Research

**Ecography 34: 94–102, 2011**
doi: 10.1111/j.1600-0587.2009.05899.x

Subject Editor: Catherine H. Graham. Accepted 23 November 2009

# A method to incorporate the effect of taxonomic uncertainty on multivariate analyses of ecological data

**Luis Cayuela, Marcelino de la Cruz and Kalle Ruokolainen**

*L. Cayuela (lcayuela@urg.es), EcoLab, Depto de Ecología, Centro Andaluz de Medio Ambiente, Univ. de Granada – Junta de Andalucía, Av. del Mediterráneo s/n, ES–18006, Spain. – M. de la Cruz, Depto de Biología Vegetal, E.U.T.I. Agrícola, Univ. Politécnica de Madrid, ES–28040 Madrid, Spain. – K. Ruokolainen, Dept of Biology, Univ. of Turku, FI–20014 Turku, Finland.*

Researchers in ecology commonly use multivariate analyses (e.g. redundancy analysis, canonical correspondence analysis, Mantel correlation, multivariate analysis of variance) to interpret patterns in biological data and relate these patterns to environmental predictors. There has been, however, little recognition of the errors associated with biological data and the influence that these may have on predictions derived from ecological hypotheses. We present a permutational method that assesses the effects of taxonomic uncertainty on the multivariate analyses typically used in the analysis of ecological data. The procedure is based on iterative randomizations that randomly re-assign non identified species in each site to any of the other species found in the remaining sites. After each re-assignment of species identities, the multivariate method at stake is run and a parameter of interest is calculated. Consequently, one can estimate a range of plausible values for the parameter of interest under different scenarios of re-assigned species identities. We demonstrate the use of our approach in the calculation of two parameters with an example involving tropical tree species from western Amazonia: 1) the Mantel correlation between compositional similarity and environmental distances between pairs of sites, and; 2) the variance explained by environmental predictors in redundancy analysis (RDA). We also investigated the effects of increasing taxonomic uncertainty (i.e. number of unidentified species), and the taxonomic resolution at which morphospecies are determined (genus-resolution, family-resolution, or fully undetermined species) on the uncertainty range of these parameters. To achieve this, we performed simulations on a tree dataset from southern Mexico by randomly selecting a portion of the species contained in the dataset and classifying them as unidentified at each level of decreasing taxonomic resolution. An analysis of covariance showed that both taxonomic uncertainty and resolution significantly influence the uncertainty range of the resulting parameters. Increasing taxonomic uncertainty expands our uncertainty of the parameters estimated both in the Mantel test and RDA. The effects of increasing taxonomic resolution, however, are not as evident. The method presented in this study improves the traditional approaches to study compositional change in ecological communities by accounting for some of the uncertainty inherent to biological data. We hope that this approach can be routinely used to estimate any parameter of interest obtained from compositional data tables when faced with taxonomic uncertainty.

Researchers commonly use multivariate analyses to interpret patterns in species data and relate these patterns to environmental predictors. Typically, datasets describing species composition are arranged as a two-dimensional matrix with the samples forming the columns and the species forming the rows. The cells of the matrix can represent the observed abundance of each species, an abundance-score, or presence-absence information. Information on species composition per sample can be used in a variety of ways to define and analyse the species communities. One can investigate the variation of species composition among samples, for instance, through the use of non-parametric multivariate analysis of variance (Anderson 2001), canonical correspondence analysis (CCA, Ter Braak 1986), and redundancy analysis (RDA, Legendre and Legendre 1998). It is also possible to first use the species by samples table to calculate species compositional dissimilarity between each pair of samples and then explain the variation in these measurements by one or several other measures of dissimilarity based on environmental, spatial or temporal variables. Numerical techniques allowing these types of analyses are Mantel test (Mantel 1967), multiple regression on distance matrices (Legendre et al. 1994), and generalized dissimilarity modelling (Ferrier et al. 2007). All these methodological approaches have been applied to different ecological questions, including the mechanisms through which regional biotas are formed (Williams 1996, Moritz et al. 2001, Graham et al. 2006, McKnight et al. 2007), the delineation of biotic regions or biotic transitions (Williams 1996, Williams et al. 1999), or the analysis of distance decay of similarity, i.e. the decrease in compositional similarity with increasing geographic distance between sites (Nekola and White 1999,

Condit et al. 2002, Tuomisto et al. 2003, Qian et al. 2005, Tuomisto and Ruokolainen 2006, Davidar et al. 2007).

In any scientific enterprise, it is important to be able to estimate the uncertainty of the obtained results. Failing to quantify and understand the variation in model predictions due to measurement errors and uncertainty can lead to assumptions that are not valid and ultimately result in erroneous practical decisions (Regan et al. 2002). One recurrent source of uncertainty in explaining variation in species composition or species compositional dissimilarities is the inability to identify a specimen to a scientifically named species. In this case, family or even genus might be identified, but the species identity is indicated with number or some other identifier of a so called morphospecies. The lack of a taxonomic name for the species is not a problem if it is possible to cross-check all the specimens included in the study so that one, and only one, morphospecies name consistently refers to a unique entity that can be interpreted to represent a biological species. However, in some cases the cross-checking of specimens is not practically possible, for example, if the study includes inventories carried out by several different (groups of) investigators who have deposited their specimens in different museums. Even if inventories have been carried out by the same investigators, there can still be situations, especially if the biota is taxonomically poorly known, when it is practically impossible to successfully use the morphospecies approach. This taxonomic uncertainty is common in studies on species-rich and/or little studied systems, such as soils, tropical forests and freshwater invertebrate assemblages (Prance 1994, Brown and Lomolino 1998, Heino and Soininen 2007).

