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, and

Department of Botany and Zoology, Faculty of Science, Masaryk University, Kotlářská 2, Brno 61137, Czech Republic

email: zeleny.david@gmail.com

 # Words: 10 718

 # References: 50

Abstract

A common way to analyse relationship between matrix of species attributes (like functional traits of indicator values) and sample attributes (e.g. environmental variables) via the matrix of species composition is by calculating community-weighted mean of species attributes (CWM) and relating it to sample attributes by correlation, regression, ANOVA or other method. This *weighted-mean approach* is used in number of ecological fields (e.g. functional and vegetation ecology, biogeography, hydrobiology or paleolimnology), and represents an alternative to other methods relating species and sample attributes via species composition matrix (like the fourth-corner problem and RLQ analysis).

Here, I point out two important problems of weighted-mean approach: 1) in certain cases, which I discuss in detail, the method yields highly biased results in terms of both effect size and significance of the relationship between CWM and sample attributes, and 2) this bias is contingent upon beta diversity of species composition matrix. CWM values calculated from samples of communities sharing some species are not independent from each other and this lack of independence influences the number of effective degrees of freedom. This is usually lower than actual number of samples entering the analysis, and the difference further increases with decreasing compositional heterogeneity of the dataset. Discrepancy between number of effective degrees of freedom and number of samples in analysis turns into biased effect sizes and inflated Type I error rate in case that significance of the relationship is tested by standard tests, a problem which is analogous to analysis of two spatially autocorrelated variables.

Consequences of the bias is that reported results of studies using rather homogeneous (although not necessarily small) compositional datasets may be overly optimistic, and results of

studies based on datasets differing by their compositional heterogeneity are not directly comparable. I describe the reason for this bias and suggest guidelines how to decide in which situations the bias is actually a problem for interpretation of results. I also introduce analytical solution accounting for the bias, test its validity on simulated data and compare it with an alternative approach based on the *fourth-corner* approach.

Introduction

Weighted-mean approach is a method to analyse link 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 (Fig. 1). *Species attributes* are species properties (traits), behaviour (species ecological optima) or phylogenetic age, while *sample attributes* are characteristics of community samples measured in the field (environmental variables) or derived from matrix of species composition (species richness or positions of samples in ordination diagrams).

Weighted-mean approach is used in wide range of ecological fields. In functional ecology, testing the effect of environmental variables on changes in CWM is one of the approaches demonstrating the effect of environmental filtering on trait-mediated community assembly (Diaz *et al.* 1998; Shipley 2010; Laliberté *et al.* 2012). Similarly, CWM are 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 (Diaz et al. 2007). In biogeography, grid-based means of species properties (like animal body size, range size or evolutionary age) are linked to macroclimate or diversity (Blackburn & Hawkins 2004,

Hawkins & Diniz-Filho 2006, Hawkins et al. 2014). Vegetation ecologists use species indicator values (e.g. those of Ellenberg et al. 1992 or Landolt 1977) to estimate habitat conditions from calculated mean species indicator values of vegetation samples and relate them to soil, light or climatic variables (Schaffers & Sýkora 2000, Wamelink et al. 2002, 2005). In hydrobiology, reliability of saprobic index of Sládeček (1973) based on weighted mean of diatom indicator values, or similar indices (e.g. trophic diatom index, Kelly & 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 pH gradient (ter Braak & Barendregt 1986) and as one of the transfer functions, (e.g. Birks *et al.* 1990) is considered to be a tool which have "revolutionised paleolimnology" (Juggins 2013). Other, more specific examples include relating community specialization index (mean of species specialization values weighted by their dominance in community) to environmental variables (Clavero & Brotons 2010, Fajmonová et al. 2013), or attempts to verify whether plant biomass can be estimated from tabulated plant heights and species composition as mean of species heights weighted by their cover in a plot (Axmanová *et al.* 2012).

Important thing to note is that although weighted-mean approach is technically relating two sets of variables (CWM and sample attributes), three matrices are in fact involved in the computation background (notation here follows 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 abundances (or presences-absences) of p species in n samples ($n \times p$) and matrix of *species attributes* **Q** with s species attributes for p species ($s \times p$). Weighted-mean approach is

just one of possible options how to tackle the problem of relating species attributes (**Q**) with sample attributes (**R**) via matrix of species composition (**L**): it combines **Q** with **L** into matrix of weighted-means **M** and relates it to **R** (Fig. 1). Alternative solution is to combine matrix of sample attributes **R** with species composition **L** by calculating weighted-mean of sample attributes (optima of individual species along given sample attribute or species centroids) and relate these values to species attributes **Q** (e.g. ter Braak & Looman 1986). 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), related ordination method called RLQ analysis (Dolédec *et al.* 1996) and other alternatives (Jamil et al. 2013, Brown et al. 2014).

In weighted-mean approach, relationship between CWM and sample attributes is in most cases tested by standard parametric or permutation test. However, not all types of ecological questions, which are usually solved by weighted-mean approach, should actually be tested by standard tests. In certain situations and types of null hypotheses, weighted-mean approach combined with standard tests generates biased results, which are more optimistic than would be actually warranted by analysed data. This bias includes unreliable estimates of effect size (e.g. correlation coefficients in case of correlation or r^2 in case of linear regression) and 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, based on critical inspection of the context of the study question and tested null hypothesis, 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). Inspiration for this issue can be seen in application of the fourth-corner

approach (Legendre et al. 1997), for which Dray & Legendre (2008) demonstrated the problem of deciding 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 situation of wrong choice. For *weighted-mean approach*, this issue was shown by Zelený & 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 also by Peres-Neto et al. (2012) in the context of metacommunity phylogenetics. Both studies proposed numerical solutions: Zelený & Schaffers (2012) introduced modified permutation test, based on permuting species instead of sample attributes, and Peres-Neto et al. (2012) suggested to use *sequential test* (ter Braak et al. 2012) using the *fourth-corner* statistic (Legendre et al. 1997). Additionally, Šmilauer & Lepš (2014) touched on this issue in the context of CWM-RDA method (Kleyer et al. 2012).

In this study, first I review categories of questions and null hypotheses which are commonly analysed by weighted-mean approach. Using simulated data, I show for which category there is a risk of biased results if tested by standard tests, and describe in detail what exactly causes this bias. Namely, I argue that the bias is caused by mismatch between number of samples in weighted-mean analysis and 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. I will also demonstrate that the amount of bias depends on the compositional heterogeneity (beta diversity) of the species composition matrix in a way that with increasing heterogeneity the bias decreases, which makes comparison of results between datasets of different compositional heterogeneity difficult. Note that for numerical simplicity, I ignore

intraspecific variation in species attributes. Finally, I will review methods available for solving the problem of inflated Type I error rate in weighted mean approach (namely *modified permutation test* of Zelený & Schaffers 2012 and solution based on combining *fourth-corner statistic* and *sequential permutation test* in weighted regression as introduced by Peres-Neto et al. 2012) and introduce novel solution, here called *two-step permutation test*. Although the examples, ecological interpretations and reasoning used here are focused on relationship of species functional traits or species indicator values with sample attributes analysed by weighted-mean approach, the general context is valid also for other types of species and sample attributes linked by weighted-mean approach.

