Running head: Bias in CWM analysis

Bias in community-weighted mean analysis relating species attributes to sample attributes: justification and remedy

David Zelený

Institute of Ecology and Evolutionary Biology, National Taiwan University, No. 1, Sec. 4, Roosevelt Rd., Taipei 10617, Taiwan

email: zeleny@ntu.edu.tw

 # Words: 8966

 # References: 40

Abstract

One way to analyze the relationship between species attributes (e.g. functional traits) and sample attributes (e.g. environmental variables) via the matrix of species composition is by calculating the community-weighted mean of species attributes (CWM) and relating it to sample attributes by correlation, regression, ANOVA etc. This weighted-mean approach is used in a number of ecological fields (e.g. functional and vegetation ecology, biogeography, hydrobiology or paleolimnology), and represents an alternative to other methods used to relate species and sample attributes via the species composition matrix such as the fourth-corner approach.

The problem with the weighted-mean approach is that in certain cases it yields biased results in terms of both effect size and significance, and this bias is contingent upon the beta diversity of the species composition matrix. The reason is that CWM values calculated from samples of communities sharing some species are not independent from each other. This lack of independence influences the number of effective degrees of freedom, which is usually lower than the actual number of samples, and the difference further increases with decreasing beta diversity of the data set. Discrepancy between the number of effective degrees of freedom and the number of samples in analysis turns into biased effect sizes and an inflated Type I error rate in those cases where the significance of the relationship is tested by standard tests, a problem which is analogous to analysis of two spatially autocorrelated variables. Consequently, reported results of studies using rather homogeneous (although not necessarily small) compositional data sets may be overly optimistic, and results of studies based on data sets differing by their beta diversity are not directly comparable.

Here, I introduce guidelines on how to decide in which situation the bias is actually a problem when interpreting results, recognizing that there are several types of species and sample attributes with different properties and that ecological hypotheses commonly tested by the weighted-mean approach fall into one of three broad categories. I also compare available analytical solutions accounting for the bias (namely modified permutation test and sequential permutation test using the fourth-corner statistic) and suggest rules for their use.

Key Words: degrees of freedom; fourth-corner approach; functional traits; modified permutation test; sequential permutation test; species indicator values

Introduction

Weighted-mean approach is a method to analyze the relationship between species attributes and sample attributes by calculating community-weighted means of species attributes (CWM), which can be directly related to sample attributes by correlation, regression, ANOVA or other methods. *Species attributes* are species properties (traits), behavior (species ecological optima) or phylogenetic age, while *sample attributes* are characteristics of community samples measured in the field (environmental variables) or derived from a matrix of species composition (like species richness or positions of samples in ordination diagrams).

The weighted-mean approach is used in a wide range of ecological fields. In functional ecology, testing the effect of environmental variables on changes in CWM is one of the approaches that demonstrates the effect of environmental filtering on trait-mediated community assembly (Díaz et al. 1998; Shipley 2010). Similarly, CWM is used to predict changes in ecosystem properties, such as biomass production or nutrient cycling (Garnier et al. 2004; Vile et al. 2006), or ecosystem services like fodder production or maintenance of soil fertility (Díaz et al. 2007). In biogeography, grid-based means of species properties (such as animal body size, range size or evolutionary age) are linked to macroclimate or diversity (Hawkins and Diniz-Filho 2006). Vegetation ecologists use species indicator values (e.g. those of Ellenberg et al. 1992) to estimate habitat conditions from calculated mean species indicator values of vegetation samples and relate them to soil, light or climatic variables (Schaffers and Sýkora 2000). In hydrobiology, reliability of the saprobic index of Sládeček (1973) based on weighted mean of diatom indicator values, or similar indices (e.g. trophic diatom index, Kelly and Whitton 1995) is evaluated by relating them to measured water quality parameters. Similarly, in paleoecology the method used to reconstruct acidification of lakes from fossil diatom assemblages preserved in lake sediments is based on weighted means of diatom optima along the pH gradient (ter Braak and Barendregt 1986), and as one of the transfer functions (e.g. Birks et al. 1990) is considered to be a tool which has "revolutionised paleolimnology" (Juggins 2013). Other, more specific examples include relating the community specialization index (mean of species specialization values weighted by their dominance in the community) to environmental variables (Clavero and Brotons 2010, Carboni et al. 2016), or attempts to verify whether plant biomass can be estimated from tabulated plant heights and species composition as the mean of species heights weighted by their cover in a plot (Axmanová et al. 2012).

Although the weighted-mean approach technically relates two sets of variables (CWM and sample attributes), three matrices are in fact involved in the computation background (notation here follows the RLQ analysis of Dolédec et al. 1996): matrix of *sample attributes* **R** with m sample attributes of n samples ($n \times m$); matrix of *species composition* **L** with abundance (or presence-absence) of p species in n samples ($n \times p$); and matrix of *species attributes* **Q** with s species attributes for p species ($s \times p$). The weighted-mean approach is just one of the possible options for relating species attributes (**Q**) to sample attributes (**R**) via a matrix of species composition (**L**): it combines **Q** with **L** into a matrix of weighted-means **M** and relates it to **R**. An alternative solution, although rarely used, is to combine a matrix of sample attributes **R** with species composition **L** by calculating the weighted-mean of sample attributes (optima of individual species along a given sample attribute or species centroids) and relate these values to species attributes **Q** (e.g. ter Braak and Looman 1986). A third option is to use methods suitable for simultaneously handling all three matrices (**R**, **L** and **Q**), such as the *fourth-corner approach*

(Legendre et al. 1997), the related ordination method, called RLQ analysis (Dolédec et al. 1996), and other alternatives (Jamil et al. 2013, Brown et al. 2014).

In the weighted-mean approach, the relationship between CWM and sample attributes, analyzed by correlation/regression/ANOVA, is often tested by a standard parametric or permutation test (called simply *standard test* throughout this study). However, not all types of ecological questions, which are usually solved by the weighted-mean approach, should actually be tested by standard test. In certain situations and types of null hypotheses, the weighted-mean approach combined with standard tests generates biased results, which are more optimistic than would be actually warranted by analyzed data. This bias includes unreliable estimates of effect size (e.g. correlation coefficients in the case of correlation, or r^2 in the case of linear regression) and an inflated Type I error rate, leading to more frequent rejection of the null hypothesis than would be expected. The key point before applying the weighted-mean approach is to explicitly decide what is actually the relationship between species attributes or sample attributes and species composition, and which of these relationships is actually fixed and which is random (more on the terms “fixed” and “random” below). This decision should be based on critical inspection of the context of the study question and tested null hypothesis. Inspiration for this issue can be seen in the application of the fourth-corner approach (Legendre et al. 1997), for which Dray and Legendre (2008) demonstrated the problem of deciding on the right permutation test (from five permutation models) to test the actual question in hand, with a risk of inflated Type I error rate in the case of a wrong choice. For the weighted-mean approach, this issue was highlighted by Zelený and Schaffers (2012) in a specific context of relating mean Ellenberg indicator values (species attributes) to sample attributes derived from species composition matrix (like ordination scores or species richness), and by Peres-Neto et al. (2012, 2016) in the context of metacommunity phylogenetics and species functional traits. These studies also proposed numerical solutions: Zelený and Schaffers (2012) introduced a *modified permutation test*, an alternative to the standard permutation test between CWM and sample attributes in which species rather than sample attributes are permuted, and Peres-Neto et al. (2012, 2016) suggested employing the *sequential test* (ter Braak et al. 2012), using the *fourth-corner* statistic (Legendre et al. 1997). Additionally, Šmilauer and Lepš (2014, p. 158) mentioned this issue in the context of the CWM-RDA method (Kleyer et al. 2012).