One way of dealing with unidentified species is to relax the taxonomic resolution to the level of genus or family (Terborgh and Andresen 1998, Valencia et al. 1998, Chessman et al. 1999, Kessler and Bach 1999, Pyke et al. 2001, Negi and Gadgil 2002, Slik et al. 2003, ter Steege et al. 2003, Murphy and Davy-Bowker 2005). Identification to genus is generally far easier than identification to species and can simplify considerably the intensive task of field sampling, particularly in tropical forests (Higgins and Ruokolainen 2004), and freshwater ecosystems (Wunsam et al. 2002, Heino and Soininen 2007). Analyses of the same inventory data separately at species level and genus level have often led to rather similar conclusions of the importance of predictor variables on compositional patterns and/or patterns of dissimilarity among the inventory sites (Kessler and Bach 1999, Higgins and Ruokolainen 2004, Heino and Soininen 2007). However, similar results obtained from at the species or genus resolution can not fully justify the relaxation of the taxonomic identification process as a solution for the problem of unidentified species. This is because similar results arise only if two separate forces – species-specific responses to abiotic and biotic factors affecting results at species level, and the degree of evolutionary conservatism in species-specific responses affecting results at genus level – act simultaneously (Wiens and Graham 2005, Losos 2008).

An alternative way of dealing with unidentified species has consisted of trimming off doubtful identifications and morphospecies (Oliveira-Filho and Ratter 1995, Pitman et al. 2001, Linares-Palomino et al. 2003). This has the advantage of ensuring taxonomic uniformity among sites but, as part of the data are discarded, the result obtained will not necessarily represent the relationship between species compositional patterns and environmental or spatial predictors. This risk has been acknowledged and therefore species-level analyses have been avoided when the researchers have felt that taxonomic uncertainty is strong, even if they had actually preferred species-level analyses (Terborgh and Andresen 1998, Pitman et al. 2008). A solution to this problem is to estimate the degree of uncertainty in the results in order to determine if it could affect the conclusions of the analyses. To our knowledge, no previous studies have attempted to do this estimation, or set up theoretical or practical guidelines to quantify uncertainty. The aim of our paper is to take the first steps, both in theory and in practice, to enable justified treatment of the effect of taxonomic uncertainty on multivariate methods commonly used in the analysis of ecological data. To achieve this, we introduce a method that incorporates the effect of taxonomic uncertainty in the estimation of any parameter of interest obtained from multivariate techniques (e.g. Mantel correlation coefficient or explained variance in RDA, CCA or non-parametric multivariate analysis of variance). The method allows an estimation of the potential range of values for each parameter when morphospecies identifications are inconsistent. The procedure is based on iterated randomizations that re-assign unidentified species in each site to any of the potential species found in the remaining sites. If the genus or family of unidentified species is known, re-assignment is done within that level of taxonomic resolution.

We selected two widely used methods in the analysis of ecological communities, Mantel tests and RDA, to illustrate the application of this approach. We used our method to explore the relationship between tree species composition and soil variables in Amazonia. Specifically, we showed how taxonomic uncertainty affected the calculation of two parameters: 1) the Mantel correlation between compositional similarity and environmental distances between pairs of sites, and; 2) the variance explained by environmental predictors in redundancy analysis (RDA). In addition, we performed simulations on a species dataset from southern Mexico to investigate how an increase in the number of undetermined species at different taxonomic resolutions affects the range of the estimated parameter. We choose these two case studies to provide a range of situations where our method can be used. Our hope is that the method outlined in this study will become an easily available tool for applying multivariate methods to ecological communities in cases where there is taxonomic uncertainty.

## The method

The method outlined here describes a general approach to account for taxonomic uncertainty when computing any parameter of interest from biological data tables. This is done by estimating credible bounds under plausible scenarios of re-assigned species identities. We have implemented the procedure in the accompanying package ''betaper'' (freely available at <http://cran.r-project.org/web/packages/betaper/index.html> and in the Supplementary material

Appendix 1), that can be run under the R environment (R Development Core Team 2009).

The input data necessary to implement the procedure are: 1) a community data matrix including the family, genus and species specific names; and usually 2) a matrix of explanatory variables (usually environmental or geographical). Species in the community data matrix are codified according to the taxonomic rank (species, genus, family, etc.) of the most accurate identification. The procedure is then implemented in two sequential steps.

Step 1. Morphospecies identified only to genus are randomly re-assigned with the same probability within the group of species and morphospecies that share the same genus, provided they are not found in the same sites. In the re-assignment of the species identity, the species considered can also receive its own identity. For instance, let's assume we have three floristic inventories. In site A we have *Eugenia* sp1 and *E. nesiotica*. In site B we have *Eugenia nesiotica*, *E. principium* and *E. salamensis*. In site C we have *Eugenia* sp2 and *E. salamensis*. *Eugenia* sp1 can be thus re-identified with equal probability as *E.* sp2, *E. principium*, *E. salamensis* or its own identity (*E.* sp1). In the latter case, this means that we assume that *E.* sp1 is a completely different species, although we do not know its true identity. On the contrary, we cannot re-identify *E.* sp1 as *E. nesiotica* because they were found in the same site, so we are quite certain that *E.* sp1 is different from *E. nesiotica*. The same is applied to species identified only to family and fully unidentified species. Note that when collating inventories are from different researchers, we must rename all unidentified species. This is because two researchers can use the same label, e.g. *E.* sp1, even though this name does not necessarily refer to the same species. For a verification of the biological identity of *E.* sp1 one would need to cross-check the vouchers bearing the same name.