Types of species and sample attributes

When thinking about possible alternative types of questions which are commonly being analysed using weighted-mean approach, it proves as useful to distinguish whether species and sample attributes are fixed or random. Terminology behind distinction into fixed and random is diverse (Gelman 2005); here I don't use this terms in the sense of ANOVA (fixed vs random terms), neither in the sense inferring their importance (fixed factors are important while random are nuisance factors). As *fixed attributes* I consider those which are specific for given dataset and related to given species or samples, and this specificity is acknowledged by the question/hypothesis being tested in a way that this link is deemed as given, and not further questioned or tested. *Random attributes* represent a subset of some larger pool of values, and their link to species composition is not acknowledged by the question being tested. In the narrow sense of permutation tests, fixed attributes should not be permuted among each other, while

random attributes can. For interpretation, effect of fixed attributes is limited only for given set of attribute values and in the context of community datasets included in the analysis and is not likely to be generalized beyond, while in case of random attributes interpretation is focused on the general effect of given attribute, not only the subset of attribute values used in the study. Species traits measured on individuals from plots of given community datasets can be considered as fixed, while species traits taken from large trait databases and measured often in completely different context should more likely be considered as random. Similarly, species richness or sample ordination scores derived from community matrix are more likely fixed sample attribute, while environmental variables measured in the field or derived from GIS layers may be considered as random. Indeed, this distinction is often dependent on author's view of the problem and on the theoretical context of the study, and the same variables can be seen as fixed or random in different context.

In the original description of the fourth-corner problem (Legendre et al. 1997), focused on linking fish behavioral and biological characteristics to environmental variables, both species and sample attributes have been considered as fixed, and random was matrix of species composition. Alternative permutation models were then used to test alternative hypotheses about the mechanisms assembling the community (like environmental control over individual species or species assemblages, lottery or random species attributes). In weighted-mean approach, decision about fixed or random nature of attributes directly influences the decision for meaningful way to test the relationship, and is therefore crucial for selecting correct statistical test. Important is to note that all hypotheses tested by weighted-mean approach make (implicit or explicit) assumption that either species or sample attributes are fixed, and their link to species

composition is *a priori* acknowledged and therefore not questioned (and not tested). In some cases, whether the species or sample attributes are better considered as fixed or random depends also on the need to generalize the results of the study. If the results should be used e.g. for local application (e.g. whether CWM of species height in given agricultural system can predict well the harvested biomass), species attributes can be seen as fixed, since the results will be used solely in the context of the studied system – if the same values are measured again for the same community, the results should be similar, but not applicable generally to other communities. If the aim is to generalize the results (e.g. to assess whether the species height itself, as tabulated in the floras, can be used as a tool to predict biomass yield), it is more reasonable to treat the species attributes as random and modify the analysis accordingly, so as even local study can contribute to more general description of such pattern.

Another useful distinction of attribute types is whether the sample or species attributes are internal or external. The main difference is that internal attributes are numerically derived from matrix of species composition, while external attributes are typically measured or estimated variables, not directly derived from species composition matrix. Internal species attributes are species optima calculated as weighted-means of sample attributes or as species scores on ordination axes, and similarly internal sample attributes are e.g. sample scores on ordination axes, species richness of individual samples or assignment of samples into groups based on compositional similarity (e.g. by numerical classification). External species attributes, in contrary, are measured traits or tabulated species indicator values, 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 should always be considered as fixed, since they refer only to the context of the dataset from which they have been derived.

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 weighted-mean approach fall into one of the three categories (see Table 1 for summary). *Category 1* assumes that while sample attributes are fixed, species attributes are random; *category 2* is opposite to the previous, with sample attributes considered random and species attributes fixed; and, finally, *category 3* 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 1 – 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, and they focus on testing the link of species attributes to species composition. The null hypothesis which is tested states that species attributes are not linked to species composition, while alternative hypothesis states that they are. Questions focused on relating CWM to internal sample attributes derived computationally from matrix of species composition fall into this category (e.g. relating mean Ellenberg indicator values to sample scores on unconstrained ordination, often used to interpret ecological meaning of ordination axes; Zelený & Schaffers 2012). Also studies with external sample attributes considered to be fixed, like experimental

treatments, fall into this category in case that their effect on species composition is acknowledged, and the question is about how does species attributes response to it. Additional level of complexity is added in studies dealing with grid data with both CWM and internal sample attributes (e.g. species richness derived from community data) spatially autocorrelated due to spatial coherence of species distribution (B. Hawkins, *pers. comm.*). Zelený & Schaffers (2012) showed that standard parametric and permutation tests has inflated Type I error rate for this category, and as an alternative introduced *modified permutation test*, permuting species attributes instead of sample attributes as further discussed in this study.

Category 2 – species attributes are fixed, sample attributes are random

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

Category 3 – both species and sample attributes are random

This category of hypotheses includes mostly observational studies without prior knowledge or expectations about link between any of matrices. The null hypothesis states that there is no link between species and sample attributes via the matrix of species composition because either species attributes or sample attributes (or both) are not linked to 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 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 CWM of functional traits to environmental variables without clear assumption that traits are functional, allowing to question whether particular trait are actually linked to species composition or not. In case of studies with species indicator values, these include relating mean indicator values to environmental variables with aim to generalize the result also out of the studied dataset (e.g. answering the question whether Ellenberg indicator values for soil reaction *per se* are good predictors of measured soil pH, i.e. not only in the context of given community dataset).

Illustration of the bias using simulated community data

In the next section, I will use simulated community data to illustrate performance of standard parametric test, if this is used to test hypotheses from each category defined above. The benefit of simulated data is the possibility to keep certain parameters fixed and to manipulate only those parameters whose effect is studied – in this case fixed are numbers of samples in the dataset, and

manipulated are links between species or sample attributes and species composition, and also compositional heterogeneity of community data.

Each artificial community dataset 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 permutation of attributes. This creates four scenarios (Fig. 2, identical with scenarios 1-4 of Dray & Legendre 2008): *scenario 1* - both sample and species attributes are linked to species composition, *scenario 2* - sample attributes are linked to species composition, but species attributes are not, *scenario 3* - species attributes are linked to species composition, but sample attributes are not, and *scenario 4* - none of species or sample attributes are linked to species composition. For hypotheses in category 1 defined above, the scenario 2 represents the null hypothesis, for category 2 scenario 3 is the null hypothesis, and for category 3 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 probability of getting significant results if the alternative hypothesis is true). Additionally, I also examined how observed bias depends on compositional heterogeneity (beta diversity) of community matrix, which influences the number of effective degrees of freedom in analysis (as explained in detail further in the section *Justification of the bias*). Note that all analyses in this paper were conducted using R-project (v. 3.2.3, R Core Team 2015); complete R scripts are available in Appendix S5, and all functions have been wrapped into R-packages *weimea* (abbreviation for *wei*ghted *mea*n; source code for v. 0.58 available as Appendix S6).

Description of simulated data

I created an algorithm generating community data which are structured by two virtual ecological gradients (by extending the original one-gradient algorithm of Fridley et al. 2007 based on concept of Minchin 1987). The first gradient has *constant* length for all generated datasets and serves as a surrogate for measured environmental variable; in the analysis, positions of samples along this gradient are used as *sample attributes*, while the optima of species along this gradient are used as *species attributes*. The length of the second gradient is *variable* and its increasing length increases the compositional heterogeneity of species composition matrix. Community samples based on this simulated community ecospace were created by randomly locating samples along the first gradient, and species composition for each sample was derived by random assignment of fixed number of individuals to species identities weighted by relative abundances of species with non-zero probability of occurrence at given location of the gradient (see Appendix S1 for further details). With short second gradient, the resulting simulated community dataset was compositionally relatively homogeneous, with samples located nearby along the first gradient (with similar value of sample attribute) sharing rather high proportion of species. Increasing the length of the second gradient increased the compositional heterogeneity of the dataset, since two nearby samples may have quite different species composition (Fig. S1 in Appendix S1). Note that although scenarios 1-4 are conceptually analogous to scenarios 1-4 in Dray & Legendre (2008) used in the context of the *fourth-corner approach*, the model generating simulated communities is different, since Dray & Legendre (2008) used one-gradient model, which generates rather homogeneous communities, while I used two-gradients model, generating set of communities of increasing compositional heterogeneity.