Here, I first define several types of species or sample attributes, differing by their origin and relationship to a matrix of species composition. Then I review categories of questions and null hypotheses that are commonly analyzed by the weighted-mean approach. I use simulated data to show for which of these categories there is a risk of biased results if tested by standard test and how this bias depends on the beta diversity of a compositional data set. I argue that the bias is caused by a mismatch between the number of samples in the weighted-mean analysis and the actual number of effective degrees of freedom, since community samples sharing some of the species with other samples do not count for the full degree of freedom in this analysis. Note that for numerical simplicity, I ignore intraspecific variation in species attributes. Finally, I review and compare methods available for solving the problem of inflated Type I error rate in the weighted-mean approach, namely the *modified permutation test* (Zelený and Schaffers 2012) and the *sequential permutation test* based on the *fourth-corner statistic* (Peres-Neto et al. 2012, 2016), and suggest guidelines for their use. Although the examples, ecological interpretations and reasoning used here are focused on the relationship of species functional traits or species indicator values with sample attributes analyzed by the weighted-mean approach, the general context is also valid for other types of species and sample attributes linked by the weighted-mean approach.

Types of species and sample attributes

When considering alternative types of questions commonly analyzed by the weighted-mean approach, it proves useful to distinguish whether species and sample attributes are *fixed* or *random*, and *internal* or *external*. The distinction between *fixed* and *random* attributes depends on whether they are specific for a given data set and this specificity is acknowledged by the question/hypothesis being tested (this link is deemed as given, and not further questioned or tested). Fixed attributes are specific and acknowledged, while random attributes represent a subset of some larger pool of values and their link to species composition is not acknowledged by the hypothesis being tested. In the narrow sense of permutation tests, fixed attributes should not be permuted among each other, while random attributes can be. For interpretation, the effect of fixed attributes is limited only to a given set of attribute values and in the context of community data sets included in the analysis and cannot be generalized beyond, while the effect of random attributes can be interpreted more broadly and also beyond the data set used in the study.

The main difference between *internal* and *external* attributes is their origin. *Internal* attributes are numerically derived from a matrix of species composition, while *external* attributes are typically measured or estimated variables, not directly derived from a species composition matrix. *Internal species* attributes are, for example, species optima calculated as the weighted-means of sample attributes or as species scores on ordination axes, and similarly *internal sample* attributes are sample scores on ordination axes, species richness of individual samples or the assignment of samples into groups based on compositional similarity (e.g. by numerical classification). *External species* attributes, on the other hand, are measured traits or tabulated species indicator values, and *external sample* attributes are measured or estimated environmental variables or experimental treatments. While the link of external species or sample attributes to species composition may be fixed or random and depends on the context, internal attributes are always fixed, since they refer only to the context of the data set from which they have been derived and their randomization would make no sense.

To give a few examples: species traits measured on individuals from plots of given community data sets can be considered as fixed and external, while species traits taken from large trait databases and measured often in a completely different context should be considered as random and external. Sample ordination scores derived from a matrix of species composition are fixed and internal sample attributes, while environmental variables measured in the field or derived from GIS layers may be considered as random and external. Indeed, the distinction between *fixed* and *random* is often arbitrary and depends on the authors' decision and the theoretical context of the study, and the same variables can be seen as fixed or random in different contexts. For example, if results are expected to have local validity (e.g. whether the CWM of species height in a given agricultural system can predict the harvested biomass), species attributes can be seen as fixed; if the species height will be measured again in the same community, results will be similar, but not generally applicable to other communities. If the aim is to generalize results (e.g. to assess whether the species height itself, as tabulated in the national floras, can be used as a tool to predict biomass yield), species attributes should be treated as random and the analysis should be modified accordingly, so that even a local study can contribute to a more general description of this relationship.

In the original description of the fourth-corner problem (Legendre et al. 1997), both species and sample attributes were considered as fixed, while the matrix of species composition was considered as random, and different permutation models were applied to test alternative hypotheses. In the weighted-mean approach, the decision as to whether attributes are fixed or random also influences the choice of a meaningful way to test the relationship, and is therefore crucial in the selection of the correct statistical test. All hypotheses (as discussed further) make an implicit or explicit assumption that either species or sample attributes are fixed, with a link to species composition acknowledged *a priori* and not further questioned (and also not tested).

Types of hypotheses tested by weighted-mean approach

Considering the distinction between fixed and random (sample or species) attributes, questions and hypotheses commonly tested by the weighted-mean approach fall into one of the three categories (see Table 1 for summary). *Category A* assumes that while sample attributes are fixed, species attributes are random; *category B* is opposite to the previous category, with sample attributes considered random and species attributes fixed; and, finally, *category C* assumes that both species and sample attributes are random. Below, I review in detail individual categories, with examples of ecological questions/hypotheses for each of them.

Category A – species attributes are random, sample attributes are fixed

Hypotheses in this category explicitly acknowledge the link between sample attributes and species composition, or the link is implicit from the context or numerical background of the study. The null hypothesis states that species attributes are not linked to species composition, while the alternative hypothesis states that they are. Questions focused on relating CWM to internal sample attributes (i.e. those derived numerically from the matrix of species composition) fall into this category (e.g. relating mean Ellenberg indicator values to sample scores in unconstrained ordination to interpret the ecological meaning of ordination axes; Zelený and Schaffers 2012). In addition, studies with external sample attributes considered to be fixed, such as experimental treatments, fall into this category in the case when their effect on species composition is acknowledged, and the question is about how species attributes respond to it. An additional level of complexity is added in studies dealing with grid data where both CWM and internal sample attributes (e.g. species richness derived from community data) are spatially autocorrelated due to the spatial coherence of species distribution (B. Hawkins, *pers. comm.*). Zelený and Schaffers (2012) showed that standard tests have inflated the Type I error rate for this category of hypotheses, and as an alternative introduced the *modified permutation test*, permuting species attributes instead of sample attributes, as further discussed in this study.

