ECOLOGICAL SOCIETY OF AMERICA

Ecology/Ecological Monographs/Ecological Applications

PREPRINT

The doi for this manuscript and the correct format for citing the paper are given at the top of the online (html) abstract.

Once the final published version of this paper is posted online, it will replace the preliminary version at the specified doi.

Consensus RDA across dissimilarity coefficients for canonical ordination of community composition data

F. Guillaume Blanchet[1, 2, 3, 4], Pierre Legendre[5], J. A. Colin Bergeron[3] and Fangliang He[3]

[1] Department of Biology, Section of Ecology, University of Turku, FIN-20014 Turku, Finland

[2] Mathematical Biology Group, Department of Biosciences, University of Helsinki, FIN-00014 Helsinki, Finland

[3] Department of Renewable resources, University of Alberta, 751 General Services Building, Edmonton, Alberta, Canada T6G 2H1

[5] Département de sciences biologiques, Université de Montréal, C.P. 6128, succursale Centre-ville, Montréal, Québec, Canada H3C 3J7

[4] E-mail: guillaume.blanchet@helsinki.fi

Corresponding address: F. Guillaume Blanchet, Department of Biology, Section of Ecology, University of Turku, FIN-20014 Turku, Finland

Abstract. Understanding how habitat structures species assemblages in a community is one of the main goals of community ecology. To relate community patterns to particular factors defining habitat conditions, ecologists often use canonical ordinations such as canonical redundancy analysis (RDA). It is a common practice to use dissimilarity coefficients to perform canonical ordinations through distance-based RDA (db-RDA) or transformation-based RDA (tb-RDA). Dissimilarity coefficients are measures of resemblance where the information about species communities is condensed into a symmetric square matrix of dissimilarities among sites. In this paper we compared 16 of the most commonly used dissimilarity coefficients to evaluate if the species abundance distribution (SAD) of a community can be used to select an appropriate coefficient. Of these, 11 are designed to be used primarily with abundance data although they can also be used with presence-absence data, whereas five can only be applied to presence-absence data. Using simulations, we compared the explained variance of RDAs differing only by their coefficients to evaluate how the abundance patterns of communities influence coefficient choice. We found that coefficients are largely equivalent, independently of the community SAD. In light of these findings, we propose the consensus RDA method, a new canonical ordination procedure that performs a consensus of RDAs across several coefficients. This new method focuses on the common relations found by independent RDAs differing only by their dissimilarity coefficients; this ensures the absence of a coefficient-related bias when interpreting the canonical ordination result. Also, because in our simulations the presence-absence data were directly derived from the abundance data, we were able to evaluate if the information in presence-absence data was equivalent to that in abundance data. We found that although some information was lost by converting abundance data into presence-absence, both data formats may be complementary. When applying consensus RDA to abundance and presence-absence data

independently, a more complete understanding and interpretation of the ecological patterns is obtained. An ecological example illustrating consensus RDA and the conclusions of our simulations is presented, using Carabidae data collected at the Ecosystem Management Emulating Natural Disturbances (EMEND) project in northwestern Alberta, Canada.

Key words. Dissimilarity coefficient, canonical redundancy analysis (RDA), Carabidae, species-abundance distribution (SAD).

INTRODUCTION

The species composition of an ecological community is heavily influenced by local variation in habitats. In theory, this intimate species-habitat relationship is due to evolutionary adaptations of the species to their environment; because of these adaptations, species have ecological niches (Hutchinson 1957), meaning that they are found at locations where they encounter appropriate living conditions. Whittaker (1967) illustrated this idea using the concept of environmental gradients where different species use distinct sections of the same gradient in a manner analogous to the dispersion of niches envisioned under Hutchinson's multivariate niche concept.

Numerous studies have shown that most communities that use a complex configuration of local habitats are composed of a few common species, plus a large proportion of less abundant or rare species. In contrast, species-poor communities with no dominant species are generally affected by only a few habitat gradients (Loreau 2010, Chapter 2). Thus, we may suggest that the complexity of species-habitat relationships influences the species abundance structure of a community.

Variation in species abundance and the effects of multi-habitat gradients on this variation have been studied extensively. A common approach for depicting variation in species abundance is the species abundance distribution (SAD), which ranks species in terms of the number of individuals of each species observed in sampling units from a community. SADs were mathematically described in the earlier in the ecological literature (Fisher et al. 1943 and Preston 1948). McGill et al. (2007) reviewed the various types of SADs and explained the utilities of SADs in describing and comparing communities.

At the community level, species-habitat relationships are often described using ordinations. Unconstrained ordinations such as principal component analysis (PCA, Pearson 1901), correspondence analysis (CA, Roux and Roux 1967), detrended CA (Hill and Gauch 1980), non-metric multidimensional scaling (Shepard 1962), and principal coordinate analysis (PCoA, Gower 1966) have been widely used to study associations between species and habitat factors (Legendre and Legendre 2012, Chapter 9). More recently, constrained ordinations such as canonical redundancy analysis (RDA, Rao 1964) and canonical correspondence analysis (CCA, ter Braak 1986, 1987) have been used to more directly evaluate how specific habitat components affect species assemblages. It is well known that RDA is not well-suited to the analysis of species abundance data collected along long gradients, which contain many zeros, because the Euclidean distance preserved in RDA does not have the property of being double-zero asymmetrical (Legendre and Legendre 2012, Subsection 7.2.2). Two variants of RDA have also been proposed to ecologists during the last 15 years: distance-base RDA (db-RDA, Legendre and Anderson 1999), which is a constrained version of PCoA, and transformation-based RDA (tb-RDA, Legendre and Gallagher 2001). Note that a PCA carried out on transformed data (tb-PCA) is the unconstrained version tb-RDA. The transformations used in tb-PCA and tb-RDA make

these ordination methods preserve one of the distances that is appropriate for the analysis of community composition data (Legendre and Legendre 2012, Sections 7.7, 9.1.10 and 11.1.5). In contrast with earlier approaches where the dissimilarity coefficient underlying the canonical ordination was fixed, db-RDA and tb-RDA make it possible to use an array of dissimilarity coefficients and data transformations to perform canonical ordinations, offering much more flexibility for the analysis of community data. A coefficient assesses the resemblance in species composition among sampled sites by condensing the community data into a symmetric square matrix of resemblance among sites. For example, the Euclidean distance (Table 1) computes Pythagoras' formula between all pairs of sites, which results in a symmetric square matrix where the species information is compared between two sites and condensed into a distance value.