Step 2. Step 1 is iterated $n$ times. As a result, $n$ matrices are obtained, all of which contain the same number of sites but a variable number of species depending on the resulting re-assignment of morphospecies.

Multivariate analyses such as non-parametric multivariate analysis of variance, Mantel test, CCA or RDA, can then be applied to each of these matrices of species per sites, provided that a matrix of explanatory variables is available. Application of any of these analyses to the $n$ matrices computed in the former steps will allow the estimation of $n$ parameters of interest, thus allowing calculation of credible bounds for such parameters.

### Implementation of the procedure in two multivariate analyses

We selected two widely used methods in the analysis of ecological communities, namely Mantel test and RDA, to illustrate our approach. In the Mantel test, one parameter of interest is the Pearson's correlation coefficient, $r$, between the species compositional distance matrix and the geographical or environmental distance matrix. The square of Pearson's $r$ is the percentage of variation in the community dissimilarities explained by geographical or environmental distances. In subsequent applications of this method, the Sørensen's coefficient (Legendre and Legendre 1998) will be used to calculate species compositional distance between sites, although other coefficients, such as Jaccard's index, can also be used (Magurran 1988). Rank-based measures of association, such as Kendall's or Spearman's correlation, can also be specified in the procedure. Computation of Mantel test under different scenarios of species re-assignments was done using function "mantel" in the R package "vegan" (Oksanen et al. 2008).

In RDA, or any other canonical ordination method, the aim is to explain the variance of the community composition matrix by the environmental matrix (Legendre and Legendre 1998). This is calculated like a coefficient of determination ($R^2$) in multiple regression, dividing the sum of all canonical eigenvalues by the total variation in the biological data matrix. Another parameter of interest in RDA is the F statistic, i.e. the ratio of constrained and unconstrained total inertia, each divided by their respective ranks. This F statistic is the basis for tests of significance of canonical ordinations. Computation of RDA under different scenarios of species re-assignments was done using function "rda" in the R package "vegan" (Oksanen et al. 2008).

## Application to a case study from western Amazonia

### Description of the dataset

The data from western Amazonia included tree inventories at nine lowland sites (ca 100–150 m a.s.l.) near Iquitos, Peru. The sites were selected to represent regional variations in geology and were distributed along a soil nutrient gradient ranging from nutrient poor loamy soils to richer clayey soils. Each of the nine inventories consisted of four $20 \times 20$ m plots (0.16 ha total area) distributed along 1.3-km transects. At each site, all woody, free-standing stems of $>2.5$ cm dbh were identified to species or morphospecies (Higgins and Ruokolainen 2004). The full inventories included 3980 individuals from 1188 species or morphospecies, between 284 and 477 genera (263 identified genera and from 21 to 214 unidentified genera), and between 78 and 93 families (77 identified families and from 1 to 16 unidentified families). As explanatory variables in RDA we used base cation concentrations (Ca, K, Mg, Na) and in the Mantel test, following Ruokolainen et al. (2007), we had just one environmental distance data matrix that presented Euclidean distances in the logarithm of the sum of cations at each site.

A large proportion of species were not having a scientific name. Out of the 1188 species and morphospecies identified, 475 had been identified only to genus (40.0%), 112 to family (9.4%), and 15 lacked even family names (1.3%). All specimens were in reality cross-checked and therefore the morphospecies can be taken to represent biological species. However, for the purposes of our study, we decided to regard every inventory as a separate effort without any comparison of the specimens with other inventories. After this assumption, we can use the dataset to illustrate how our method allows one to take taxonomic uncertainty into account in analyses of taxonomically non-harmonised data collected by different researchers.

### Mantel test

We calculated the Mantel correlation between floristic similarities (Sørensen index) and soil cation distances 1000 times simulating a situation in which there had not been any cross-checking of specimens from different plots (Fig. 1). This gave us maximum (−0.694) and minimum (−0.538) values for the correlation, together with the median (−0.616) and the range of 95% of values above and below the median [−0.562, −0.662]. For a dataset in which all the morphospecies have been trimmed off from the analysis, the correlation was slightly above the mean of the simulated cases ($r = -0.646$), but still within the 95% range of estimated values.

### Redundancy analysis (RDA)

Using the same simulations of re-assigned species identities conducted in the previous analysis, we calculated the proportion of the variance of the community composition matrix that was explained by the environmental matrix in RDA 1000 times (Fig. 2). This gave us a maximum of 0.514 and minimum of 0.475. The median was 0.493 and the range for 95% of values above and below the median was 0.479–0.507. After trimming off morphospecies from the analysis, the explained variance results below the minimum obtained through permutations ($R^2 = 0.455$).