Category B – species attributes are fixed, sample attributes are random

Hypotheses in the second category explicitly assume that the species attributes are linked to species composition. The null hypothesis states that sample attributes are not linked to species composition, while the alternative hypothesis states that they are. Examples are trait-based studies asking whether species traits can explain the effect of environmental filtering on species abundance in a community. These studies operate with an assumption that species traits (as species attributes) are functional, i.e. they influence the abundance of species in a community, and the question being evaluated is whether the sample attributes (environmental factor) act as an environmental filter on species abundance. Descriptive studies without ambitions to be more generalized also fall into this category – e.g. the relationship between the CWM of species indicator values (e.g. mean Ellenberg indicator values) and measured environmental variables, if the interpretation is restricted only for the community data set included in the study. Finally, studies using internal species attributes (derived from species composition, e.g. as the weighted-mean of sample attributes or as scores on ordination axes) also belong to this category.

Category C – both species and sample attributes are random

This category of hypotheses includes mostly observational studies without prior knowledge or expectations about a link between any of the matrices. The null hypothesis states that there is no link between species and sample attributes via the matrix of species composition because either the species attributes or the sample attributes (or both) are not linked to the matrix of species composition. To reject this null hypothesis means to prove that both species and sample attributes are actually linked to species composition. Empirical studies describing the general relationship between sample attributes and species attributes, without explicitly or implicitly acknowledging some underlying assumptions or mechanisms, belong to this category. Examples are studies relating the CWM of functional traits to environmental variables without a clear assumption that traits are functional, allowing to question whether particular traits are actually linked to species composition or not. In the case of studies with species indicator values, these include relating mean indicator values to environmental variables with the aim of generalizing the result also beyond the scope of the studied community data set (e.g. answering the question of whether Ellenberg indicator values for soil reaction *per se* are good predictors of measured soil pH, i.e. not only in the context of a given community data set).

Illustration of the bias and its dependence on beta diversity

If hypotheses from categories A and C are tested by the weighted-mean approach based on correlation or regression and combined with the standard test, results may be highly biased, both in terms of the estimated model parameters and the inflated Type I error rate (Zelený and Schaffers 2012, Peres-Neto et al. 2016), and the magnitude of the bias changes with the beta diversity of the compositional data set. I will illustrate this bias using simulated community data, in which the set of communities with increasing beta diversity will be generated and accompanied by matrices of species and sample attributes related (or not) to species composition (creating four scenarios relevant to hypotheses in categories A, B and C). The same simulated data set will be later used to demonstrate the performance of available statistical solutions.

Description of 2D simulated community data set

Each simulated community data set includes the set of three matrices (sample attributes **R**, species composition **L,** and species attributes **Q**), with the link between species or sample attributes and species composition (or both) broken by the permutation of attributes. This creates four scenarios (Fig. 1, identical with scenarios 1–4 of Dray and Legendre 2008): 1) both sample and species attributes linked to species composition; 2) sample attributes linked to species composition, species attributes not; 3) species attributes linked to species composition, sample attributes not; 4) none of species or sample attributes linked to species composition. For hypotheses in category A defined above, scenario 2 represents the null hypothesis, for category B scenario 3 is the null hypothesis, and for category C the scenarios 2, 3 and 4 represent alternative states of null hypothesis (Table 1). Scenario 1 represents the power test for all three categories (i.e. it measures the probability of getting significant results if the alternative hypothesis is true).

Additionally, I also examined how observed bias depends on the beta diversity of the species composition matrix, which influences the number of effective degrees of freedom in analysis (as explained in detail in the section *Justification of the bias*). An algorithm generating community data is structured by two virtual ecological gradients, and will be called *2D simulated community data set* throughout this paper (this is an extension of the original one-gradient algorithm of Fridley et al. 2007). The first gradient has *constant* length for all generated data sets and serves as a surrogate for the measured environmental variable; in the analysis, positions of samples along this gradient are used as *sample attributes*, while the simulated species optima along this gradient are used as *species attributes*. The length of the second gradient is *variable*, and increasing its length increased the beta diversity of the data set (Appendix S1: Table S1 and Fig. S1). The length of the first gradient was arbitrarily set to 1000 units and the range of species niche widths was between 500 and 1000 units. The length of the second gradient varied between 1000 to 10 000 units; for simplicity, here I assume that 1000 units of the second gradient represents one community, i.e. enlarging the second gradient from 1000 to 10 000 units (by steps of 1000 units) generates a set of data sets with 1 to 10 communities. Community samples were created by randomly choosing locations along the first and second gradient, and the species composition for each sample was derived by the random assignment of a fixed number of individuals to species identities weighted by the relative abundance of species with non-zero probability of occurrence at a given location of the gradient (see Appendix S1 for further details).

Note that the model generating the 2D simulated community data is different from the one generating the simulated data sets used by Dray and Legendre (2008) and Peres-Neto et al. (2016), which used only one environmental gradient and generated rather homogeneous communities (Appendix S1: Table S1 vs Appendix S4: Table S2). The other difference is how each algorithm achieves the increase in beta diversity: while in the 2D simulated data set this is done by prolonging the second virtual gradient (which increases gamma diversity while keeping the mean alpha diversity rather constant), in the 1D simulated data set of Dray and Legendre (2008) the beta diversity is increased by narrowing the niche breadth of individual species (keeping the gamma diversity of the data set constant but decreasing the mean alpha diversity). For comparison with other published studies, all analyses were also repeated with the 1D simulated community data generated according to Dray and Legendre (2008), with results available in Appendix S4.

All analyses were conducted using R-project (v. 3.3.1, R Core Team 2015); complete R scripts are available in Data S1 and all functions are in R-packages *weimea* (abbreviation for *wei*ghted *mea*n; source code of v. 0.60 in Data S2).

Weighted-mean approach with standard test applied on simulated data

For each of the four scenarios (1–4) I created ten levels of beta diversity, and for each combination of *scenario* × *level of beta diversity* I created 1000 datasets (4 scenarios × 10 levels of beta diversity × 1000 replications = 40 000 data sets). For each data set I calculated the CWM of species attributes, related it to sample attributes using Pearson's *r* correlation and tested its significance using the parametric *t*-test (for additional results for least-square regression and r^2 see Appendix S2: Fig. S1). For each level of community beta diversity in each scenario, I counted the proportion of correlations significant at $\alpha = 0.05$ (note that this proportion is identical to the proportion of significant regressions).