Choosing a dissimilarity coefficient well suited to study specific communities and particular ecological questions is a problem often faced by ecologists because of the overwhelming number of coefficients available in the literature. As an example, Legendre and Legendre (2012, Chapter 7) describe 26 coefficients (distances and (dis)similarities) designed specifically for studying species assemblages. Although they propose theory-based guidelines and decision keys to help choose among coefficients (e.g., Legendre and Legendre 2012, Section 7.6), it often happens that more than one coefficient can be used to answer a particular ecological question. When such situations occur, Legendre and Gallagher (2001) suggest selecting the coefficient that yields the highest fraction of explained variance in canonical ordination; in other words, let the data determine which coefficient to use. Under this procedure, the abundance structure of a community can influence the selection of a coefficient used to describe it.

Although variation in SADs complicates coefficient selection, little is known about how variations in SADs affect the performance of coefficients. In this study, we compare the

performance of dissimilarity coefficients commonly used in canonical ordination and beta diversity studies of community composition data and use simulations to evaluate the sensitivity of the coefficients to varying SADs. The comparisons are made for communities described either in terms of abundance or presence-absence data. The analysis meets two objectives. Firstly, by comparing the performance of coefficients within data type, we show that the choice of a coefficient based on the proportion of explained variance may influence the resulting interpretation of the species-habitat relationship. To solve this problem, we propose a new technique that computes a consensus among the canonical ordination results obtained from several coefficients. Secondly, by comparing coefficients between data types, we evaluate the extent to which information in abundance data is preserved after transformation to presence-absence data. We illustrate these effects using ground beetle (Carabidae) data from a boreal forest in northwestern Alberta, Canada.

DEFINING A COMMUNITY WITH A SAD

There are many ways to display a SAD. In this paper, we use a variation of Preston's (1948) graphs to describe species abundance distributions where the abundance classes are arranged along the abscissa and increase according to a geometric progression, such that their lower bounds are 2^k where k represents the successive integers from 0 and up. This approach was recommended by Gray et al. (2006) as the SAD construction that most accurately represents the species-abundance pattern of an ecological community. These graphs can be compared visually, making them effective tools to differentiate communities.

The twenty-five graphs shown in Figure 1 present a range of possible SADs; most of which can be found in nature. All of them will later be employed to simulate site-by-species abundance matrices. For all SADs, the number of species was fixed to 20 but the total abundance

varied from 261 (the sum of the abundance classes' lower limits for all species of the community depicted by Figure 1a) to 20460 (the sum of the abundance classes' upper limits for all species of the community depicted in Figure 1j). Therefore, the SADs of Figure 1 represent a huge variation of species-abundance distributions, as would typically be observed in real communities (see Dewdney [2000] for a comparison of 50 SADs constructed from many different species communities). SADs were selected to represent a broad range of species abundance patterns found in natural communities.

Figure 1a-b present communities with many rare species and no common species. Note that communities with a similar SAD structure but with a larger number of rare species are often found in nature; however because the SADs in Figure 1 will later be used to define the abundance of species in simulated communities, the SADs in Figure 1a-b are the most extreme cases that would not generate empty sites in the site-by-species table.

Ecologists sometimes remove species with low abundances because the many zeros introduced by including these rare species can be troublesome for data analysis, especially with methods based on Euclidean distances, as explained by Legendre and Legendre (2012, Subsection 7.4.1). For example, in the classical Oribatid mite study of Borcard et al. (1992), 14 poorly represented species which, together, summed to 50 individuals, were removed from the data matrix before analysis by CCA. Depending on the group of organisms studied, removing rare species can yield SADs similar to what is found in Figure 1c-g, m-o, u-v.

In a recent paper, Gaston (2010) emphasized the importance of studying common instead of rare species. In light of that work, we included a few SADs (Figure 1h-j, w-y) corresponding to communities composed mainly of common species. Other SADs have been found to characterize well certain groups of organisms. For example, boreal carabid communities often

present bimodal SADs (Niemelä 1993) such as those in Figure 1k-l. Finally, the SADs presented in Figure 3p-t are mainly theoretical and unlikely to be found in nature. We included them because analysis of such extreme cases may lead to a better understanding of dissimilarity coefficients.

RDA AND DISSIMILARITY COEFFICIENTS

In this study, we used the RDA framework to compare commonly used dissimilarity coefficients (Table 1), all of which can be used within db-RDA. Although most models were constructed through db-RDA, the chord, χ^2, Hellinger, Ochiai, and distance between species profiles coefficients were applied in tb-RDA because it is computationally more efficient. These five coefficients are mathematically equivalent in tb-RDA and db-RDA (Legendre and Legendre 2012, Section 7.7).

For presence-absence data, the Euclidean distance is equal to the square root of the complement of the simple-matching coefficient (first entry of Table 1) multiplied by the number of species p: $\sqrt{p(1 - \text{simple matching coefficient})}$; the formula reduces to $\sqrt{b + c}$ (see Table 2 for the meaning of b and c). This relationship was shown by Gower (1966) when he described PCA based on binary descriptors. A PCA based on binary data produces the same ordination as the principal coordinate analysis of a matrix of $\sqrt{1 - \text{simple matching coefficient}}$; between the two ordinations, the coordinates are strictly proportional and differ by a constant factor of $\sqrt{p}$. The same relationship holds when binary descriptors are used in an RDA because it is the canonical extension of PCA. As a consequence, RDA based on binary data is equivalent to db-RDA of a matrix of $\sqrt{1 - \text{simple matching coefficient}}$ and no data transformation is required.