Weighted-mean approach with standard parametric tests applied on simulated data

Using the algorithm described above, for each of the four scenarios (1-4) I created ten levels of compositional heterogeneity, and for each combination of *scenario* × *level of heterogeneity* I created 1000 datasets (4 scenarios × 10 levels of heterogeneity × 1000 replications = 40 000 datasets). For each dataset I calculated CWM of species attributes, related it to sample attributes using Pearson's *r* correlation and tested its significance using parametric *t*-test (for additional results for least-square regression and r^2 see Fig. S2 in Appendix S2). For each level of community heterogeneity in each scenario I counted proportion of correlations significant at α = 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 (Fig. 3) reveals the bias – the correlation coefficient deviates from zero more than in other cases (Fig. 3), and the test of significance shows inflated Type I error rate (Fig. 4). This bias is decreasing with increasing heterogeneity of the species composition matrix (Fig. 3 & 4, Scenario 2): for the most homogeneous dataset (*number of communities* = 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 dataset with high beta diversity (*number of communities* = 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 in average between -0.278 and 0.281, with expected number of significant results being close to 5%). Similarly inflated are values of coefficient of determination (r^2; Fig. S2, Scenario 2) calculated by least-square linear regression.

Dray & Legendre (2008) showed that the *fourth-corner approach*, if tested by the permutation test based on reshuffling sample attributes (or rows of species composition matrix, respectively, model 2 in their paper) also reveals inflated Type I error rate in case of scenario 2. I applied the fourth-corner analysis also on the simulated community data described above. Results show, in line with Dray & Legendre (2008), biased values of the fourth-corner statistic and inflated Type I error rate for the model 2 permutation test (Figs. S3 & S4, Appendix S2), with the bias (and inflated Type I error rate) decreasing with increasing compositional heterogeneity (Fig S3 & S4 with Scenario 2, Appendix S2).

Justification of the bias

Simulation study above showed that if hypotheses in category 1 or 3, for which scenario 2 is relevant, are tested by standard parametric or permutation tests, results are prone to biased estimates of model parameters and inflated Type I error rate, and the bias is contingent upon compositional heterogeneity of the community dataset. In this section, I explore the reasons for this bias, which will be used as a theoretical base for solution proposed in the next section, and also explain the link of bias to compositional heterogeneity. Note that only hypotheses in category 1 and 3 are influenced by this bias, since for the category 2 with fixed species and random sample attributes, scenario 2 is not relevant and the bias therefore does not occur.

A peculiar feature of CWM of species attributes is their “numerical burden”, namely that they are calculated from species attributes assigned to individual species and from species composition of individual samples, and therefore inherit part of information from both sources. The numerical difference between calculated CWM values of two community samples is indeed

constrained by a difference in species composition of these two samples - if they have identical species composition (or identical relative species abundances), their calculated weighted-means do not differ (because they cannot), and if their species composition differs only slightly, the difference in their weighted-mean values is likely to be small. This property of weighted-mean values, which are not independent from each other, has notable consequences for analysis with sample attributes, if these are themselves related to species composition. In fact the situation is analogous to the analysis of two spatially autocorrelated variables, just the autocorrelation is not happening in the geographical space, but in the compositional space (more about this analogy in the next section). Two values of 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 - difference in their CWM are predictable (to some extent) from difference in their species composition in a way that the more similar is species composition of two community samples, the more similar must be also their calculated CWM. This problem scales up to the dataset level: in case of two compositional datasets with the same number of samples used in weighted-mean approach, the dataset which is compositionally more homogeneous has lower number of effective degrees of freedom compared to the more heterogeneous one.

Although there are many ways how to quantify compositional heterogeneity of the dataset, promising is to use beta diversity measure based on Whittaker's index of association (Whittaker 1952; in Legendre & Legendre 2012 as D_9). This index can be numerically derived from differences in species composition between two calculated CWM values, and therefore quantifies the dissimilarity in species composition which is directly related to weighted-mean approach

analysis (Appendix S3). To obtain single value of beta diversity for given dataset, one may use beta diversity metric of Legendre & De Cáceres (2013) quantifying the variation in species composition, which can be calculated also from symmetric matrix of dissimilarities among all pairs of samples (here using Whittaker's index of association). Advantage of this beta diversity metric is that it is independent on the size of the dataset, and the underlying dissimilarity coefficient is directly related to weighted-mean approach.

Below, I will first illustrate what I mean by the differences in *effective number of degrees of freedom* in analysis. I use a simple example focused on relationship between community-weighted mean of traits with environmental variables, and compare two contrasting sampling designs differing substantially by number of effective degrees of freedom they bring to the analysis. Then, I elaborate in more detail the analogy to analysis of two spatially autocorrelated variables, since it offers deeper insight to the problem and inspiration for potential solutions.

Two sampling designs: difference in effective number of degrees of freedom

The following example will illustrate the mismatch between number of effective degrees of freedom and number of samples in analysis, using species functional traits. The weighted mean of species functional traits, if combined with environmental variables, is used to investigate whether the environment filters species into community via their trait properties. Assumption behind is that the traits are functional, meaning that they directly influence the probability of species occurrence in community under given environmental conditions, and this hypothesis therefore classifies into the category 2 described above. For example, plant species specific leaf area (SLA, Reich et al. 1992) is known to be related to the plant requirements for light, and shade tolerant species restricted to more shady habitats have larger thinner leafs with higher SLA

values (e.g., Lambers et al. 2008). One may therefore expect that light acts as an environmental filter for species entering given community, and this filtering is (at least partly) happening because of species specific SLA. To support this reasoning, let's conduct imaginary experiment: collect a dataset about composition of understory species in the forest and analyse it by the weighted-mean approach. Let's keep things simple in this example and assume that we will compare two forest types, one with open and one with closed canopy, to see whether the closed canopy filters the understory species with high SLA. When preparing the design of data collection, we have two options. The first is to choose two vegetation types (e.g. at the level of association) with contrasting canopy openness, search for them in the study area (this could be just one forest complex with mosaic of both vegetation types, or a larger region where these vegetation types occur), sample their species composition and measure the light in the canopy and SLA of individual species. The second option is to sample open- and closed-canopy forests without any restrictions about their species composition, possibly in wider area. Following the first sampling design we get dataset where samples of open-canopy forest are compositionally similar to each other, as well as samples made in closed-canopy forest (but the composition of open-canopy forest is different from the close-canopy one). In the second scenario we probably get samples which all are having rather different species composition.

A simplified example how the species composition of six samples collected by these two sampling designs would look likes in extreme cases is on Fig. 5. In the case of the first sampling design, three and three samples have exactly the same species composition, while in the case of the second design, species composition of each sample is distinctly different from the others (no species are shared among any pair of samples). In the first case, three and three samples have the

same weighted-mean values since their species composition is identical. In the second case, three and three samples have also the same weighted-mean values, not because of identical species composition, but because samples from the same canopy cover category (open or closed) are made to have similar distribution of trait values (although the species taxonomic identities are completely different).

