, standardizations and distributional assumptions, and state them explicitly when presenting results. We continue to add new methods as we find them useful, and to simplify the usage of existing methods (where this does not break a decade of accumulated code).

We can now suggest ways to approach the ecological questions presented earlier. First, use the multivariate correlogram to describe the spatial structure of species composition and environmental variables. If all relationships are linear or monotonic, standard Mantel tests can be used. If the relationships are nonlinear, as in the simulated "bump" dataset, piecewise removal of spatial variation can be used for space-free analysis. While the multivariate correlogram method can be used for data with linear relationships, we recommend using the simplest appropriate statistical method so as not to sacrifice statistical power. Whichever method is used, it is necessary to remove the effects of space when examining the strength of relationships between environment and species composition so that spurious effects caused by the effects of space on both can be avoided.

If there is reason to expect nonlinear relationships between species and environment, a piecewise approach can be used again. Finally, a multivariate partial correlogram of compositional dissimilarity can be calculated after removing the effects of all environmental variables so that the spatial scale of unmeasured factors can be identified. This additional information can help to identify these factors or to guide future research. There are many more possible applications of this framework; once these methods become readily available each researcher can apply those most suitable for each system of interest.

References

Anderson MJ, Legendre P (1999). "An Empirical Comparison of Permutation Methods for Tests of Partial Regression Coefficients in a Linear Model." *Journal of Statistical Computer Simulation*, **62**, 271–303.

Castellano S, Balletto E (2002). "Is the Partial Mantel Test Inadequate?" *Evolution*, **56**, 1871–1873.

Chessel D, Dufour AB, Thioulouse J (2004). "The **ade4** Package – I: One-Table Methods." *R News*, **4**(1), 5–10.

Dietz EJ (1983). "Permutation Tests for Association Between Two Distance Matrices." *Systematic Zoology*, **32**, 21–26.

Diniz-Filho JAF, Bini LM (1996). "Assessing the Relationship Between Multivariate Community Structure and Environmental Variables." *Marine Ecology Progress Series*, **143**, 303–306.

Dutilleul P, Stockwell JD, Frigon D, Legendre P (2000). "The Mantel Test versus Pearson's Correlation Analysis: Assessment of the Differences for Biological and Environmental Studies." *Journal of Agricultural, Biological, and Environmental Statistics*, **5**, 131–150.

Faust K, Romney AK (1985). "The Effect of Skewed Distributions on Matrix Permutation Tests." *British Journal of Mathematical and Statistical Psychology*, **38**, 152–160.

Fortin MJ, Gurevitch J (1993). "Mantel Tests: Spatial Structure in Field Experiments." In SM Scheiner, J Gurevitch (eds.), "Design and Analysis of Ecological Experiments," pp. 342–359. Chapman and Hall.

Fortin MJ, Payette S (2002). "How to Test the Significance of the Relation between Spatially Autocorrelated Data at the Landscape Scale: A Case Study Using Fire and Forest Maps." *Ecoscience*, **9**, 213–218.

Jackson DA, Somers KM (1989). "Are Probability Estimates from the Permutation Model of Mantel's Test Stable?" *Canadian Journal of Zoology*, **67**, 766–769.

Koenig WD (1999). "Spatial Autocorrelation of Ecological Phenomena." *Trends in Ecology and Evolution*, **14**, 22–26.

Leduc A, Drapeau P, Bergeron Y, Legendre P (1992). "Study of Spatial Components of Forest Cover Using Partial Mantel Tests and Path Analysis." *Journal of Vegetation Science*, **3**, 69–78.

Lefkovitch LP (1984). "A Nonparametric Method for Comparing Dissimilarity Matrices, a General Method of Biogeographical Distance, and their Application." *American Naturalist*, **123**, 484–499.

Legendre P (2000). "Comparison of Permutation Methods for the Partial Correlation and Partial Mantel Tests." *Journal of Statistical Computation and Simulation*, **67**, 37–73.

Legendre P, Fortin MJ (1989). "Spatial Pattern and Ecological Analysis." *Vegetatio*, **80**, 107–138.

Legendre P, Legendre L (1998). *Numerical Ecology*. Elsevier.

Legendre P, Trousselier M (1998). "Aquatic Heterotrophic Bacteria: Modeling in the Presence of Spatial Autocorrelation." *Limnology and Oceanography*, **33**, 1055–1067.

Luo J, Fox B (1996). "A Review of the Mantel Test in Dietary Studies: Effect of Sample Size and Inequality of Sample Sizes." *Wildlife Research*, **23**, 267–288.

Manly BJF (1986). "Randomization and Regression Methods for Testing for Associations with Geographical, Environmental and Biological Distances Between Populations." *Research in Population Ecology*, **28**, 201–218.

Manly BJF (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman and Hall, second edition.

Mantel N (1967). "The Detection of Disease Clustering and a Generalized Regression Approach." *Cancer Research*, **27**, 209–220.

Oberrath R, Bohning-Gaese K (2001). "The Signed Mantel Test to Cope with Autocorrelation in Comparative Analyses." *Journal of Applied Statistics*, **28**, 725–736.

Oden NL, Sokal RR (1986). "Directional Autocorrelation: An Extension of Spatial Correlograms to Two Dimensions." *Systematic Zoology*, **35**, 608–617.

Oden NL, Sokal RR (1992). "An Investigation of Three-matrix Permutation Tests." *Journal of Classification*, **9**, 275–290.

Oksanen J, Kindt R, Legendre P, O'Hara RB (2006). **vegan**: *Community Ecology Package*. R package version 1.8-3, URL http://CRAN.R-project.org/.

Raufaste N, Rousset F (2001). "Are Partial Mantel Tests Adequate?" *Evolution*, **55**, 1703–1705.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Rossi JP (1996). "Statistical Tool for Soil Biology. XI. Autocorrelogram and Mantel Test." *European Journal of Soil Biology*, **32**, 195–203.

Smouse PA, Long JC (1992). "Matrix Correlation Analysis in Anthropology and Genetics." *Yearbook of Physical Anthropology*, **35**, 187–213.

Smouse PE, Long JC, Sokal RR (1986). "Multiple Regression and Correlation Extensions of the Mantel Test of Matrix Correspondence." *Systematic Zoology*, **35**, 627–632.

Sokal RR, Oden NL, Walker J, Waddle DM (1997). "Using Distance Matrices to Choose Between Competing Theories and an Application to the Origin of Modern Humans." *Journal of Human Evolution*, **32**, 501–522.

Somerfield PJ, Clarke KR, Olsgard F (2002). "A Comparison of the Power of Categorical and Correlational Tests Applied to Community Ecology Data from Gradient Studies." *Journal of Animal Ecology*, **71**, 581–593.

Tracy BF, Sanderson MA (2000). "Patterns of Plant Species Richness in Pasture Lands of the Northeast United States." *Plant Ecology*, **149**, 169–180.

Urban D, Goslee S, Pierce K, Lookingbill T (2002). "Extending Community Ecology to Landscapes." *Ecoscience*, **9**(2), 200–212.

Affiliation:

Sarah C. Goslee
USDA-ARS Pasture Systems and Watershed Management Research Unit
Bldg. 3702 Curtin Rd.
University Park, PA 16802, United States of America
E-mail: Sarah.Goslee@ars.usda.gov