By using the RDA framework for all coefficients, we were able to compare our simulation results directly. In particular, we used the χ^2 distance through the tb-RDA approach instead of calculating CCAs. In practice, tb-RDA with the χ^2 distance coefficient and CCA yield very similar, although not identical, ordination results (Legendre and Gallagher 2001).

An RDA is computed by regressing the community matrix **Y**, composed of *p* species, on a matrix of *m* explanatory variables **X** observed at the same *n* sites. This is carried out by a sum of squares minimization, leading to

$$\mathbf{B} = (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y}$$
$$\hat{\mathbf{Y}} = \mathbf{XB} \qquad (1)$$

where *t* indicates the transpose and –1 the inverse of a matrix. **X** must either be centred by columns, or contain a column of 1's to estimate the regression intercepts. In Equation 1, **B** is the matrix of regression coefficients of all species in **Y** on the explanatory variables **X**. The residuals of the models are obtained through Equation 2:

$$\mathbf{Y}_{\mathbf{res}} = \mathbf{Y} - \hat{\mathbf{Y}} \qquad (2)$$

By performing a PCA on $\hat{\mathbf{Y}}$, a matrix of eigenvectors **U** defining the species scores and a diagonal matrix of eigenvalues **Λ** are obtained. The site scores can then be computed using **X** (3.1) or **Y** (3.2).

$$\mathbf{Z} = \mathbf{XBU} = \hat{\mathbf{Y}}\mathbf{U} \qquad (3.1)$$

$$\mathbf{F} = \mathbf{YU} \qquad (3.2)$$

If required, the canonical coefficients can be calculated following Equation 4:

$$\mathbf{C} = \mathbf{BU} \qquad (4)$$

A more detailed description of the RDA algebra is available in Legendre and Legendre (2012, Section 11.1).

These calculations are exactly the same for tb-RDA, with the exception that the community matrix **Y** is pre-transformed before calculating an RDA, using any of the transformations proposed by Legendre and Gallagher (2001). In db-RDA, a dissimilarity coefficient is applied to a community matrix, yielding a dissimilarity matrix. A PCoA is then calculated on this dissimilarity matrix and all the eigenvectors given by the PCoA are used as the **Y** matrix in an RDA (Legendre and Anderson 1999). In db-RDA, the sites scores (Equation 3) and canonical coefficients (Equation 4) are readily obtained. However, the species scores need to be calculated *a posteriori*. We used the procedure proposed in the vegan package (Oksanen et al. 2012) to calculate the species scores:

$$\mathbf{U} = \frac{\mathbf{Y}^t\mathbf{Z}\mathbf{\Lambda}^{-1/2}}{\sqrt{n-1}} \quad (5)$$

All binary similarity coefficients with the exception of the Raup-Crick coefficient were transformed into dissimilarities using $\sqrt{1-\text{coefficient}}$ because Gower and Legendre (1986) have shown that this transformation makes them metric as well as Euclidean. This is important because a PCoA of these transformed coefficients does not produce negative eigenvalues that would have to be corrected for before performing the RDA. Thus, this transformation facilitates the calculations. In contrast, the probabilistic nature of the Raup-Crick coefficient makes it special: on the one hand, the P-value behaves like a dissimilarity, increasing as sites become more different in species composition; on the other hand, two sites with exactly the same species will not necessarily result in a dissimilarity of 0 for this coefficient; neither will two sites with completely different species automatically lead to a dissimilarity of 1. We decided to include it in

our analyses because the probabilistic nature of the Raup-Crick coefficients may offer a solution to the double-zero problem.

The double-zero problem stems from the difficulty of relating two sites where a species has not been found (Legendre and Legendre 2012, Subsection 7.2.2). Double-zero asymmetrical dissimilarity coefficients are designed to ignore double zeros altogether whereas double-x (where $x > 0$) reduce the dissimilarity; for binary dissimilarity coefficients, this amounts to ignoring the value d (Table 2) in the calculation of the coefficients. Conversely, double-zero symmetrical coefficients treat double zeros as any other double-x value, which reduces the dissimilarity. For example, for the simple-matching coefficient, which is double-zero symmetrical, double zeros (value d in Table 2) are considered as an indication of similarity in the same way as double presences (value a). Double-zero symmetrical coefficients should be used only when the goal of a study is to evaluate total changes in a community, for instance under the influence of pollution. Studies focussing on the impact of predation or disturbances may also find symmetrical coefficients interesting because the absence of a species at two sites is ecologically meaningful and should be considered (Anderson et al. 2011). For studies focussing on the variation in community composition among sites (i.e. beta diversity), double-zero asymmetrical coefficients should be preferred (Legendre and De Cáceres 2013).

In the present study, we performed simulations that reflected species variation in undisturbed communities where predation was not considered. The Euclidean and simple matching coefficients are ill adapted to these types of ecological problems because they are double-zero symmetrical (Legendre and Legendre 2012, Subsection 7.4.1). We decided to include them when comparing coefficients within data types because both coefficients have been

used, perhaps wrongfully, to study ecological communities through the use of RDA on abundance or presence-absence data (ter Braak and Verdonsschot 1995).

The Jaccard, Sørensen, and simple-matching coefficients were computed with the ade4 package (Dray and Dufour, 2007). All other calculations were performed with the vegan package (Oksanen et al. 2012) with the exception of the Raup-Crick coefficient, which was programmed independently using the McCoy et al. (1986) permutation procedure. We used the McCoy et al. (1986) permutation approach because Legendre and Legendre (2012, Subsection 7.3.5) found that it was better at recognizing significant site associations, compared to the original permutation procedure of Raup and Crick (1979). All analyses were carried out using the R statistical language (R Development Core Team 2012).

SIMULATING COMMUNITIES WITH VARYING SPECIES ABUNDANCES

In our simulations, we constructed eight explanatory variables at 49 sites structured as a regular grid comprising 7×7 sites, using the RsimSSDCOMPAS package (Ouellette and Legendre 2011) within the R statistical language. These explanatory variables (matrix **X**) define linear gradients, waves, large patches, or random patterns. They are presented in Figure A1 of Appendix A with a detailed description of how they were constructed. The same eight descriptors were used for all simulations.