Indeed, the real data would fall somewhere in between these two cases, but this example illustrates well the concept of effective degrees of freedom in weighted-mean analysis. The relationship of the light availability to weighted-mean of SLA is exactly the same in both cases (if analysed e.g. by one-way ANOVA in this situation, when environmental variable is categorical with two levels, open-canopy forest with more light and close-canopy forest with less light). The difference is in the number of effective samples, which each sampling design offers for answering our question whether light serves as an environmental filter of species occurrence in the community via species SLA. Three community samples with completely different species composition, yet similar CWM of SLA (low for open- and high for closed-canopy forest) offers considerably better information about interaction between and SLA *per se*, then do the three samples with identical species composition (and hence also identical low or high CWM of SLA). Also, what if our assumption is wrong and the SLA is in reality not a functional trait, and hence belongs to category 3 instead of category 2? Let's replace the real trait values by random one (by randomizing species attributes among species in table), calculate CWM and test their difference between open- and closed-canopy stands (and repeat this process 1000 times). In case of the first sampling design, 87% of tests detect significant difference (874 significant results of one-way ANOVA at $P < 0.05$), while in case of the second sampling design the probability is near the

expected Type I error rate of 5% (54 significant results out of 1000). This agrees with the result of weighted-mean approach applied on simulated data saying that inflation of Type I error rate is high for homogeneous datasets (the case of the first sampling design) and decreases with increasing compositional heterogeneity (being effectively zero in dataset with all samples completely dissimilar as in case of the second sampling design).

Analogy to the analysis of spatially autocorrelated variables

The situation when *weighted-mean approach* is based on species composition data, in which some pairs of samples have the same or similar species composition and sample attributes are related to species composition, resembles an analysis of two spatially autocorrelated variables. In case of spatially autocorrelated variables, samples located more close to each other in geographical space have more similar values than expected if the values are randomly selected (Legendre & Legendre 2012). In case of weighted means, it is not the proximity in geographical space, but the proximity in compositional space, which reflects distances between samples expressed as their compositional dissimilarity.

It has been shown that when analysing two positively *spatially autocorrelated* variables, spatial autocorrelation biases the results of statistical tests, inflating the type I error rate and thus resulting into too optimistic results (Legendre 1993). The problem is not with autocorrelation of individual variables themselves, but with spatially autocorrelated residuals when analysing their relationship (e.g. by linear regression). From the point of view of the degrees of freedom, samples located nearby in geographical space are not statistically independent, behaving to a certain degree as pseudoreplications (Legendre & Legendre 2012). A new observation does not bring completely new information, because its value can be partly derived from the value of a

nearby site, and the effective number of samples (i.e., 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, appropriate for a given sample size, disparity between the real number and effective number of samples leads to selection of narrower confidence intervals and hence a higher probability of obtaining significant results (Bivand 1980; Legendre 1993).

The same reasoning applies also for analysis of two *compositionally autocorrelated* variables. Two weighted-mean values calculated from two samples with similar species composition do not bring two full degrees of freedom to the analysis, as would be case of two weighted-mean values calculated from samples with distinctively different species composition. In a simple example with two sampling designs above, if we want to know whether species SLA really increases with decreasing light in the understory, we would learn more about this from two samples in the shaded understory which have different species composition and yet both have high CWM of SLA, than from two samples which both have similar species composition (and hence also similar CWM of SLA).

If sample attributes are not related to species composition, than the problem with effective number of degrees of freedom is not present; although weighted-means are still compositionally autocorrelated, sample attributes are not – in case of spatially autocorrelated variables this is analogous to situation when one variable is spatially autocorrelated, but the other is not, in which case the bias caused by autocorrelation doesn’t appear. The situation which requires attention due to potential bias is therefore limited to cases when sample attributes are linked to species composition (i.e. they are fixed). This is the case of all internal sample

attributes derived from matrix of species composition, since they are linked to matrix of species composition (fixed) due to their numerical origin, and also the case of some of external sample attributes, if these are considered to be fixed (for examples see section *Types of species and sample attributes*).

Proposed solutions

Analogy between the bias in weighted-mean approach to the bias in analysis of spatially autocorrelated variables suggests potential toolbox for solving the problem. A simple option would be to stratify the dataset to reduce redundancy in species composition among samples, i.e. from pairs of samples with similar species composition remove one of them. Although methods for stratification based on species composition are available (e.g. Lengyel et al. 2011), it potentially results into throwing out a large number of expensive data. Alternative option would be to apply some correction for effective degrees of freedom in analysis, analogously to Dutiluel's method introduced for estimating effective number of samples in case of autocorrelated variables (Dutilleul 1993). The option I will further investigate here is based on comparison of results obtained by weighted-mean approach with those generated by a null model.

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

Comparison with results of a null model is an analogy 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 & Dale 2005). Alternatively, one can generate random variables with the same degree of spatial autocorrelation as of the original

variable (Deblauwe et al. 2012). In case of compositionally autocorrelated variables used for weighted-mean analysis, such variables can be generated for weighted-mean values, by calculating weighted mean from randomized (or randomly generated) species attributes. Such weighted-mean of randomized species attributes inherits the same level of compositional autocorrelation as have the weighted-mean values of the real species attributes, 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) for each weighted-mean of randomized species attributes related to original sample attributes, and compare the observed statistic (relating the weighted-mean of real species attributes to sample attributes) to this null distribution. This is identical with the modified permutation test, introduced to test the relationship between weighted mean of species attributes and sample attributes by Zelený & Schaffers (2012) in case of relating mean Ellenberg indicator values with variables derived from ordination/classification based on the same species composition dataset.

To illustrate behaviour of the modified permutation test, I used the set of artificial community data as above, calculated the correlation between weighted-mean of species attributes and sample attributes for all four scenarios in communities of increasing heterogeneity, and tested the significance of this correlation using modified permutation test. Results show that in contrast to standard permutation test, inflated Type I error rate in case of the scenario 2 disappears (Fig 6b). At the same time, in case of scenario 3 (species composition related to species attributes, but not to sample attributes) the test is overly conservative for homogeneous datasets. Additional power analysis (Appendix S4) reveals that the modified permutation test loses the power with decreasing sample size and mainly with decreasing number of species

which are being permuted (Fig S5 in Appendix S4). Modified permutation test seems therefore suitable for testing hypotheses in the category 1, 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 is, however, not optimal for hypotheses in the category 3, which assume that both species and sample attributes are not fixed (not linked to species composition), since in scenario 3, which is also relevant as an alternative null hypothesis for this category, the results are overly conservative (although only for the most homogeneous dataset, Fig. 6).

Use of the fourth-corner statistic and the sequential test

Dray & 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 abundances, 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), who presented algorithm how to use the *fourth-corner* statistic in weighted-mean approach. In short, both **R** and **Q** matrices are first centred by weighted mean of row sums of **L** (in case of **R**) and column sums of **L** (in case of **Q**), and rescaled; then, the fourth-corner *r* statistic is the slope of regression between weighted mean of standardized **Q** and standardized **R**, weighted by row sums of **L**. Advantage of the fourth-corner statistic is an option to use *sequential permutation test* introduced by ter Braak et al. (2012), which gives unbiased test of significance for all scenarios (for application on the simulated community data used above, see Fig. S3 & S4 in Appendix S2). This sequential permutation test combines results based on permuting sample attributes (model 2) and species attributes (model 4); if the first one is significant, than the second test is done, and

overall significance of the result is equal to the higher of these two test's *P*-values. Disadvantage, on the other side, is the fact that the combination of fourth-corner statistic and sequential test applies only to the regression between standardized (centred and rescaled) species and sample attributes, weighted by row sums of species composition matrix (**L**), and (to my knowledge) it cannot be used to test correlation, non-weighted regression or ANOVA between non-centred and standardized CWM and sample attributes.