From the three scenarios with no direct link between species and sample attributes (scenarios 2, 3 and 4), analysis of data generated by scenario 2 reveals the bias – the correlation coefficient deviates from zero more than in other cases (Fig. 2), and the test of significance shows an inflated Type I error rate (Fig. 3). This bias decreases with increasing beta diversity of the species composition matrix (Fig. 2 & 3, Scenario 2): for the most homogeneous data set (*level of beta diversity* = 1), the range of Pearson's *r* correlation coefficients (expressed as 2.5% and 97.5% quantiles) is between -0.751 and 0.751, with 60% of correlations significant, while for the most heterogeneous data set with a high beta diversity (*level of beta diversity* = 10) the range of Pearson's *r* values is between -0.381 and 0.354, with 15% of correlations significant (compared to 2.5 and 97.5% quantile range values of *r* observed in scenarios 3 and 4 being on average between -0.278 and 0.281, with the expected number of significant results being close to 5%). Similarly inflated are the values of coefficient of determination (r^2; Appendix S2: Fig. S1, Scenario 2) calculated by least-square linear regression. Applying the standard test on the simulated community data set of Dray and Legendre (2008) shows analogously biased results (Appendix S4: Table S1 and Fig. S2).

Justification of the bias

Since the CWM of species attributes are calculated from species attributes assigned to individual species and from species composition of individual samples, they inherit some information from both sources. The numerical difference between the calculated CWM values of two community samples is necessarily constrained by a difference in these samples' species composition: two samples with identical species composition (or identical relative species abundances) have identical calculated weighted-means, and two samples with slightly different species composition will have CWM values rather similar. Non-independence of CWM values has consequences for the analysis with sample attributes, if these are themselves in some way related to species composition. Two values of the CWM calculated from community samples sharing some species do not bring two independent degrees of freedom into analysis, because samples used for their calculation are not independent and the difference in their CWM is predictable (to some extent) from the difference in their species composition. This problem scales up to the data set level: in case of two compositional data sets with the same number of samples used in the weighted-mean approach, the data set that is compositionally more homogeneous has a lower number of effective degrees of freedom compared to the more heterogeneous one.

If CWM values are calculated from species composition data in which some samples share some species, and at the same time sample attributes are (in some way) related to species composition, analysis of the CWM with sample attributes resembles the analysis of two spatially autocorrelated variables. Samples of spatially autocorrelated variables located nearby in geographical space have more similar values than expected if the values are randomly selected and are therefore not statistically independent (Legendre and Legendre 2012). A new observation does not bring completely new information, because its value can be partly derived from the value observed in a nearby site, and the effective number of samples (i.e. the effective number of degrees of freedom) is lower than the real number of samples. Since for standard parametric tests the number of degrees of freedom is important for choosing the correct statistical distribution for a given sample size, disparity between the real number and effective number of samples leads to the selection of narrower confidence intervals and hence a higher probability of obtaining significant results (Bivand 1980, Legendre 1993).

In the case of the weighted-mean approach, it is not the proximity in a geographical space, but the proximity in a compositional space, which reflects distances between samples expressed as their compositional dissimilarity. The bias is not present if one or both of the variables (CWM and sample attributes) are not autocorrelated in the compositional space. This happens if either sample attributes are not related to the species composition matrix, or in the improbable case of a species composition matrix having so high a beta diversity that individual samples do not share any species and calculated CWM values are therefore not related to species composition. In case of spatially autocorrelated variables this is analogous to the situation where only one or none of the variable are spatially autocorrelated, in which case the bias caused by autocorrelation does not appear.

The bias in the community-weighted mean approach is therefore limited to cases where species composition of samples is at least partly overlapping and sample attributes are linked to species composition (i.e. they are fixed). This is true for all *internal* sample attributes derived from the matrix of species composition, since they are linked to the matrix of species composition due to their numerical origin, and also for some of *external* sample attributes, if these are considered to be fixed (for examples see section *Types of species and sample attributes*).

The dissimilarity in species composition of two samples directly related to the difference in their CWM values can be quantified by Whittaker's index of association (Whittaker 1952, in Legendre and Legendre 2012 as D_9), which can be numerically derived from differences between two calculated CWM values (see proof in Appendix S3). The single value of beta diversity for a given data set can be obtained using the beta diversity metric of Legendre and De Cáceres (2013), also calculated from the symmetric matrix of dissimilarities (measured by Whittaker's index of association) among all pairs of samples. This beta diversity metric is not dependent on the size of the data set, and the underlying Whittaker's index of association is directly related to the weighted-mean approach.

Available solutions and their comparison

To my knowledge two approaches have been introduced that attempt to solve the bias in the weighted-mean approach, namely the *modified permutation test* introduced (in the context of the CWM of species indicator values) by Zelený and Schaffers (2012), and the *sequential permutation test* using the *fourth-corner statistic*, introduced first in the electronic appendix of Peres-Neto et al. (2012, Appendix A) and later in a more elaborated version in Peres-Neto et al. (2016). Here I review the strengths and weaknesses of both approaches, test their performance using simulated community data and suggest guidelines for their use. Both a 2D simulation community data set and a 1D simulated data set according to Dray and Legendre (2008) have been used, with results of the former reported in the main paper and the latter in Appendix S4.

Note that hypotheses in category B are not prone to bias if the weighted-mean approach with standard test is used, and reviewed solutions are therefore relevant only for categories A and C.

Modified permutation test: comparison with the results of a null model

Comparison with results of a null model is analogous to testing the relationship between autocorrelated variables using toroidal shift, when one variable is permuted in a way that it preserves the original degree of spatial autocorrelation (Fortin and Dale 2005). Alternatively, one can generate random variables with the same degree of spatial autocorrelation as that of the original variable (Deblauwe et al. 2012). In the case of weighted-mean analysis with variables autocorrelated in the compositional space, such variables can be generated by calculating the CWM from randomized (or randomly generated) species attributes (CWM_{rand}). CWM_{rand} inherits the same level of compositional autocorrelation as those of the CWM values of the real species attributes (CWM_{obs}), because they are calculated by the same algorithm from the same species composition matrix. One can generate the null distribution of a test statistic (like *t*-value for correlation or *F*-value for regression) by repeated calculation of CWM_{rand} each time with newly randomized (or newly generated) species attribute values, and compare the observed statistic (relating CWM_{obs} to sample attributes) to this null distribution. This is identical to the *modified permutation test*, introduced to test the relationship between the CWM of species attributes with sample attributes by Zelený and Schaffers (2012) in the context of relating the CWM of species indicator values (e.g. those tabulated in Ellenberg et al. 1992) with internal variables (e.g. species richness or ordination scores based on the same species composition data set).