In a simulated community, each of the 20 species had a different underlying structure constructed by combining pairs of the eight explanatory variables presented above. This structure remained constant for all simulated communities. The reference structure $\mathbf{y}_{ref}$ of a species was constructed following Equation 6, where ω is a weight, $\mathbf{x}_i$ and $\mathbf{x}_j$ are two of the eight explanatory variables, and $\boldsymbol{\varepsilon}$ is an error vector of standard normal deviates.

$$\mathbf{y}_{ref} = \omega\,(\mathbf{x}_i + \mathbf{x}_j) + \varepsilon \quad (6)$$

The weight ω acts as a regression coefficient to influence the abundance of each species in the community, which is directly related to the size of the absolute value of ω (i.e. $|\omega|$). A value of ω was predefined for each species. A large $|\omega|$ generates species with larger abundances. Half of the species were constructed with positive weights and the other half with negative weights.

Ten species were characterized by strong links ($\omega = 2$ or -2) with the explanatory variables defining them. In ecological terms, a large absolute weight represents a species that has a strong relationship with the measured environmental variables. The other ten simulated species had smaller weights representing medium (two species with $\omega = 1$ or -1), weak (four species with $\omega = 0.5$ or -0.5), or very weak (four species with $\omega = 0.1$ or -0.1) relationships between a species and the descriptors controlling it.

As will be explained at the end of this section, additional sets of communities were simulated where the error $\boldsymbol{\varepsilon}$ was smaller, giving more importance to species with lower absolute weights. Note that Equation 6 without the error term $\boldsymbol{\varepsilon}$ represents the true pattern defining a species. The reference structure of each species was determined following a predefined combination of ω, $\mathbf{x}_i$, and $\mathbf{x}_j$ (Table A1). Also, the explanatory variables used to construct each species were carefully selected in such a way that each one was independently used to create five different species, making all explanatory variables equally important in the simulated community.

To construct a species, we transformed $\mathbf{y}_{ref}$ for it to range from 0 to 1 in order to use the information it encompasses as a probability distribution. Equation 7.1 was used if ω was positive

and Equation 7.2 if ω was negative. In these two Equations, $|\mathbf{y}_{ref}|$ is the absolute value of $\mathbf{y}_{ref}$ and $\mathbf{y}_{prob}$ defines the probabilities of sampling a species at each of the 49 sites in the sampling area.

$$\mathbf{y}_{prob} = \frac{|\mathbf{y}_{ref}|}{\sum |\mathbf{y}_{ref}|} \quad (7.1)$$

$$\mathbf{y}_{prob} = \frac{1}{|\mathbf{y}_{ref}|} \times \frac{1}{\sum |\mathbf{y}_{ref}|} \quad (7.2)$$

Equation 7.1 defines the probability of sampling a species directly related to the patterns in $\mathbf{y}_{ref}$ whereas Equation 7.2 defines the probability of sampling a species inversely related to the patterns in $\mathbf{y}_{ref}$. If the probability of sampling a species is high for a site in proportion to the other sites, it is more likely for at least one individual of that species to be found at the site.

As explained in section *Defining a community with a SAD*, the abundance patterns of each simulated community (defined as a group of species living in heterogenous environment) followed one of the predefined SADs presented in Figure 1. The SAD is a commonly used tool to rank species, based on the abundance of each species sampled from a community (McGill et al. 2007). The structure of these SADs were unaffected by the other steps of the simulation; SADs remained constant throughout the simulations. Each species was assigned to a bin of the SAD in order for the abundance distribution of the community to be reproduced when summing the number of individuals for each species in the site-by-species table. To define the exact abundance of a species in a simulated community, we randomly sampled the number of individuals of that species within its SAD bin boundary. Each number of individuals had the same probability of being selected within a particular bin boundary. To allocate these individuals to specific sites, we sampled, with replacement, the sites using the species probability

distribution $\mathbf{y}_{prob}$. Because $\mathbf{y}_{prob}$ is constructed from $\mathbf{y}_{ref}$, which has an underlying normal error, $\mathbf{y}_{prob}$ also follows a normal distribution.

By repeating this procedure for the 20 species, we obtained a site-by-species table representing one simulated community. We constructed 1000 communities for each of the 25 SADs in Figure 1. Four other sets of 25 000 communities were also constructed where the error terms $\boldsymbol{\varepsilon}$ in Equation 6 were standard normal deviates with standard deviations of 0.001, 0.25, 0.5, and 2. In all, we simulated 125 000 communities describing the abundance of species at each site.

To create site-by-species presence-absence tables, we transformed all abundances larger than 0 to 1s for all species abundance community data generated above.

COMPARING DISSIMILARITY COEFFICIENTS USING EXPLAINED VARIANCE

The amount of explained variance in canonical ordinations was estimated with the coefficient of determination (R^2) and the comparison of dissimilarity coefficients was carried out following the procedure proposed by Legendre and Gallagher (2001). Coefficients of determination were calculated by dividing the total variance in $\hat{\mathbf{Y}}$ (which, incidentally, is also the sum of the canonical eigenvalues) by the total variance in $\mathbf{Y}$ (which is also the sum of all eigenvalues, canonical and non-canonical). R^2 values range from 0 to 1. For example, if a model yields an R^2 of 0.2, it should be understood that this model explains 20% of the variance of the response.

In the present study, only the canonical eigenvalues associated with significant canonical axes ($P \leq 0.05$ after 999 random permutations) were considered in the calculation of R^2. Figure 2 compares the performance of RDAs for different dissimilarity coefficients for each of the 25

SADs presented in Figure 1. The RDAs were carried out on the simulated species abundances constructed with the smallest error (normal distribution with a standard deviation of 0.001). Results of simulations with larger error are presented in Figures B1-B4. All simulations yielded the same conclusions (see next paragraph) regardless of the error size. The only difference between the sets of simulations is that larger error when constructing species is associated with lower R^2. The inverse relation between error term and variance explained, which is consistant for all coefficients compared, suggests that the amount of error does not favour (or disfavour) any coefficient. Note that if all canonical eigenvalues are used to calculate the R^2 instead of using only the significant eigenvalues, the conclusions are unchanged because the fractions of the explained variance corresponding to the non-significant canonical axes are too small to markedly affect the results. The variance explained by all the non-significant canonical axes considered together is above 0.1 only in extreme cases and is usually around 0.06. The variance explained by a single non-significant canonical axis is usually less than 0.025.