## Simulating different scenarios of taxonomic uncertainty

### Setting the scenarios

To test the effect of increasing number of undetermined taxa and changing level of taxonomic resolution on the estimated parameters (Mantel correlation and RDA explained variance), we used a dataset including tree species data obtained from 16 forest fragments in the Highlands of Chiapas, southern Mexico. Over 19 000 trees >10-cm stem diameter were identified to species or morphospecies. The morphospecies identities were cross-checked among the inventory sites and therefore we assume that each morphospecies name corresponds to one distinct biological species. Following this assumption, we established a baseline with which to compare different scenarios of (simulated) taxonomic uncertainty. Overall, there were a total of 231 native tree species, 143 genera, and 72 families. Details of the sampling procedure can be found in Cayuela et al. (2006). A matrix of species abundance per fragment and geographical coordinates of the fragments' centroids are provided in the Supplementary material Appendix 2 and as example datasets in the accompanying package "betaper".

We randomly selected 5% of the species, and classified them as unidentified at each level of decreasing taxonomic resolution: 1) genus-resolution; 2) family-resolution; and 3) fully undetermined species. We repeated the procedure with

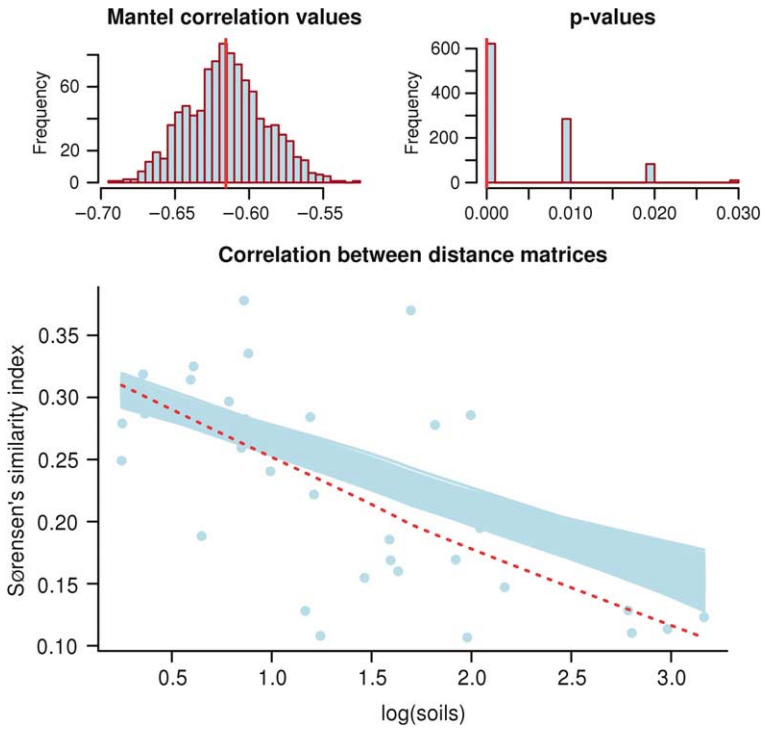


Figure 1. Correlation between Sørensen's similarity index and soil distances among nine inventory plots in western Amazonia (bottom). Soil distances between the 36 possible pairs of plots were calculated using the logarithm of the sum of soil cations. Locally-weighted polynomial regression lines are presented for the 1000 correlations performed between floristic and soil distances (solid blue lines). The red dashed red line represents the locally-weighted polynomial regression line obtained when using the fully identified part of the dataset. Histograms are shown for estimated Mantel correlations (top left) and p-values (top right, p-value in each correlation estimated through 100 randomisations) respectively (solid red lines indicate the median).

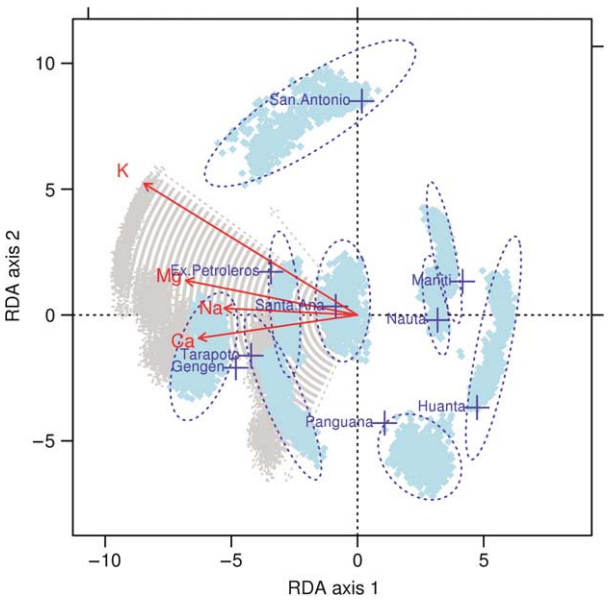


Figure 2. RDA ordination plot of samples (blue points) and linear constraints of soil variables (grey arrows) in the western Amazonia dataset for 1000 simulations of species identities re-assignments. Dispersion ellipses have been drawn depicting 99% confidence levels for the position of each locality. Blue crosses and red arrows indicate the RDA values of samples and linear constraints, respectively, for the fully identified part of the dataset.

10, 15, 20, 25, 30, 35, 40, 45 and 50% of species respectively. In each of these ten different levels of taxonomic uncertainty, 100 simulations were performed following the procedure described above in steps 1–2, and the resulting matrices with different species re-assignments were used as input in Mantel correlation (between floristic similarity matrices based on Sørensen index and Euclidean geographical distance matrices) and RDA (latitude and longitude coordinates as explanatory variables). In each case, results were compared with those resulting from trimming off morphospecies.