Two-step permutation test

As an analogy to the sequential test used together with the fourth-corner statistic, here I introduce two-step permutation test, which gives unbiased results for relationship between CWM and sample attributes for range of statistical metrics (*t*-value for correlation and *F*-value for linear regression tested here). The test is based on combination of standard and modified permutation test; while both tests give unbiased results for scenario 4, standard test gives unbiased results also for scenario 3 (in which sample attributes are not related to species composition), while modified permutation test gives unbiased results for scenario 2 (where sample attributes are related to species composition). The idea behind the two-step permutation test is to first test whether sample attributes (**R**) are related to matrix of species composition (**L**), without considering (or even knowing) the values of species attributes (**Q**). This could be achieved e.g. by constrained ordination, when sample attributes are used as explanatory variables explaining variation in species composition. Here I introduce more general solution (called *LR permutation test* within this paper as a notice that relationship between **R** and **L** matrices is tested), which can be directly connected to particular test statistic (e.g. *t*-value for correlation). The LR permutation test consists of the following steps: (i) generate artificial set of species attributes as species centroids

calculated from real sample attributes, (ii) use these species attributes to calculate CWM, (iii) calculate observed test statistic for relationship between artificial CWM and real sample attributes, and (iv) test this relationship. The test is based on comparing observed values of the test statistic (calculated in step iii) with the null distribution of the test statistic, generated in the following way: 1) randomize sample attributes, 2) use them to calculate species attributes as species centroids from weighted-means of sample attributes, 3) use these species attributes to calculate CWM, and 4) relate these calculated CWM with randomized sample attributes from step 1) to obtain the expected test statistic; repeat steps 1) to 4) *n*-times (*n* = number of permutations). If this test is significant, it means that the sample attributes are related to matrix of species composition, and relationship of CWM with sample attributes is consequently tested by modified permutation test. If the test is not significant, standard permutation test is used. When applied on the set of artificial communities used above, this sequential test gives unbiased values of Type I error rate for all three scenarios (2, 3 and 4) and for all levels of compositional heterogeneity (Fig. 6).

Discussion

Main motivation of this study was to show that results of weighted-mean approach critically depend on the correct decision about 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 optimal strategy for testing the significance of relationship between CWM and sample attributes. The decision about appropriate category is not

always straightforward, although the decision whether species attributes should be considered as fixed or random changes classification of the hypothesis from category 2 (with recommended standard parametric or permutation test) into category 3 (with two-step permutation test). For example, trait studies, which are testing whether environment is filtering the species into community via their functional traits, routinely assume that such traits are functional, and in weighted-mean approach are therefore considered as fixed (category 2). 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 (Fox 2012, Mlambo 2014). In case of compositionally relatively homogeneous datasets, even the traits with no ecological meaning may show high and significant relationship to environmental variables if tested by standard tests. I believe that this calls for revision of such commonly applied practice.

Differences in effective degrees of freedom among datasets complicate comparison of results between studies based on datasets of different compositional heterogeneities. Studies conducted on datasets of relatively low beta diversity may obtain stronger and more likely significant relationship between weighted-mean of species attributes and sample attributes than studies on datasets of relatively higher beta diversity, even in case that the real link of species attributes to species composition is missing (Figs. 3 & 4, Scenario 2). This situation is analogous to biased estimates of model parameters and inflated Type I error rate in analysis of spatially autocorrelated variables. An option how to deal with this problem is to routinely report, in each case-study using weighted-mean approach, some standardized value of compositional heterogeneity. Although this would not remove bias in results of these studies, it would at least

allow for comparison of the potential for bias among different studies. Good metric for this purpose should be independent on the sample size, and should pertain dissimilarity in species composition which is relevant for differences in community-weighted means; here, I suggested beta diversity measure based on Whittaker's index of association (Appendix S3) following the approach summarized by Legendre & De Cáceres (2013).

Specific question is how to deal with missing values of species attributes for some of the species. Should species with missing species attributes remain in the matrix of species attributes and species composition? And in a case of the modified and two-step permutation tests, should the missing values be kept and permuted among species? The analogy to spatial autocorrelation issue offers clear answers for these questions. Species with missing attribute values are not used for weighted-mean calculation, so they do not contribute to the compositional autocorrelation of weighted-mean values. The point of the modified (and subsequently also two-step) 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, matrix of species composition, which inherits the compositional autocorrelation into weighted-mean values, should remain the same also 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 matrices, because permuting missing values would cause the weighted mean of permuted species attributes being calculated every time with different species composition matrix (the species which in given permutation run would be assigned missing values will not be included in this weighted-mean calculation). The solution is hence to remove species with missing species attributes from both species attributes and species composition matrix, and in the case of

modified permutation test to permute only existing species attribute values. In case that more species attributes are analyzed (e.g. three different functional traits, or six different species indicator values) and species has missing species attribute value for some attributes and not for the others, the species should be removed from species composition matrix only for the purpose of calculating and testing weighted mean of that species attributes for which the species value is missing, and not for the others.

Power test using simulated dataset showed that the power of both two-step as well as modified permutation test decreases with decreasing number of species in the dataset (and less strongly also with decreasing number of samples). This makes these tests less suitable for smaller and relatively homogeneous datasets 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. Similarly, both two-step and modified permutation tests are overly conservative for scenario 3. For modified permutation test this is not a problem, since for the hypotheses for which the scenario 3 is null hypothesis (category 2) the modified permutation test is not recommended method (see Table 1). The two-step permutation test is also overly conservative, but only only in case of the most homogeneous community dataset, and with increasing compositional heterogeneity this issue diminishes (Fig. 6 and Table S2).

In this study, I explicitly ignored intraspecific variation in species attributes, focusing only on use of dataset-wide mean species attribute values. Indeed, intraspecific variation may be important; e.g. in the context of functional traits, the intra-specific variation gains an increasing attention (Albert et al. 2012), and relevant question is whether the inclusion of intra-specific variation (e.g. by including trait values which are sample-specific, not dataset-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 question which is 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 estimates of ecological conditions for individual sites) or CWM of traits (as one of the functional-diversity metrics and as a community-level trait value), and use of weighted-mean approach is fully justified. Yet, in other cases, when the question is focused on relating individual species-attributes to sample attributes, weighted-mean approach may not be the best analytical choice. Use of alternative options, like fourth-corner or RLQ analysis, for which the problem of inflated Type I error rate and choice of suitable permutation test have been already solved, can be a better solution.

Conclusions

In this study, I attempted to draw attention to the problem in 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 gains increasing momentum (e.g. in functional ecology with CWM of species functional traits as one of the functional-diversity indices), I suggest that time is ripe to critically asses in which situations and for which types of hypotheses the commonly used standard parametric or

permutation tests are not appropriate, since they yield results which may be overly optimistic. I offer simple guidelines how to decide whether in given context of a study the standard methodology gives correct or biased results, and suggest solutions available in case that 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 on the previous versions of this manuscript, which motivated me to heavily rework it.

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.

Blackburn, T. M., and B. A. Hawkins. 2004. Bergmann's rule and the mammal fauna of northern North America. Ecography 27:715–724.

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.

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.

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

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.

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.

Fajmonová, Z., D. Zelený, V. Syrovátka, G. Vončina, and M. Hájek. 2013. Distribution of habitat specialists in semi-natural grasslands. Journal of Vegetation Science 24:616–627.

Fortin, M.-J., and M. R. T. Dale. 2005. Spatial Analysis. A Guide for Ecologists. Cambridge University Press, New York, USA

Fox, J. W. 2012. When should we expect microbial phenotypic traits to predict microbial abundances? Frontiers in Microbiology 3:268.

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.

Gelman, A. 2005. Analysis of variance – why is it more important than ever. The Annals of Statistics 33:1–53.

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.

Hawkins, B. A., M. Rueda, T. F. Rangel, R. Field, and J. A. F. Diniz-Filho. 2014. Community phylogenetics at the biogeographic scale: cold tolerance, niche conservatism and the structure of North American forests. Journal of Biogeography 41:23–28.

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.