In the simulation results presented in Figure 2, the most striking feature is that the confidence intervals for all double-zero asymmetrical coefficients overlap considerably. Moreover, detailed inspection of the results shows that independent of the SAD structures, a community having a high R^2 for one coefficient generally also has high R^2 for other coefficients.

The R^2 values for the Euclidean distance differ most from the other coefficients, although its confidence intervals still overlap with the other coefficients (Figure 2, top panel). This is because the Euclidean distance is a double-zero symmetrical coefficient. For the same reason, the confidence intervals are much wider for the Euclidean distance than for any other coefficient. At sites with the same environmental conditions, one should expect to find the same species, but species abundances usually vary. Although these variations in abundance may have important

implications when species are rare, they should have only negligible effect on the results when species are common. In that instance, the Euclidean distance considers common and rare species similarly. The results associated with the Euclidean distance suggest that double-zero symmetrical coefficients should only be used to address ecological questions where double-zeros are ecologically meaningful, as suggested by Anderson et al. (2011).

Ecologists should also be careful in using the distance between species profiles, especially in the presence of many common species, because it seems to loose explanatory power in these circumstances (Figure 1h-j, w-y). Legendre and De Cáceres (2013) have shown that the distance between species profiles lacks some of the important properties necessary for coefficients that are used to assess beta diversity. For this reason, the distance between species profiles suffers from the same problem as the Euclidean distance in the presence of common species, but to a lesser extent.

When comparing dissimilarity coefficients with simulated presence-absence data, the R^2 coefficients are very similar between coefficients across the different SADs (Figure 3). Results for the Raup-Crick coefficient were the only exception, although its confidence intervals still overlap importantly with the others. It yields a somewhat lower R^2 when there are many common species (Figure 3, central pannel). Because a high R^2 for the Raup-Crick coefficient is generally associated with a high R^2 of the other coefficients, it may be that the Raup-Crick coefficient does not as effectively capture patterns as the other coefficients when many common species are sampled (Figure 1 h-j, y). These results are consistent with Legendre and Legendre (2012, Subsection 7.3.5) who showed that the statistical power of the Raup-Crick coefficient to detect significant association between pairs of sites is low even when McCoy's et al. (1986) permutation procedure is used.

We were surprised that the simple-matching coefficient produced results equivalent to other coefficients (Figure 3, central pannel). We expected it to be burdened by the same problems as the Euclidean distance because the simple-matching coefficient is the presence-absence equivalent of the Euclidean distance, making it a double-zero symmetrical coefficient. However, it seems that when abundances are considered, the importance of double-zeros increases. If a single species is sampled in large abundances at two sites, the Euclidean distance between these sites for that particular species is likely to be somewhat far from 0 even though it is clear that these sites are quite similar. The Euclidean distance thus over-emphasizes the differences between two sites where a species is found in large but unequal abundances, a problem that does not exist for the simple-matching coefficient because the species will be recorded as present (or 1) for both sites, yielding a distance of exactly 0.

Another aspect of our simulations is the increase in explained variance with the number of common species (progression of R^2 from SAD a to j in Figure 1). This trend is consistent for all coefficients compared (with the exception of the Euclidean, species profile, and Raup-Crick coefficients, discussed above), in abundance and presence-absence data alike, although it is weaker for presence-absence data (Figures B5-B8). Similar conclusions were found with communities simulated with larger error (Figures B1-B8).

A NEW WAY TO PERFORM CANONICAL ORDINATIONS

The previous simulations have shown that within data types, double-zero asymmetrical coefficients yield similar value of R^2, for each SAD compared (Figures 2 and 3 and Figures B1-B8). This is shown by the substantial overlap between confidence intervals of all coefficients calculated for any particular SAD. Each coefficient has particularities making it more appropriate for specific ecological situations or research questions, and less so for others. With

the wealth of coefficients available in the ecological literature, it is common for more than one coefficient to be appropriate for a particular ecological study. In that context, the question "Which dissimilarity coefficient should be used?" remains incompletely answered.

Here, we propose a three-step procedure to handle this problem. Even though most of the information highlighted by the different coefficients is often quite similar, the mathematical properties of each coefficient emphasize certain characteristic in the data that other coefficients do not, and vice versa.

• In that respect, the first step is to compare coefficients and evaluate how much the information they explain diverges. This is accomplished by comparing all aspects of the canonical ordination models (i.e., the sites, the species, and the canonical coefficients), not only the variance explained.

• Secondly, a selection may be carried out among the coefficients, if necessary. The RDA models constructed using coefficients that differ markedly from the others should be considered separately, or their use should be revaluated. The differences between RDA models can be in the ordination of the sites, the site-species relationships, and/or the relationships between canonical coefficients and the sites and species. In a nutshell, potential differences among RDA models should be sought in all aspect of the models. Comparison and selection of coefficients is recommended because if a coefficient is markedly different from the others, its inclusion in the consensus step (next) may blur ecological relationships that could appear if this coefficient was removed.

• Thirdly, the information common to RDA models that only differ by their dissimilarity coefficients should be synthesized. It is important to focus only on the information shared by the different RDA models to ensure that no misguided ecological interpretations are made. Because

it is difficult to extract common information by an examination of independent canonical ordination triplots, we propose a new method that computes a consensus among canonical ordinations that differ only by the coefficients used to construct them. The consensus focuses on the patterns found by all RDA models, leaving out the information extracted by only one or a few coefficients. We call this new approach "consensus RDA". A detailed explanation of how these three steps are carried out is presented below.