We conducted an analysis of covariance to test the effects of the level of taxonomic uncertainty, and taxonomic resolution (morphospecies identified to genus, morphospecies identified to family, fully undetermined species) on the range of observed correlations. To estimate this range we calculated the difference between the minimum and maximum estimated parameter (Mantel correlation and RDA explained variance) in each scenario. These values, together with 95% confidence intervals, are shown in Fig. 3.

### Results

Our results show that there was a significant positive effect of the amount of taxonomic uncertainty on the range of estimated parameters, i.e. potential $r$ values in Mantel test (Table 1) and $R^2$ in RDA (Table 2). The greater was taxonomic uncertainty, the greater was the range of the estimated parameter. The taxonomic resolution at which morphospecies had been identified had a statistically significant effect only on the estimated range of Mantel correlation coefficient, $r$. Here, no differences were found between morphospecies identified to genus and to family (see estimated coefficients in Table 1), but fully undetermined morphospecies clearly increased uncertainty on $r$.

## Discussion

The best way to remove the problem posed by taxonomic uncertainty is to cross-check all the specimens included in the study in order to find every individual belonging to the same morphospecies. Cross-checking of specimens may, however, be overly demanding in terms of human resources and monetary costs. As a consequence, researchers must find a trade-off between data quality and quantity. In such situations, it is the role of statistics to ensure that the available data are properly interpreted, so that the possible effect of taxonomic uncertainty is taken into account. The traditional way to handle the problem of unidentified species is to simply leave them out of any statistical considerations. However, the larger the proportion of unidentified species, the greater the potential risk of erroneous conclusions if they are based solely on identified species. Therefore, a central question is to estimate the size of this risk. We developed a method for estimating the range of values of any parameter obtained from multivariate approaches typically used to investigate the relationship between species composition and environmental predictors under the assumption that the distribution of unidentified species over the sampled communities can follow any pattern which is possible for the lowest identified taxonomical rank – usually genus or family – of the species.

Our results suggest that there are at least two different characteristics that affect the strength of taxonomic uncertainty on the value of the parameter of interest. These are: 1) the number and frequency (and abundance if it is quantified) of unidentified taxa in relation to identified ones in the data table, and 2) the taxonomic resolution of the unidentified taxa (are they identified to genus, family, or some higher taxonomic rank).

It is quite obvious that the larger the proportion of unidentified taxa in a species data table, the more uncertainty there will be about the patterns in species composition and how these patterns relate to environmental predictors. The results of the analysis of covariance performed on the simulations ran on the tree dataset from southern Mexico clearly follow this logic. Yet it was interesting to observe that, despite fairly large amounts of the species may remain unidentified (up to at least 50%), the estimated parameters (Mantel correlation and RDA explained variance) can still stay within a relatively narrow range. This suggests that at least in this particular dataset, the species are ecologically rather redundant. In other words, their distributions are controlled by roughly the same factors so that it does not matter so much which species are taken to the multivariate analysis – the resulting pattern of floristic similarities and differences among the communities remains about the same (Mantel test), as well as the species distribution patterns among the sites (RDA). When thinking about the effect of our randomization procedure on the value(s) of the parameter(s) of interest, it is also important to notice that our method does not only randomize the ecological responses of the unidentified

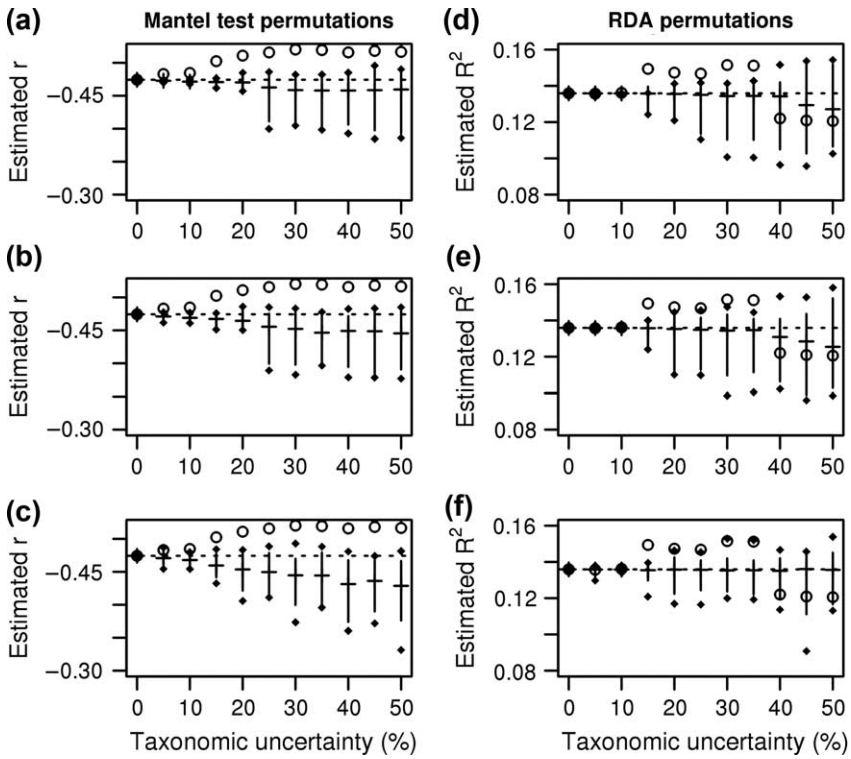


Figure 3. Uncertainty range of the predicted correlation, *r*, between floristic and geographical distances in Mantel tests at increasing levels of taxonomic uncertainty (from 0 to 50%) under different simulated scenarios: (a) undetermined species identified to genus resolution; (b) undetermined species identified to family resolution; (c) fully undetermined species. The same scenarios of taxonomic uncertainty are analysed for the predicted squared R in RDA ((d), (e), (f), respectively). Crosses represent mean values, vertical lines represent 95% confidence intervals, black dots represent minimum and maximum values, and circles represent the estimated parameter after taking out morphospecies from the data matrix.