Lambers, H., F. S. Chapin III, and T. L. Pons. 2008. Plant Physiological Ecology, 2nd Edition. Springer, New York, USA.

Laliberté, E, B. Shipley, D. A. Norton, and D. Scott. 2012. Which plant traits determine abundance under long term shifts in soil resource availability and grazing intensity? Journal of Ecology 100:662–677.

Landolt, E. 1977. Ökologische Zeigerwerte zur Schweizer Flora. Veröffentlichungen des Geobotanischen Institutes der Eidgenössischen Technischen Hochschule, Stiftung Rübel, Zürich, 64:1–208.

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. 2012. 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.

Minchin, P. R. 1987. Simulation of multidimensional community patterns: towards a comprehensive model. Vegetatio 71:145–156.

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.

Reich, P. B., M. B. Walters, and D. S. Ellsworth. 1992. Leaf life-span in relation to leaf, plant and stand characteristics among diverse ecosystems. Ecological Monographs 62:365–392.

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.

Wamelink, G. W. W., V. Joosten, H. F. van Dobben, and F. Berendse. 2002. Validity of Ellenberg indicator values judged from physico-chemical field measurements. Journal of Vegetation Science 13:269–278.

Wamelink, G. W. W., P. W. Goedhart, H. F. van Dobben, and F. Berendse. 2005. Plant species as predictors of soil pH: Replacing expert judgment with measurements. Journal of Vegetation Science 16:461–470.

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.

Supplementary materials

Appendix S1. Description of an algorithm generating artificial community data along two environmental gradients.

Appendix S2. Weighted-mean approach applied on simulated data: additional results.

Appendix S3. Dissimilarity index between two CWM values and beta diversity assessment.

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

Appendix S5. R-code for all analyses.

Appendix S6. Source code for the R library *weimea*, version v. 0.58 (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 *weighted-mean* approach. For each situation, corresponding assumption about link between sample attributes (**R**) or species attributes (**Q**) to species composition (**L**) is given, as well as null vs alternative hypothesis, scenario relevant in the context of given category (see Fig. 2), and recommended test.

Category	Assumption	Null hypothesis	Alternative hypothesis	Relevant scenario	Recommended test
1	sample attributes fixed (**R** <--> **L**)	**Q** <-//-> **L**	**Q** <--> **L**	Scenario 2	modified permutation test
2	species attributes fixed (**Q** <--> **L**)	**R** <-//-> **L**	**R** <--> **L**	Scenario 3	standard parametric or permutation test
3	no assumptions	**R** <-//-> **Q**, i.e. **R** <-//-> **L** and/or **Q** <-//-> **L**	**R** <--> **Q**, i.e. **R** <--> **L** and **Q** <--> **L**	Scenarios 2, 3 and 4	two-step permutation test

Figure captions

Figure 1. Computational schema of the weighted-mean approach to analyse relationship between species attributes and sample attributes via matrix of species composition. **R** - matrix of sample attributes (e.g. environmental variables), **L** - matrix of species composition ($\mathbf{L_s}$ – **L** standardized by sample totals to simplify the equation), **Q** - matrix of species attributes (e.g. traits, species indicator values), **M** - matrix of weighted means of species attributes (e.g. CWM). The colour gradient within the matrix M (weighted mean of species attributes) from dark to light grey illustrates that this matrix includes information from both matrix of species composition (dark grey) and matrix of species attributes (light grey).

Figure 2. Schema showing conceptual differences between scenarios 1-4 in 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) of attributes are considered random, without the link to species composition. In 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, matrix of species attributes is transposed (**Q'**) to match the dimension of matrix of species composition (**L**).

Figure 3. Pearson's *r* correlation coefficients among CWM and sample attributes for each of the four scenarios and ten levels of compositional heterogeneity of species matrix (1000 correlations for each combination have been conducted). Grey horizontal bars are outliers.

Figure 4. 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 compositional heterogeneity of species matrix, 1000 tests have been conducted.

Figure 5. Simplified example of two community datasets (each with six samples) collected in plots of two different environmental conditions (A and B, e.g. open- vs closed-canopy forest). First sampling design (a) restricts choice into only two vegetation types (one open- and one closed-canopy forest), and results into three and three plots with identical species composition. Second sampling design (b) does not restrict the sampling by choice of community type (any forest with open- or closed-canopy can be sampled), resulting in situation when none of six samples share any species in common. For each dataset, three matrices are presented: sample × species compositional matrix, matrix of sample attributes (in this case with two-level categorical variable) and matrix of species attributes (quantitative variable in range 1 to 5). x - presence of species in the sample.

Figure 6. Proportion of significant correlations ($P < 0.05$) between CWM and sample attributes, tested by three different permutation tests: standard, modified and two-step. For each of the four scenarios and ten levels of compositional heterogeneity of species matrix, 1000 tests have been conducted.

Figure 1

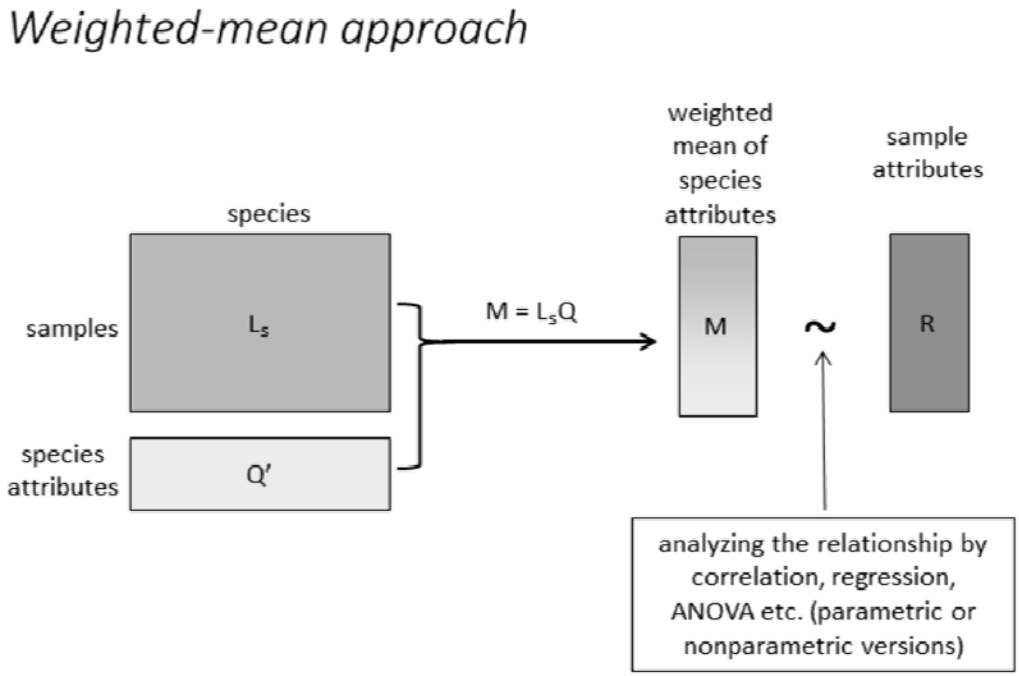

Weighted-mean approach
weighted mean of species attributes
sample attributes
species
samples
L_s
$M = L_sQ$
M
~
R
species attributes
Q'
analyzing the relationship by correlation, regression, ANOVA etc. (parametric or nonparametric versions)