Comparison of RDA models.—To compare RDA models that only differ by their coefficients, the first step is to isolate the significant components found in each **Z** matrix (site scores calculated using the explanatory variables, Equation 3.1), e.g. the axes with a P-value ≤ 0.05. Model comparisons rely on the **Z** matrices, which contain the canonical ordination coordinates of the sites; the variance of each canonical axis in **Z** is equal to its associated eigenvalue when the distances among sites are preserved in the ordination results (RDA scaling 1). In the RDA framework, the canonical eigenvalues measure the variance explained by the canonical axes.

We correlated the significant canonical axes of the **Z** matrix obtained for each dissimilarity index to those obtained with the other indices using RV coefficients (Escoufier 1973, Robert and Escoufier 1976). The RV coefficient is a multivariate generalization of the Pearson correlation that correlates two matrices with corresponding rows (sites). It produces values that range between 0 (no correlation) and 1 (perfect correlation). The RV coefficients for all pairs of dissimilarity indices were assembled in a matrix of pairwise RV coefficients. Using this matrix, we drew a minimum spanning tree (MST, Legendre and Legendre 2012, Section 8.2) to compare dissimilarity indices. This required the matrix of RV coefficients to be transformed into a dissimilarity matrix. We used (1 – RV) to perform the transformation because it ensured

that the correlation information brought by the RV coefficients was conserved. These dissimilarities ranged from 0 to 1.

Selection of RDA models.—After examination of the MST, a selection of concordant dissimilarity indices can be made. We leave it at the discretion of users to decide how dissimilarity indices should be selected. For example, the dissimilarity indices linked by the longest MST branches can be removed if these branches are much longer than the average branch. If the longest branch in the MST links two groups of dissimilarity indices, it may be interesting to calculate two consensus RDAs, one for each group of indices.

Consensus RDA.—To calculate a consensus RDA, the significant components of the **Z** matrices selected to compare RDA models are used again (Figure 4b). Of course, only the **Z** matrices from coefficients that have been selected in the previous step should be considered. In consensus RDA, all significant components are grouped in a large matrix (Figure 4c). Using this large matrix as a response in an RDA where the matrix of explanatory variables is **X** (Figure 4c), compute the consensus RDA site scores $\mathbf{Z}^*$ and the consensus RDA canonical coefficients $\mathbf{C}^*$. This RDA also yields eigenvalues, which express the amount of variance represented by each $\mathbf{Z}^*$ component, and more generally by each axis of the consensus RDA. These eigenvalues can be used to measure the strength of the consensus. The consensus RDA species scores $\mathbf{U}^*$ need to be calculated following Equation 5. In other words, $\mathbf{U}^*$ is obtained following the same procedure as in db-RDA.

When performing an RDA, the results can be presented in either a distance (scaling 1) or a correlation (scaling 2) triplot. Scaling can also be used in consensus RDA. All the calculations presented above are carried out using the scaling 1 matrices **Z** because, as explained in the subsection *Comparison of RDA models*, the consensus method relies on a property of **Z** that is

only present in scaling 1. To obtain a consensus result in scaling 2, the consensus site scores (matrices $\mathbf{Z}^*$) need to be rescaled following $\mathbf{Z}^*(\mathbf{\Lambda}^*)^{-1/2}$. A similar procedure is used to apply scaling 2 on the species scores consensus ($\mathbf{U}^*(\mathbf{\Lambda}^*)^{-1/2}$).

An interesting aspect of this new method is that as long as the dissimilarity indices represent the only aspect that differs between the different RDAs, a consensus RDA can be computed. This also includes partial RDAs.

All calculations necessary to obtain a consensus RDA rely on the **Z** matrices of the different RDA models constructed with a group of relevant dissimilarity indices. The components in these **Z** matrices contain the fitted site scores for the RDAs; they do not include the residuals components of **Y**, which are part of the **F** matrices (ter Braak 1994; Legendre and Legendre 2012, Subsection 11.1.3). Because it is often more interesting to study an RDA triplot in a projection where residuals are not included, a consensus of **F** matrices was not incorporated in our description of consensus RDA.

The explanations of how to perform consensus RDA indicate that any number of axes can be used for any of the RDAs that are considered in the calculation of the consensus. However, it is not clear whether all, or only the significant canonical axes, should be used in a consensus RDA to obtain the model that best explains the community data. To evaluate which approach should be used, the simulated site-by-species tables presented in section *Simulating communities with varying species abundances* were used. Each site-by-species table was correlated with $\mathbf{Z}^*$ (consensus site scores), which was calculated using all canonical axes, using the RV coefficient. We then compared these RV coefficients with RV coefficients correlating the site-by-species tables with the consensus site scores calculated using only the significant axes. The comparisons

were carried out using both abundance and presence-absence simulated data. All dissimilarity coefficients discussed in this paper were used in the construction of the consensus site scores.

The results in Figure 5 were obtained using abundance data where the error was the largest (ε in Equation 6 followed a Normal distribution with a mean = 0 and a standard deviation = 2), which yielded the largest variation in the comparisons made. In Figure 5 (note the narrow range of the ordinate), the differences between the RV coefficients calculated using all canonical axes and the RV coefficients computed using only significant axes ranges almost always between 0.05 and -0.02. Although, for certain extreme cases, slightly more information can be obtained using all canonical axes, in the majority of situations very little information is gained (or sometimes lost) from using all canonical axes instead of only the significant ones. Results from the simulations where communities were generated with smaller error terms are presented in Appendix C. In these simulations, presence-absence and abundance data were considered. For abundance data, the results yield the same conclusions as the one presented in Figure 5. For presence-absence data, it is slightly better to use all canonical axes; however the information gain is minimal. In doubtful cases, the best solution is found by comparing a consensus RDA obtained using all canonical axes with a consensus RDA constructed with only the significant axes and choosing the solution that yields the largest RV coefficient. This approach ensures that the result of the consensus RDA always represents the largest amount of information from the community data.

A comparison of dissimilarity indices and a consensus RDA are presented in the *Ecological illustration* section, for abundance and presence-absence data.