species, but it also adds noise to the ecological responses of the identified species through the random reassignments of the identifications so that a named species can be recorded to occur in environmental conditions that are in reality outside its limits of ecological tolerance. Therefore, 50% of unidentified species means that much less than 50% of the species in the dataset actually retain their originally observed pattern of distribution in each random run.

The Amazonian data differs from the Mexican data in the sense that unidentified species represented the taxonomically problematic cases and not taxa that were assigned as unidentified by a random draw. In reality, taxonomically unknown species may well be an ecologically non-random selection of species – at least they apparently have smaller ranges of distribution than identified species (Ruokolainen et al. 2002). Therefore, the Amazonian data probably imitate better than the Mexican data a real situation in which our method can be used.

The robustness of the estimated parameters for the two example datasets is well in line with previous independent studies which have considered how subsamples of community data can reveal the pattern of similarities and differences visible in the complete dataset (Kessler and Bach 1999, Vellend et al. 2008). This suggests that biological

Table 1. Analysis of covariance testing the effects of taxonomic uncertainty (covariable), and taxonomic resolution on the range of potential values of the estimated Mantel correlation between floristic and geographical distances. Significant values at $p < 0.01$ are highlighted in bold.

| Analysis of covariance table | | DF | Sum | Sq mean | F value | Pr (>F) |
|---|---|---|---|---|---|---|
| Taxonomic uncertainty | | 1 | 0.013 | 0.013 | 79.860 | **<0.001** |
| Taxonomic resolution | | 2 | 0.007 | 0.004 | 22.099 | **<0.001** |
| Residuals | | 29 | 0.005 | 0.000 | | |
| **Estimated coefficients** | | | **Estimate** | **SE** | **t value** | **Pr (>\|t\|)** |
| Intercept* | | | 0.002 | 0.005 | 0.330 | 0.743 |
| Taxonomic uncertainty | | | 0.001 | 0.000 | 8.936 | **<0.001** |
| Taxonomic resolution: | to genus | | −0.007 | 0.006 | −1.241 | 0.225 |
| | fully unidentified | | 0.028 | 0.005 | 5.036 | **<0.001** |

*The intercept incorporates the effects of one level of the factor taxonomic resolution (to family).

Table 2. Analysis of covariance testing the effects of taxonomic uncertainty (covariable), and taxonomic resolution on the range of potential values of the estimated variability of floristic composition explained by geographical coordinates in RDA. Significant values at $p < 0.01$ are highlighted in bold.

| Analysis of covariance table | DF | Sum | Sq mean | F value | Pr (>F) |
|---|---|---|---|---|---|
| Taxonomic uncertainty | 1 | 0.012 | 0.012 | 249.99 | **<0.001** |
| Taxonomic resolution | 2 | 0.000 | 0.000 | 2.081 | 0.143 |
| Residuals | 29 | | | | |

| Estimated coefficients | | Estimate | SE | t value | Pr (>\|t\|) |
|---|---|---|---|---|---|
| Intercept* | | 0.002 | 0.003 | 0.565 | 0.577 |
| Taxonomic uncertainty | | 0.001 | 0.000 | 15.811 | **<0.001** |
| Taxonomic resolution: | to genus | −0.003 | 0.003 | −1.116 | 0.274 |
| | fully unidentified | −0.006 | 0.003 | −2.037 | 0.051 |

*The intercept incorporates the effects of one level of the factor taxonomic resolution (to family).

communities in general may be characterized by a fairly large amount of ecological redundancy in species' responses to environmental characteristics. It appears conceivable that a redundant community will have a narrow range of values for the estimated parameters obtained through the randomization process. Also, in a redundant system the parameter of interest obtained after trimming off the unidentified species should have a value close to, but more extreme than, the most extreme limit of the range of values produced via randomizations. This is expected because the randomization process is unlikely to improve the statistical relationship between environment and species in a redundant system, but a decrease should occur easily. If the relationship is rather weak and/or individual species behave ecologically quite differently, it would be easier to obtain higher parameter values through the randomization process. Our results, however, cannot provide a rigorous test of this idea and further research with simulated datasets of different degrees of ecological redundancy would therefore be needed to properly address these issues.

The taxonomic resolution at which morphospecies were determined also affected the outcome of the distance matrix correlation in Mantel tests in our simulations with the Mexican data (but not in the outcome of RDA). Morphospecies identified to genus or to family had a smaller influence on the range of estimated correlation values than morphospecies without any identified level of taxonomic hierarchy. This is probably due to the fact that a specimen with genus identification can change the distribution pattern of only relatively small number of congeneric identified species. On the other hand, a specimen without genus or family identification can change the distribution pattern of any other identified species in the data table. Therefore, morphospecies with only a relatively high identified taxonomic rank can potentially have a stronger effect on the range of the correlation.