I illustrated the behavior of the *modified permutation test* using the 2D simulated community data sets. I calculated the correlation of the CWM of species attributes with sample attributes for all four scenarios in communities of increasing beta diversity, and tested the significance of this correlation using the modified permutation test. Results show that in contrast to the standard permutation test, the originally inflated Type I error rate in the case of scenario 2 disappears (Fig. 4). At the same time, in the case of scenario 3 the test is slightly conservative for homogeneous data sets. The same conclusion applies if the modified permutation test is used on Dray and Legendre's simulated community data set, in which the results for scenario 3 are even more conservative (almost no significant correlations, Appendix S4: Table S1 and Fig. S2a), since the community data set has rather low beta diversity (Appendix S4: Table S2). Additional detail power analysis on the simulated community data set with added random noise reveals that the modified permutation test loses power when both sample size and species number decrease (Appendix S4: Fig. S1a), and also with a decrease in the beta diversity of the data set (due to increased species tolerance, Appendix S4: Fig S1b).

The modified permutation test is suitable for testing hypotheses in category A, which assume that species attributes are random, while sample attributes are fixed (linked to species composition) and for which scenario 2 is relevant for testing the null hypothesis. It should, however, not be used for testing the hypotheses in the category B and C, since for both categories is relevant scenario 3 with fixed species attributes, which should not randomized (which is what modified permutation test is doing).

Sequential permutation test with the fourth-corner statistic

Dray and Legendre (2008) noted that the fourth-corner statistic *r*, introduced by Legendre et al. (1997), is "equal to the slope of the linear model, weighted by total species abundance, with the niche centroids as the response variable and the species trait as the explanatory variable." This analogy was further elaborated by Peres-Neto et al. (2012, Appendix A) and Peres-Neto et al. (2016), who presented an algorithm for how to use the *fourth-corner* statistic *r* in the weighted-mean approach. In short, both **R** and **Q** matrices are first centered by the weighted mean of row sums of **L** (in the case of **R**) and column sums of **L** (in the case of **Q**), and then standardized. Then, the fourth-corner *r* statistic is the slope of weighted regression between the weighted mean of centered plus standardized **Q** and centered plus standardized **R**, weighted by row sums of **L**. The main advantage of the fourth-corner statistic is the option to use the *sequential permutation test* introduced by ter Braak et al. (2012), which combines results of tests based on permuting sample attributes (model 2 in Legendre et al. 1997) and species attributes (model 4). If the first test is significant, then the second test is done, and overall significance of the result is equal to the higher of these two tests' *P*-values. When applied to the 2D simulated community data set, this test gives unbiased results for all scenarios (Appendix S2: Fig. S3), although being more conservative in the case of homogeneous data sets in scenario 4, which is relevant for questions in category C. Results calculated on the simulated data set of Dray and Legendre (2008) confirm this finding (Appendix S4: Table S1 and Fig. S2b). Power analysis (Appendix S4: Fig. S1c,d) reveals a performance very similar to that of the modified permutation test. The sequential test with the fourth-corner statistic is therefore suitable for testing hypotheses from all three categories, although in the case of category B it is not needed (standard permutation test gives unbiased results) and in the case of category C it is overly conservative for homogeneous community data sets (scenario 4 on Fig. 4). A disadvantage is that the sequential test with the fourth-corner statistic is restricted only to the weighted regression/correlation between centered and standardized species and sample attributes, weighted by row sums of a species composition matrix (**L**), and is therefore more like a special case of weighted-mean approach (which also includes other methods such as non-weighted regression, correlation or ANOVA and does not require standardizing species and sample attributes).

Discussion

The main motivation of this study was to show that the results of the weighted-mean approach critically depend on the correct decision being made regarding the test used for statistical inference. To help in this decision process, I suggested that each hypothesis can be classified into one of the three categories, given the explicit (or implicit) assumptions about the role of species and sample attributes. For each category, I suggested an optimal strategy for testing the significance of the relationship between the CWM and sample attributes, summarized in Table 1. The choice of the appropriate category is not always straightforward. For example, trait studies testing whether an environment is filtering the species into a community via their functional traits routinely assume that such traits are functional, and in the weighted-mean approach are therefore considered as fixed (category B). However, this assumption may not always be justified; traits included in these analyses are often those readily available in databases and/or relatively easy to measure, but these do not necessarily need to be really the functional ones (Mlambo 2014). In case of compositionally relatively homogeneous data sets, even the traits with no ecological meaning may show a high and significant relationship to environmental variables if tested by standard tests. I believe that this calls for a revision of such commonly applied practice.

The analogy between the bias in the weighted-mean approach to the bias in the analysis of spatially autocorrelated variables suggests some other alternatives to reduce or remove the bias. One is to stratify the data set to reduce redundancy in species composition among samples and increase the overall beta diversity of the compositional dataset, e.g. by removing one sample from pairs of samples with similar species composition. Although methods for stratification based on species composition are available (e.g. Lengyel et al. 2011), this potentially results in throwing out a large number of expensive data. Another option would be to apply some correction for effective degrees of freedom in analysis, analogous to a method estimating the effective number of samples in the case of autocorrelated variables (Dutilleul 1993), or to apply methods capable of dealing with autocorrelated residuals (analogy of geographically weighted regressions).

The analogy of the weighted-mean approach to the analysis of spatially autocorrelated variables also provides a solution to the question of how to deal with missing values for some of the species. Species with missing attribute values are not used for weighted-mean calculation, so they do not contribute to the compositional autocorrelation of CWM values. The point of the modified permutation test is to generate random variables with the same compositionally autocorrelated structure as the weighted mean calculated from the original species attributes. For this, the matrix of species composition, which inherits the compositional autocorrelation into weighted-mean values, should also remain the same for calculation of weighted means from randomly generated species attribute values. This would not be the case if the species with missing attribute values remains in both the composition and species attributes matrices, because permuting missing values would cause the weighted mean of permuted species attributes to be calculated every time with different species composition matrix (the species which in a given permutation run would be assigned missing values will not be included in this weighted-mean calculation). The solution is to remove species with missing species attributes from both the species attributes and the species composition matrix, and in the case of the modified permutation test to permute only existing species attribute values. In the case where more species attributes are analyzed (e.g. three different functional traits) and the species has missing species attribute values for some attributes and not for others, the species should be removed from the species composition matrix only for the purpose of calculating and testing the weighted mean of that species' attributes for which the species value is missing, and not for the others. Although not explicitly mentioned in the studies describing the sequential test with the fourth-corner *r* (Peres-Neto et al. 2012, 2016), I suggest that the same should also be done in the case of this approach.

The power test using the simulated data set showed that the power of both the modified permutation test and the sequential permutation test with the fourth-corner statistic decreases with a decrease in the number of species and/or number of samples. This makes these tests less suitable for smaller and relatively homogeneous data sets with few species (e.g. less than 40), since the probability of Type II error (i.e. not rejecting the null hypothesis, which is false) strongly increases. Additionally, in the case of a relatively homogeneous compositional data set the modified permutation test is overly conservative for scenario 3, while the sequential permutation test with the fourth-corner statistic is conservative for scenario 4. Both tests are therefore less suitable for testing hypotheses in category C in the case of a relatively homogeneous compositional data set.