SHOULD WE USE PRESENCE-ABSENCE DATA?

Modelling presence-absence data is more challenging than abundance data because information on species abundances is missing. The results of our simulations confirm this statement; the R^2 values are consistently higher for abundance (Figures 2, B1-B4) than for presence-absence data (Figures 3, B5-B8). This result is not surprising because one would expect to obtain better species-environment linear models when using more informative data. This finding remains the same irrespective of the error level in the data (Figures 2, B1-B9). However, comparison between presence-absence and abundance data using R^2 does not reflect how well the true species structure is modelled. To compare the canonical ordination results of abundance and presence-absence data, we first need to measure how much information from the true species (Equation 6 without the error term) structure is extracted by the canonical analyses. As explained at the end of section *RDA and dissimilarity coefficients*, the Euclidean and simple matching coefficients are double-zero symmetrical; they are designed to answer ecological questions where double-zeros are ecologically meaningful. In our simulations, double-zeros do not necessarily reflect a strong similarity between sites. For this reason, double-zero symmetrical coefficients were not included in the comparison between abundance and presence-absence data. For both data types, we calculated RV coefficients between the true species structure (Equation 6 without the error term) and the significant canonical axes.

We regrouped all RV coefficient results within data type and compared the grouped abundance to the grouped presence-absence results (Figure 6). According to the results obtained by comparing dissimilarity coefficients within data type (Figures 2, 3, B1-B9), it is valid to group dissimilarity coefficients used on the same data type because no dissimilarity coefficient dominates over the others for any SAD. Figure 6 illustrates the grouped results for simulations

where the error is the smallest (standard deviation = 0.001). What is striking about these results is that when there are many common species (Figure 1, i-j, y), the amount of information extracted by canonical ordinations is much less for presence-absence than for abundance data. These conclusions can be extended to situations where there are at least as many common as there are rare species (Figure 6, g, h, l, t) because the overlap between confidence intervals is small in these situations. This suggests that for communities with at least as many common as rare species, the information lost in canonical ordinations on occurrences should not be interpreted in the same way as results obtained from canonical ordinations on abundance data. Similar results were obtained for data simulated with larger errors (Figures D1-D4). We will show in the *Ecological illustration* section how these findings apply to real ecological data.

Ecological illustration: Carabidae of northwestern Alberta

To show how the previous findings may be applied in real ecological situations, we extend the analysis to a data set about ground beetles (Carabidae) sampled at 192 sites in a never-harvested mature boreal mixedwood forest (see Bergeron et al. 2011, Blanchet et al. 2013). In this illustration, we aim at finding how trees influence the ground beetle community in the boreal forest. This question has already been approached with the same data by Bergeron et al. (2011). The difference here is that we used consensus RDA based on several dissimilarity indices detailed 3 paragraphs below, for abundance and presence-absence data. Bergeron et al. (2011) performed all their analyses using a single dissimilarity calculated on abundance data.

The sites, which formed a near-regular grid in an area of 70 km^2, were located in the Ecosystem Management Emulating Natural Disturbances (EMEND) experimental area in northwestern Alberta, Canada. Each site contained three pitfall traps (Spence and Niemelä 1994) located on the perimeter of a 15 m radius circle. From the center of the circle, a trap points due

north while the other two are separated by 120 degrees. The community data are composed of 37 ground beetle species sampled throughout the summer of 2003. Beetle abundances were divided by the number of days each trap was active to remove the effect of trap disturbance and of non-demonic intrusions (Hurlbert 1984). Presence-absence data for each site were obtained by transforming all abundances larger than 0 to 1.

As explanatory variables, the relative basal areas of the 25 trees closest to the centre of each site were used. Eight tree species were present in the experimental area and the relative basal area of each species was used as an explanatory variable. Further analysis of this data set may be found in Blanchet et al. (2013) and in Bergeron et al. (2011, 2012). The Hellinger distance was used by Blanchet et al. (2013) and Bergeron et al. (2012), and the percentage difference distance was employed by Bergeron et al. (2011). Note that Bergeron et al. (2011) used non-metric multidimentional scaling (Legendre and Legendre 2011, Section 9.4) to study carabids, unlike Blanchet et al. (2013) and Bergeron et al. (2012) who used tb-RDAs.

In this ecological illustration, we compare canonical ordinations calculated on abundance and presence-absence data, considering results from all dissimilarity coefficients used in our simulations, with the exception of the double-zero symmetrical coefficients. We did not use double-zero symmetrical coefficients because they consider double-zeros (the absence of a species at two sites) as informative, which may lead to wrongful interpretations. The carabid dataset used in this illustration was sampled to study how habitat variation influenced the ground beetle community. Blanchet et al. (2013) have shown that this community is mostly unaffected by anthropogenic disturbances. In this context, Anderson et al. (2011) explained that double-zeros are not necessarily ecologically meaningful, making the use of double-zero symmetrical coefficients inappropriate for studying this particular carabid community.

An RV comparison of the RDA models constructed with different dissimilarity coefficients is presented using MSTs in Figure 7b for abundance data and Figure 7e for presence-absence data. Each MST was constructed from a dissimilarity matrix of RV coefficients correlating all pairs of RDA models obtained from the different coefficients following the procedure presented in the section *A new way to perform canonical ordinations*. As a reference, we included in Table 3 the amount of variance explained (R^2) by the full db-RDA or tb-RDA models based upon the different dissimilarity coefficients. We used the full RDA models because the final consensus RDA results were more informative than when only the significant axes were used. This was true for abundance and presence-absence data.

We found that for both abundance and presence-absence data, the RDA model constructed using the χ^2 distance was the most different from the others (Figure 7b, e). This is likely due to *Notiophilus directus*, a rare species found with low abundance at three sites where *Pterostichus punctatissimus* and *Miscodera arctica* (one site), or only *Pterostichus punctatissimus* (two sites), were encountered. Legendre and Legendre (2012, Subsection 7.4.1) and Greenacre (2013) explained that the χ^2 distance gives higher weights to species represented by only a few individuals at sites where only a few other species are found. Legendre and De Cáceres (2013) also found that the χ^2 distance lacked an important property for analysis of community composition data. Because we did not want to give undue importance to rare species, we did not further consider the χ^2 distance in the analyses of this carabid community.