All the examples used in this article concern plants, a consequence of our field of expertise. Nevertheless, the proposed methodology can be applied to any taxon and ecosystem. For freshwater ecosystems, for example, taxonomic uncertainty is particularly problematic given their high levels of biodiversity compared to their areal extent and the absence of species-level information for many taxonomic groups (Heino and Soininen 2007). Taxonomic uncertainty is also widespread in microbial communities (Mitchell and Meisterfeld 2005, Fraser et al. 2009, Heger et al. 2009). Many bacteria types are identified by genetic polymorphism and, therefore, no one can really know exactly what the locus banding pattern means and what bacteria species they are dealing with (O. Steinitz pers. comm.).

Finally, our approach can be easily transferred to other multivariate methods, such as non-parametric multivariate analysis of variance (Anderson 2001), CCA (Ter Braak 1986), analysis of similarities (Clarke 1993) or generalized dissimilarity modelling (Ferrier et al. 2007). Some of these methods have been already implemented in the accompanying package "betaper". In addition, we must also acknowledge the fact that data are samples and not populations. Therefore, we can presume that the real ranges for the estimated parameters are possibly wider than the ones obtained because of the effect of sampling a population. In fact, one way to see the problem of taxonomical uncertainty is to think what would be the maximum possible effect that the reallocation of species identities of unidentified taxa can have on the value of any parameter of interest, if the parameter value is first calculated on the basis of identified species only. Our method looks for this effect through a random process of reallocating the species identities. However, a random process would probably not find the absolute maximum effect. To find this maximum effect one would first need to define a specific kind of relation between the predictor and response variable and thereafter use this relation as a constraint so that the reallocated species identities would minimally fit to the relation. We leave this approach for eventual further work on the subject. Finally, other kinds of uncertainty could be also considered, such as that derived from periodic revisions of species phylogeny (Isaac et al. 2004). Nonetheless, incorporating this source of uncertainty to multivariate methods still remains a challenge.

*Acknowledgements* – Special thanks to Esteban Alvárez, who inspired this research. We are particularly indebted to Lucía Gálvez, Richard Condit, Ana Justel, Catherine H. Graham and two anonymous reviewers for their insightful comments on previous versions of the manuscript. LC was supported by the European Commission REFORLAN project (INCO contract 2006-032132), the Andalusian Regional Government project GESBOME (P06-RNM-1890), and project BIOTREE (BIOCON08_044) funded by Fundación BBVA. MC was supported

by the projects REMEDINAL (S-0505/AMB-0335) funded by the Comunidad de Madrid, and A/012436/07 funded by Agencia Española de Cooperación Internacional.

## References

Anderson, M. J. 2001. A new method for non-parametric multivariate analysis of variance. – Austral Ecol. 26: 32–46.

Brown, J. H. and Lomolino, M. V. 1998. Biogeography, 2nd ed. – Sinauer.

Cayuela, L. et al. 2006. Fragmentation, disturbance and tree diversity conservation in tropical montane forests. – J. Appl. Ecol. 43: 1172–1181.

Chessman, B. et al. 1999. Predicting diatom communities at the genus level for the rapid biological assessment of rivers. – Freshwater Biol. 41: 317–331.

Clarke, K. R. 1993. Non-parametric multivariate analysis of changes in community structure. – Aust. J. Ecol. 18: 117–143.

Condit, R. et al. 2002. Beta-diversity in tropical forest trees. – Science 295: 666–669.

Davidar, P. et al. 2007. The effect of climatic gradients, topographic variation and species traits on the beta diversity of rain forest trees. – Global Ecol. Biogeogr. 16: 510–518.

Ferrier, S. et al. 2007. Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. – Divers. Distrib. 13: 252–264.

Fraser, C. et al. 2009. The bacterial species challenge: making sense of genetic and ecological diversity. – Science 323: 741–746.

Graham, C. H. et al. 2006. Habitat history improves prediction of biodiversity in rainforest fauna. – Proc. Natl Acad. Sci. USA 103: 632–636.

Heger, T. J. et al. 2009. The curse of taxonomic uncertainty in biogeographical studies of free-living terrestrial protists: a case study of testate amoebae from Amsterdam Island. – J. Biogeogr. 36: 1551–1560.

Heino, J. and Soininen, J. 2007. Are higher taxa adequate surrogates for species-level assemblage patterns and species richness in stream organisms? – Biol. Conserv. 137: 78–89.

Higgins, M. A. and Ruokolainen, K. 2004. Rapid tropical forest inventory: a comparison of techniques based on inventory data from western Amazonia. – Conserv. Biol. 18: 799–811.

Isaac, N. J. B. et al. 2004. Taxonomic inflation: its influence on macroecology and conservation. – Trends Ecol. Evol. 19: 464–469.

Kessler, M. and Bach, K. 1999. Using indicator families for vegetation classification in species-rich Neotropical forest. – Phytocoenologia 29: 485–502.

Legendre, P. and Legendre, L. 1998. Numerical ecology. – Elsevier.

Legendre, P. et al. 1994. Modeling brain evolution from behavior: a permutational regression approach. – Evolution 48: 1487–1499.

Linares-Palomino, R. et al. 2003. The phytogeography of the seasonally dry tropical forests in Equatorial Pacific South America. – Candollea 58: 473–499.

Losos, J. B. 2008. Rejoinder to Wiens (2008): phylogenetic niche conservatism, its occurrence and importance. – Ecol. Lett. 11: 1005–1007.