In this study, I explicitly ignored intraspecific variation in species attributes, focusing only on the use of data set-wide mean species attribute values. Indeed, intraspecific variation may be important; e.g. in the context of functional traits, the intraspecific variation gains increasing attention (Albert et al. 2012), and a relevant question is whether the inclusion of intraspecific variation (e.g. by including trait values that are sample-specific, not data set-wide) influences the potential bias reported in this study or not. This question requires further examination, which goes beyond this study, but in my opinion including another source of variation (species-level variation in species attributes) does not remove the problem of the bias itself, but makes the estimation of the bias and its correction more complex.

Finally, relevant consideration is whether the weighted-mean approach is actually the best analytical solution for the question being explored. In some cases, the question is explicitly focused on relating community-level values of species attributes, like mean Ellenberg indicator values (serving as an estimate of ecological conditions for individual sites) or the CWM of traits (as one of the functional-diversity metrics and as a community-level trait value), and the use of the weighted-mean approach is fully justified. Yet, in other cases, when the question is focused on relating individual species-attributes to sample attributes, the weighted-mean approach may not be the best analytical choice. The use of alternative options, such as the fourth-corner or RLQ analysis, for which the problem of inflated Type I error rate and choice of suitable permutation test have already been solved, can be a better solution.

Conclusions

In this study, I attempted to draw attention to the problem of the weighted-mean approach, which I believe is largely overlooked and generally not acknowledged, although it represents a source of potentially serious misinterpretations. Since in certain fields the weighted-mean approach is gaining increasing momentum (e.g. in functional ecology with the CWM of species functional traits as one of the functional-diversity indices), I suggest that the time is ripe to critically assess in which situations and for which types of hypotheses the commonly used standard parametric or permutation tests are inappropriate, since they yield results that may be overly optimistic. I offer simple guidelines on how to decide whether, in a given context of a study, the standard methodology gives correct or biased results, and I review available solutions for those cases where it does not.

Acknowledgements

This study was supported by the Czech Science Foundation (P505/12/1022). My thanks go to Bill Shipley and Cajo ter Braak for critical comments as reviewers of the previous versions of this manuscript, which motivated me to heavily rework it, and also to Pedro Peres-Neto and Stephen Dray for discussion of differences between the modified permutation test solution and the fourth-corner one during the ISEC 2014 conference in Montpellier.

Literature cited

Albert, C. H., F. de Bello, S. Lavorel, and W. Thuiller. 2012. On the importance of intraspecific variability for the quantification of functional diversity. *Oikos* **121**:116–126.

Axmanová, I., et al. 2012. Estimation of herbaceous biomass from species composition and cover. *Applied Vegetation Science* **15**:580–589.

Birks, H. J. B., J.M. Line, S. Juggins, A. C. Stevenson, and C. J. F. ter Braak. 1990. Diatoms and pH reconstruction. *Philosophical Transactions of the Royal Society B Biological Sciences* **327**:263–278.

Bivand, R. 1980. A Monte Carlo study of correlation coefficient estimation with spatially autocorrelated observations. *Quaestiones Geographicae* **6**:5–10.

Brown, A. M., D. I. Warton, N. R. Andrew, M. Binns, G. Cassis, and G. Helois. 2014. The fourth-corner solution – using predictive models to understand how species traits interact with the environment. *Methods in Ecology and Evolution* **5**:344–352.

Carboni, M., D. Zelený, and A. Acosta. 2016. Measuring ecological specialization along a natural stress gradient using a set of complementary niche breadth indices. *Journal of Vegetation Science*, DOI: 10.1111/jvs.12413.

Clavero, M., and L. Brotons. 2010. Functional homogenization of bird communities along habitat gradients: accounting for niche multidimensionality. *Global Ecology and Biogeography* **19**:684–696.

Deblauwe, V., P. Kennel, and P. Couteron. 2012. Testing pairwise association between spatially autocorrelated variables: a new approach using surrogate lattice data. *PLOS One* **7**:e48766.

Díaz, S., M. Cabido, and F. Casanoves. 1998. Plant functional traits and environmental filters at a regional scale. *Journal of Vegetation Science* **9**:113–122.

Díaz, S., S. Lavorel, F. de Bello, F. Quétler, K. Grigulis, and T. M. Robson. 2007. Incorporating plant functional diversity effects in ecosystem service assessments. *Proceedings of the National Academy of Sciences USA* **104**:20684–20689.

Dolédec, S., D. Chessel, C. J. F. ter Braak, and S. Champely. 1996. Matching species traits to environmental variables: a new three-table ordination method. *Environmental and Ecological Statistics* **3**:143–166.

Dray, S., and P. Legendre. 2008. Testing the species traits-environment relationships: the fourth-corner problem revisited. *Ecology* **89** 3400–3412.

Dutilleul, P. 1993. Modifying the t test for assessing the correlation between two spatial processes. *Biometrics* **49**:305–314.

Ellenberg, H., H. E. Weber, R. Düll, V. Wirth, W. Werner, and D. Paulissen. 1992. Zeigerwerte von Pflanzen in Mitteleuropa. Second Edition. *Scripta Geobotanica* **18**:1–248.

Fortin, M.-J., and M. R. T. Dale. 2005. Spatial analysis. A guide for ecologists. Cambridge University Press, New York, USA.

Fridley, J. D., D. B. Vandermast, D. M. Kuppinger, M. Manthey, R. K. Peet. 2007. Co-occurrence based assessment of habitat generalists and specialists: a new approach for the measurement of niche width. *Journal of Ecology* **95**:707–722.

Garnier, E., et al. 2004. Plant functional markers capture ecosystem properties during secondary succession. *Ecology* **85**:2630–2637.

Hawkins, B. A., and J. A. F. Diniz-Filho. 2006. Beyond Rapoport's rule: evaluating range size patterns of New World birds in a two-dimensional framework. *Global Ecology and Biogeography* **15**:461–469.

Jamil, T., W. A. Ozinga, M. Kleyer, and C. J. F. ter Braak. 2013. Selecting traits that explain species-environment relationships: a generalized linear mixed model approach. *Journal of Vegetation Science* **24**:988–1000.

Juggins, S. 2013. Quantitative reconstructions in palaeolimnology: new paradigm or sick science? *Quaternary Science Reviews* **64**:20–32.

Kelly, M. G., and B. A. Whitton. 1995. Biological monitoring of eutrophication in rivers. *Hydrobiologia* **384**:55–67.

Kleyer, M., S. Dray, F. de Bello, J. Lepš, R. J. Pakeman, B. Strauss, W. Thuiller, and S. Lavorel. 2012. Assessing species and community functional responses to environmental gradients: which multivariate methods? *Journal of Vegetation Science* **23**:805–821.