Figure 2

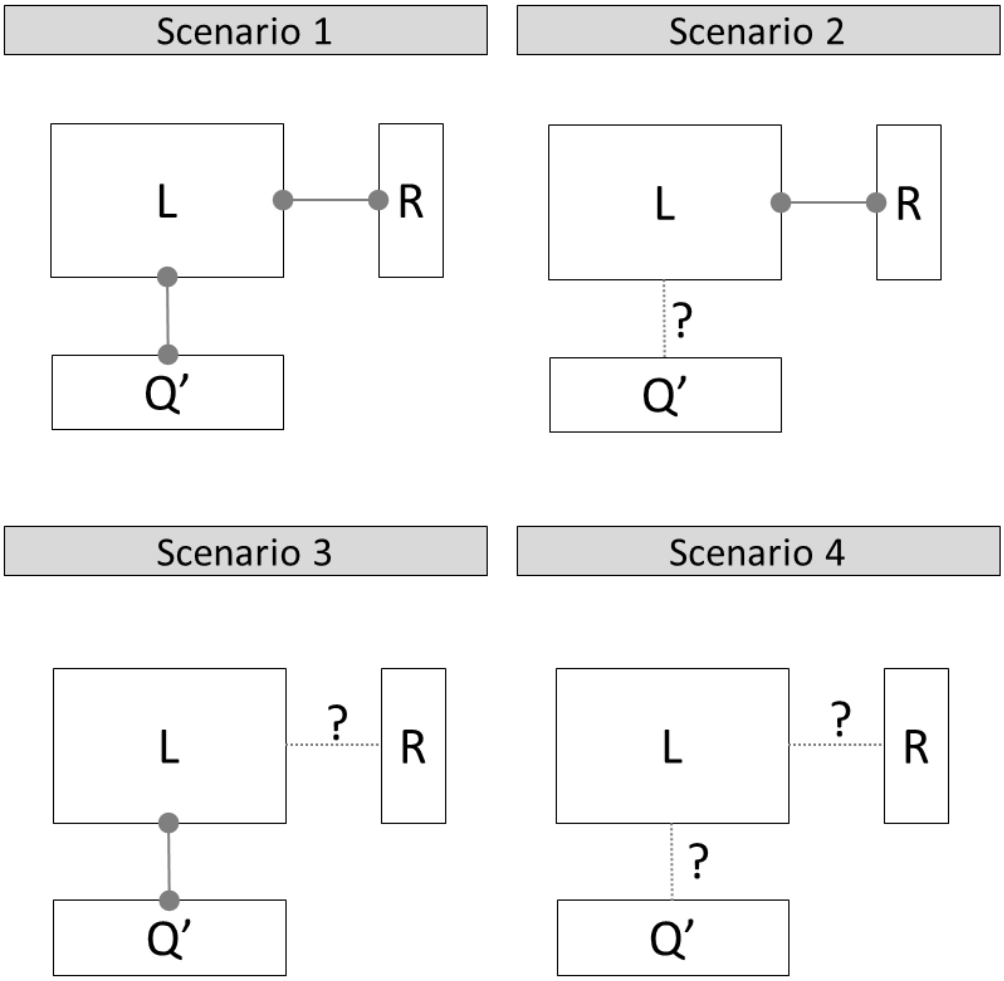


Figure 3

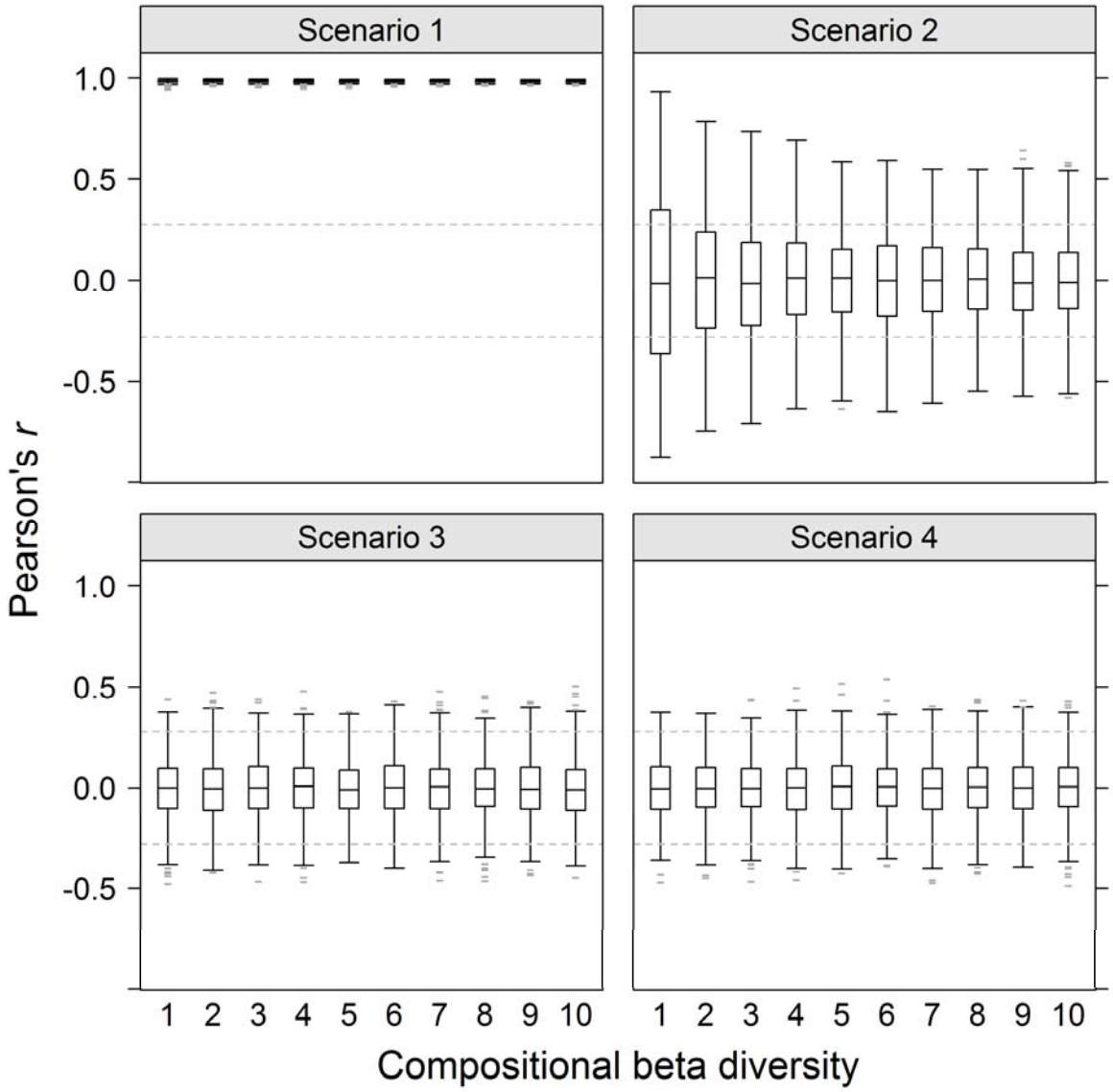

Scenario 1
Scenario 2
Scenario 3
Scenario 4
1.0
0.5
0.0
-0.5
Pearson's r
1 2 3 4 5 6 7 8 9 10
Compositional beta diversity