Magurran, A. E. 1988. Ecological diversity and its measurement. – Princeton Univ. Press.

Mantel, N. 1967. The detection of disease clustering and a generalized regression approach. – Cancer Res. 27: 209–220.

McKnight, M. W. et al. 2007. Putting beta-diversity on the map: broad scale congruence and coincidence in the extremes. – PLoS Biol. 5: e272, doi:10.1371/journal.pbio.0050272.

Mitchell, E. A. D. and Meisterfeld, R. 2005. Taxonomic confusion blurs the debate on cosmopolitanism versus local endemism of free-living protists. – Protist 156: 263–267.

Moritz, C. et al. 2001. Biogeographical concordance and efficiency of taxon indicators for establishing conservation priority in a tropical rainforest biota. – Proc. R. Soc. Lond. B 268: 1875–1881.

Murphy, J. F. and Davy-Bowker, J. 2005. Spatial structure in lotic macroinvertebrate communities in England and Wales: relationships with physicochemical and anthropogenic stress variables. – Hydrobiologia 534: 151–164.

Negi, H. R. and Gadgil, M. 2002. Cross-taxon surrogacy of biodiversity in the Indian Grahwal Himalaya. – Biol. Conserv. 105: 143–155.

Nekola, J. C. and White, P. S. 1999. The distance decay of similarity in biogeography and ecology. – J. Biogeogr. 26: 867–878.

Oksanen, J. et al. 2008. vegan: community ecology package. – R Package ver. 1.15-1, <http://vegan.r-forge.r-project.org/>.

Oliveira-Filho, A. T. and Ratter, J. A. 1995. A study of the origin of central Brazilian forests by the analysis of plant species distribution patterns. – Edinb. J. Bot. 52: 141–194.

Pitman, N. et al. 2001. Dominance and distribution of tree species in upper Amazonian terra firm forests. – Ecology 82: 2101–2117.

Pitman, N. C. A. et al. 2008. Tree community change across 700 km of lowland Amazonian Forest from the Andean foothills to Brazil. – Biotropica 40: 525–535.

Prance, G. T. 1994. A comparison of the efficacy of higher taxa and species numbers in the assessment of biodiversity in the Neotropics. – Phil. Trans. R. Soc. B 345: 89–99.

Pyke, C. R. et al. 2001. Floristic composition across a climatic gradient in a Neotropical lowland forest. – J. Veg. Sci. 12: 553–566.

Qian, H. et al. 2005. Beta diversity of angiosperms in temperate floras of eastern Asia and eastern North America. – Ecol. Lett. 8: 15–22.

R Development Core Team 2009. R: a language and environment for statistical computing. – R Foundation for Statistical Computing, Vienna, Austria, <www.R-project.org>.

Regan, H. M. et al. 2002. A taxonomy and treatment of uncertainty for ecology and conservation biology. – Ecol. Appl. 12: 618–628.

Ruokolainen, K. et al. 2002. Two biases in estimating range sizes of Amazonian plant species. – J. Trop. Ecol. 18: 935–942.

Ruokolainen, K. et al. 2007. Are floristic and edaphic patterns in Amazonian rain forests congruent for trees, pteridophytes and Melastomataceae? – J. Trop. Ecol. 23: 13–25.

Slik, J. W. F. et al. 2003. A floristic analysis of the lowland dipterocarp forests of Borneo. – J. Biogeogr. 30: 1517–1531.

Ter Braak, C. J. F. 1986. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. – Ecology 67: 1167–1179.

Ter Steege, H. et al. 2003. A spatial model of tree alpha-diversity and density for the Amazon region. – Biodivers. Conserv. 12: 2255–2276.

Terborgh, J. and Andresen, E. 1998. The composition of Amazonian forests: patterns at local and regional scales. – J. Trop. Ecol. 14: 645–664.

Tuomisto, H. and Ruokolainen, K. 2006. Analyzing or explaining beta diversity? Understanding the targets of different methods of analysis. – Ecology 87: 2697–2708.

Tuomisto, H. et al. 2003. Dispersal, environment, and floristic variation of western Amazonian forests. – Science 299: 241–244.

Valencia, R. et al. 1998. Diversity and family composition of trees in different regions of Ecuador: a sample of 18 one-ha plots. – In: Dallmeier, F. and Komiskey, J. (eds), Forest biodiversity

in North, Central and South America, and the Caribbean: research and monitoring. Man and Biosphere Series 21. Parthenon Publ. Group, pp. 569–584.
Vellend, M. et al. 2008. Using subsets of species in biodiversity surveys. – J. Appl. Ecol. 45: 161–169.
Wiens, J. J. and Graham, C. H. 2005. Niche conservatism: integrating evolution, ecology, and conservation biology. – Annu. Rev. Ecol. Evol. Syst. 36: 519–539.
Williams, P. H. 1996. Mapping variations in the strength and breadth of biogeographic transition zones using species turnover. – Proc. R. Soc. Lond. B 263: 579–588.

Download the Supplementary material as file E5899 from <www.oikos.ekol.lu.se/appendix>.

Williams, P. H. et al. 1999. Interpreting biogeographical boundaries among Afrotropical birds: spatial patterns in richness gradients and species replacement. – J. Biogeogr. 26: 459–474.
Wunsam, S. et al. 2002. Comparing diatom species, genera and size in biomonitoring: a case study from streams in the Laurentians (Québec, Canada). – Freshwater Biol. 47: 325–474.