Legendre, P. 1993. Spatial autocorrelation: trouble or new paradigm? *Ecology* **74**:1659–1673.

Legendre, P., R. Galzin, and M. L. Harmelin-Vivien. 1997. Relating behavior to habitat: solutions to the fourth-corner problem. *Ecology* **78**:547–562.

Legendre, P, and L. Legendre. 2012. Numerical ecology, Third English Edition. Elsevier Science, Amsterdam, The Netherlands.

Legendre, P., and M. De Cáceres. 2013. Beta diversity as the variance of community data: dissimilarity coefficients and partitioning. *Ecology Letters* **16**:951–963.

Lengyel, A., M. Chytrý, and L. Tichý. 2001. Heterogeneity-constrained random resampling of phytosociological databases. *Journal of Vegetation Science* **22**:175–183.

Mlambo, M. C. 2014. Not all traits are ‘functional’: insights from taxonomy and biodiversity-ecosystem functioning research. *Biodiversity and Conservation* **23**:781–790.

Peres-Neto, P. R., M. A. Leibold, and S. Dray. 2012. Assessing the effects of spatial contingency and environmental filtering on metacommunity phylogenetics. *Ecology* **93**:S14–S30.

Peres-Neto, P. R., S. Dray, and C. J. F. ter Braak. 2016. Linking trait variation to the environment: critical issues with community-weighted mean correlation resolved by the fourth-corner approach. *Ecography*, 10.1111/ecog.02302

Schaffers, A. P., and K. V. Sýkora. 2000. Reliability of Ellenberg indicator values for moisture, nitrogen and soil reaction: comparison with field measurements. *Journal of Vegetation Science* **11**:225–244.

Shipley, B. 2010. From plant traits to vegetation structure. Chance and selection in the assembly of ecological communities. Cambridge University Press, Cambridge, UK.

Sládeček, V. 1973. System of water quality from the biological point of view. *Archiv für Hydrobiologie* **7**:1–218.

Šmilauer, P, and J. Lepš. 2014. Multivariate analysis of ecological data using CANOCO 5, Second Edition. Cambridge University Press, Cambridge, UK.

ter Braak, C. J. F., and L. G. Barendregt. 1986. Weighted averaging of species indicator values: its efficiency in environmental calibration. *Mathematical Biosciences* **78**:57–72.

ter Braak, C. J. F., and C. W. N. Looman. 1986. Weighted averaging, logistic regression and the Gaussian response model. *Vegetatio* **65**:3–11.

ter Braak, C. J. F., A. Cormont, and S. Dray. 2012. Improved testing of species traits-environment relationships in the fourth-corner problem. *Ecology* **93**:1525–1526.

Vile, D., B. Shipley, and E. Garnier. 2006. Ecosystem productivity can be predicted from potential relative growth rate and species abundance. *Ecology Letters* **9**:1061–1067.

Whittaker, R.H. 1952. A study of summer foliage insect communities in the Great Smoky Mountains. *Ecological Monographs* **22:**1–44.

Zelený, D., and A. P. Schaffers. 2012. Too good to be true: pitfalls of using mean Ellenberg indicator values in vegetation analyses. *Journal of Vegetation Science* **23**:419–431.

Supplemental materials

Appendix S1. Description of an algorithm generating simulated community data along two environmental gradients (*2D simulated community dataset*).

Appendix S2. Weighted-mean approach applied to 2D simulated community data sets: additional results.

Appendix S3. The dissimilarity index between two CWM values and the beta diversity measure.

Appendix S4. Evaluation of permutation tests using 1D simulated data from Dray and Legendre (2008).

Data S1. R-code for all analyses.

Data S2. Source code for the R library *weimea*, version v. 0.62 (actual version can be found on https://github.com/zdealveindy/weimea/).

Table 1

Overview of the characteristics for the three categories of hypotheses tested by the *weighted-mean* approach. For each category, the corresponding assumption about a link between sample attributes (**R**) or species attributes (**Q**) and species composition (**L**) is provided, as well as the null vs alternative hypothesis, a scenario within the simulated data relevant in the context of a given category (see Fig. 1), and the recommended test (standard: standard parametric or permutation test; modified: modified permutation test; sequential with 4c: the sequential permutation test with the fourth-corner statistic).

Category of hypotheses		**A**	**B**	**C**
Assumption		sample attributes fixed	species attributes fixed	no assumptions
Null hypothesis		**Q** <-//-> **L**	**R** <-//-> **L**	**R** <-//-> **Q**, i.e. **R** <-//-> **L** and/or **Q** <-//-> **L**
Alternative hypothesis		**Q** <--> **L**	**R** <--> **L**	**R** <--> **Q**, i.e. **R** <--> **L** and **Q** <--> **L**
Relevant scenario(s)		Scenario 2	Scenario 3	Scenarios 2, 3 and 4
Recommended test	**standard**	no (biased result)	yes	no (biased result)
	modified	yes	no*	no*
	sequential with 4c	yes	yes (but not needed)	yes**

* species attributes in Scenario 3 are fixed and should not be permuted

** but too conservative if the beta diversity of the species composition matrix is low

<-//-> - no link between the two matrices, <--> - link between the two matrices.

Figure captions

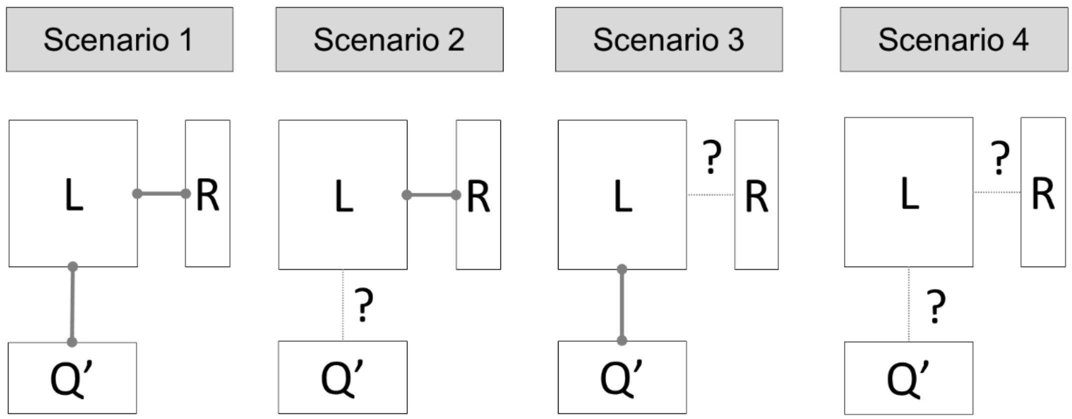
Figure 1. Conceptual differences between scenarios 1–4 in the weighted-mean approach. In scenario 1, both sample attributes (**R**) and species attributes (**Q**) are fixed, linked to matrix of species composition (**L**), while in the other three scenarios one or both attributes are considered random, without the link to species composition. In the simulated data example, the link of attributes to species composition is cancelled by permuting the values of species attributes (scenario 2), sample attributes (scenario 3) or both (scenario 4). In the schema, the matrix of species attributes is transposed (**Q'**) to match the dimension of the matrix of species composition (**L**).