Figure 4

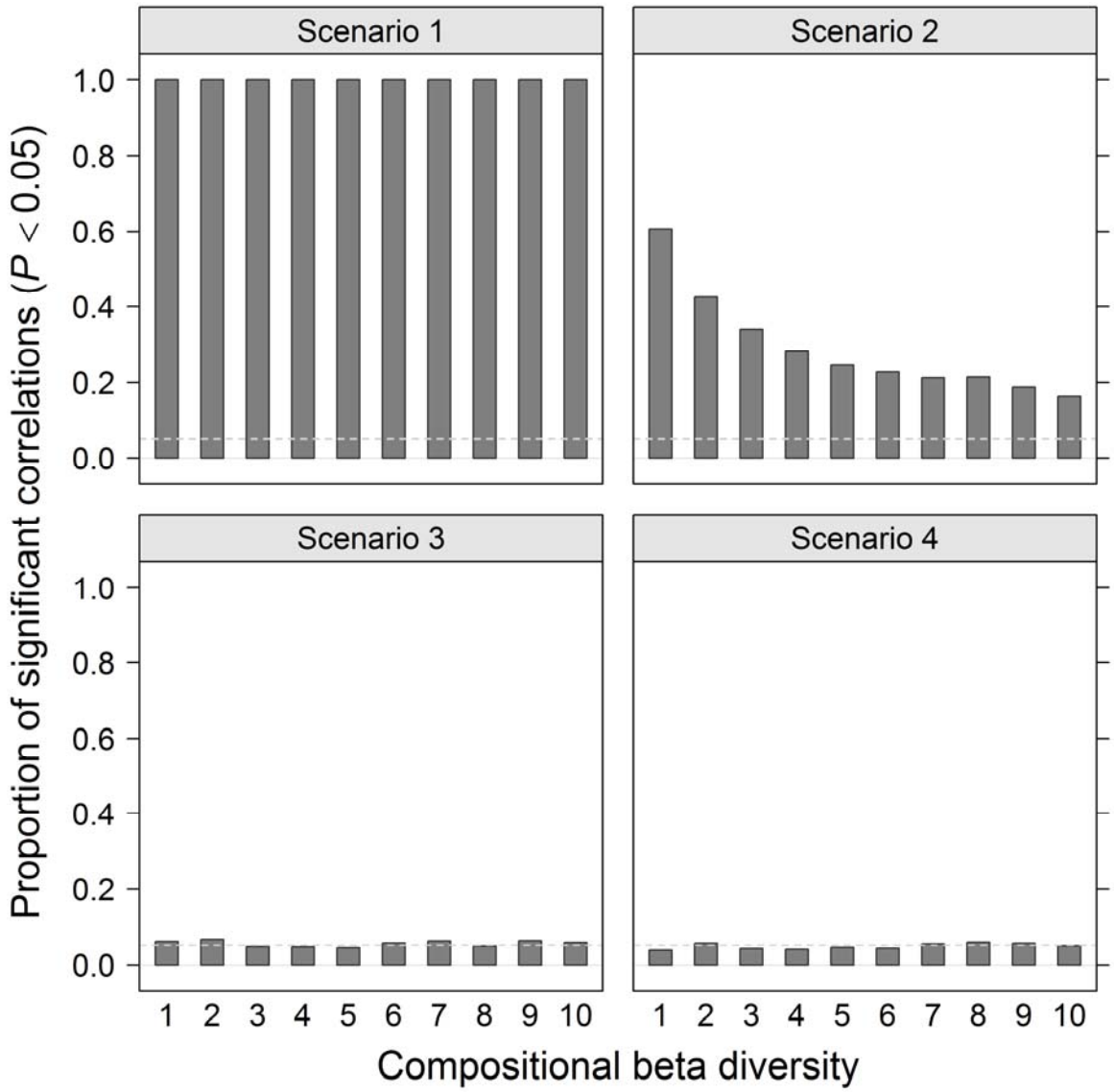
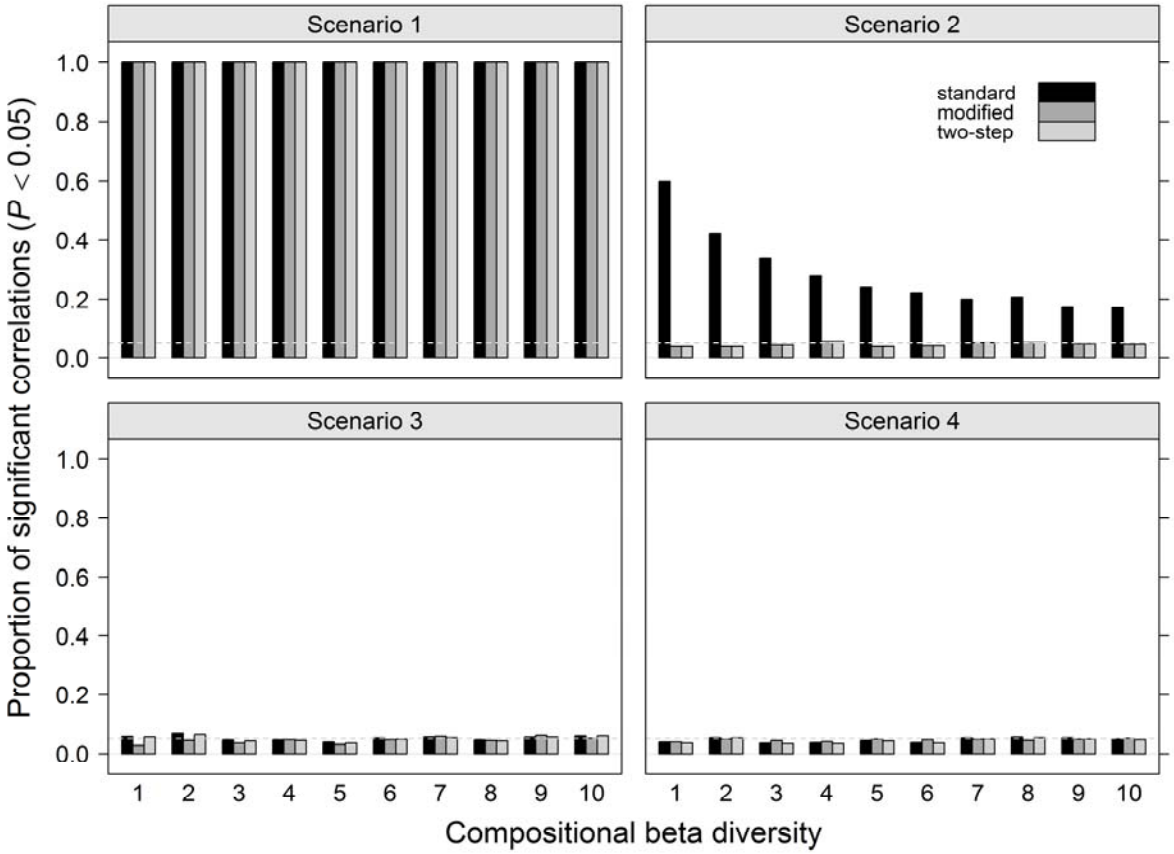

Scenario 1
Scenario 2
Scenario 3
Scenario 4
Proportion of significant correlations (P < 0.05)
1.0
0.8
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7 8 9 10
Compositional beta diversity

Figure 5

```
(a)
              species
                     11111
           12345678901234  env

sample_1   xxxxxxx         A
sample_2   xxxxxxx         A
sample_3   xxxxxxx         A
sample_4          xxxxxxx  B
sample_5          xxxxxxx  B
sample_6          xxxxxxx  B

spec.attr. 11122233344455
```

```
(b)
                              species
                     11111111112222222222333333333444
           12345678901234567890123456789012345678901 2 env

sample_1   xxxxxxx                                     A
sample_2          xxxxxxx                              A
sample_3                 xxxxxxx                       A
sample_4                        xxxxxxx                D
sample_5                               xxxxxxx         B
sample_6                                      xxxxxxx B

spec.attr. 111222311122231112223334445533444553344455
```

Figure 6


Scenario 1
Scenario 2
Scenario 3
Scenario 4
standard
modified
two-step
Proportion of significant correlations ($P < 0.05$)
1.0
0.8
0.6
0.4
0.2
0.0
1 2 3 4 5 6 7 8 9 10
Compositional beta diversity