Figure 2. Pearson's *r* correlation coefficients among CWM and sample attributes for each of the four scenarios and ten levels of beta diversity (1000 correlations for each combination have been conducted). Grey horizontal bars are outliers.

Figure 3. Proportion of significant correlations ($P < 0.05$) between CWM and sample attributes, tested by standard parametric *t*-test. For each of the four scenarios and ten levels of beta diversity, 1000 tests have been conducted.

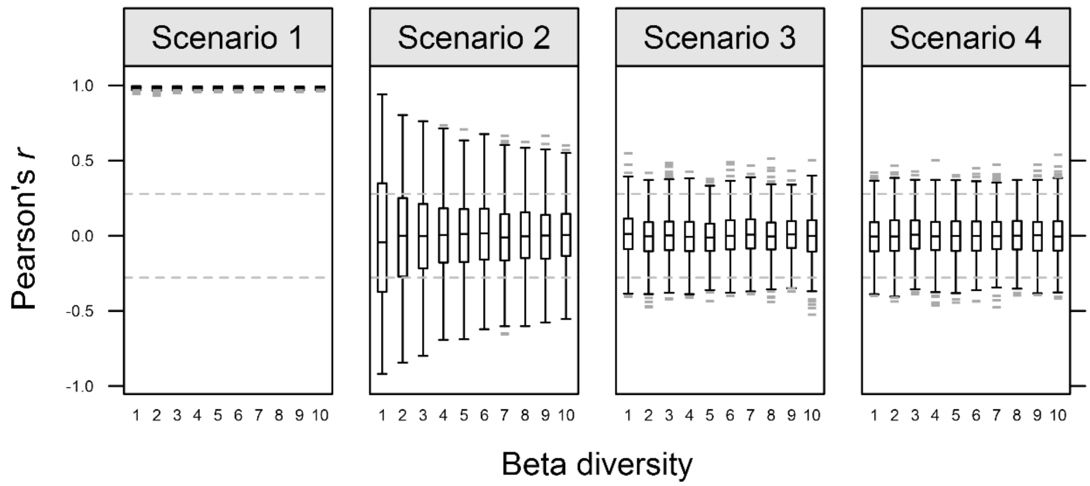
Figure 4. Proportion of significant correlations ($P < 0.05$) between CWM and sample attributes, tested by modified permutation test (white bars) and sequential test with fourth-corner *r* statistic (grey bars). For each of the four scenarios and ten levels of beta diversity, 1000 tests have been conducted for each method.

695 *Figure 1*

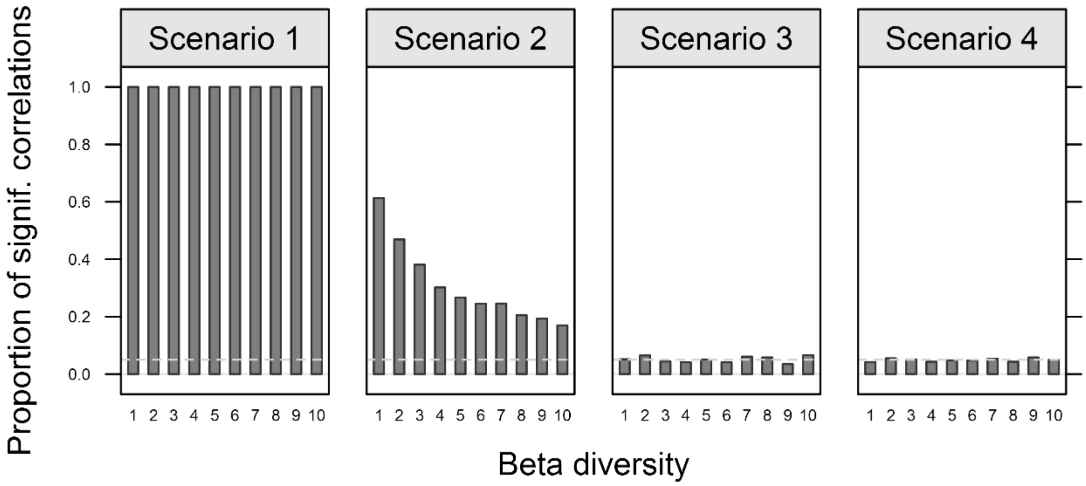


696
697

Figure 2


Scenario 1
Scenario 2
Scenario 3
Scenario 4
Pearson's r
1.0
0.5
0.0
-0.5
-1.0
1 2 3 4 5 6 7 8 9 10
Beta diversity

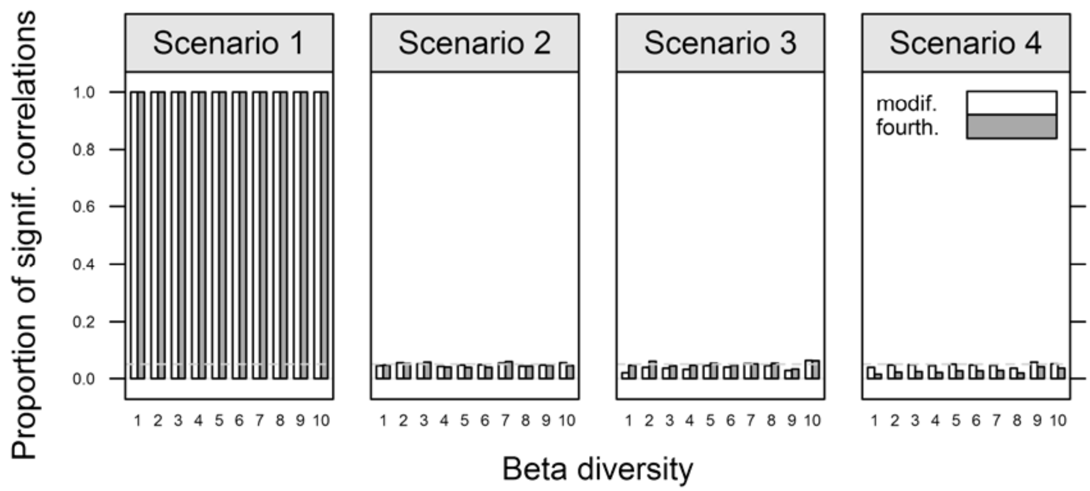
701 *Figure 3*



702

703

704 *Figure 4*



705