Using the remaining coefficients, we constructed a consensus RDA. We plotted as many species as we could in the consensus RDA triplots without loosing overall interpretability. The species not presented on the diagrams were consistently near the centre of the consensus triplots, which made it impossible to interpret the ecological relationships of these species with respect to

596 the tree basal areas. The first two axes of the consensus RDA represent 88.4% of the variance for
597 abundance data and 85.2% for presence-absence data, and thus represent well the information in
598 the different RDA models. The third consensus axes explained less than 7% of the variance, for
599 abundance as well as presence-absence data, making the information presented by any
600 subsequent axes too small to justify their use. Note that the variance in consensus RDA
601 represents the strength of the consensus, not the strength of the model, as it is the case for
602 traditional RDA.

603 Although the amount of information explained by the first two axes of the consensus
604 RDAs based on abundance (Figure 7c) and presence-absence data (Figure 7f) is similar, the
605 underlying information is different. For example, segregation of beetle species niches between
606 coniferous (Ab, Ll, Pc, Pg, and Pm) and deciduous forest (Bp, Pt, and Pb) on the positive side of
607 the abscissa is better achieved using the consensus RDA based on abundance data. Because these
608 beetle and tree species are all characteristic of upland mixedwood forest (Bergeron et al. 2011),
609 the comparion between abundance (Figure 7c) and presence-absence (Figure 7f) consensus
610 triplots suggest that beetle species occur all along the deciduous-coniferous forest gradient but it
611 is their abundance that varies according to habitat. Also, relationships between beetle and tree
612 species were not always consistent between the two consensus ordinations. For example,
613 *Calathus advena* (Calaadve) is more closely related to *Picea glauca* (Pg) in the abundance
614 ordination (Figure 7f) than it is in the presence-absence ordination (Figure 7c).

615 The result about *Agonum gratiosum* (Agongrat), *Carabus chamissonis* (Caracham),
616 *Platynus mannerheimii* (Platmann), and *Pterostichus brevicornis* (Pterbrev) are impossible to
617 interpret in the consensus RDA computed using species abundance because they are too close to
618 the triplot centre. However, in the presence-absence consensus triplot, these species are

interpretable. This may relate to the fact that presence-absence data give more weight to less abundant species than relative abundance data (Anderson et al. 2011, 2006). In this beetle assemblage, *Platynus mannerheimii* (Platmann) is of special ecological interest even if it is not a common species because it has a narrow habitat requirement that is locally restricted to old wet and productive forest (Bergeron 2011). This species, as well as *A. gratiosum*, are only found at sites dominated by *Picea mariana* (Pm) and *Larix laricina* (Ll) even if their abundance at these sites is not as high as that of the more common species. In that instance, it makes sense that the consensus RDA on presence-absence data makes these two species stand out. Such result shows that performing community analyses on both abundance and presence-absence data concurrently makes it possible to extract interesting information out of the data.

The SAD (Figure 7a) of this beetle community depicts many rare and many common species, which is typical for carabid communities (Niemelä 1993). The species presence distribution that describes species occurrence for these data highlights more sharply the two groups of species in the carabid data (Figure 7d). Our simulations suggest that a community composed of many rare and many abundant species (Figure 1, l and t) does not preserve well community patterns after having been transformed into presence-absence (Figure 6, l and t). This is reflected in the ecological analysis where abundance-based ordination achieves a better segregation of the beetle ecological niches. Although this may suggest that presence-absence ordinations are not useful on their own, differences between the abundance and the presence-absence ordinations may have an ecological foundation. It may be that differences between ordinations based on abundance and presence-absence data reflect the spatial aggregation of carabid species. It is also possible that the consensus RDA calculated on abundance data brings complementary information to the consensus RDA result obtained from presence-absence data.

642 To know if the differences between the two consensus RDAs is determined by ecological
643 processes, a detailed study of this carabid community needs to be carried out contrasting
644 presence-absence and abundance data at multiple scales using other variables characterizing the
645 habitat of Carabidae in addition to the tree basal areas.

646 In this carabid example, consensus RDA gives strong confidence in the ecological
647 associations discovered between ground beetles and trees. Here are a few examples of
648 discoveries made by Bergeron et al. (2012) that hold true in consensus RDA (Figure 7c).
649 *Agonum retractum* (Agonretr) and *Platynus decentis* (Platdece) prefer forest containing *Populus*
650 *balsamifera* (Pb), *Populus tremuloides* (Pt), and to a lesser extent to *Betula papyrifera* (Bp),
651 which are all upland deciduous trees. *Stereocerus haematopus* (Sterhaem) and *Calathus advena*
652 (Calaadve) were commonly found in coniferous forest dominated by *Picea glauca* (Pg) and
653 *Abies balsamea* (Ab). *Calathus ingratus* (Calaingr) and *Pterostichus adstrictus* (Pteadst)
654 typically occurred in both deciduous and coniferous upland forest where *Picea mariana* (Pm)
655 and *Larix laricina* (Ll) are usually absent. *Pterostichus punctatissimus* (Pterpunc) and *Picea*
656 *mariana* (Pm) present a strong ecological association. Because consensus RDA is based on many
657 dissimilarity coefficients, the ecological associations discovered between trees and ground beetle
658 species of the mixedwood boreal forest should be ecologically more meaningful and reliable and
659 not biased by the dissimilarity coefficient used in the calculation of the consensus ordination.

660 It is not the goal of this paper to present a detailed ecological study of northwestern
661 Alberta boreal carabids. However, by comparing the consensus RDA calculated on the carabid
662 abundance data (Figure 7c) with the canonical ordination results from Bergeron et al. (2012,
663 Figure 4) who also studied the relationship between carabids and tree relative basal area with the
664 same data using RDA, differences can be found that are solely attributed to the dissimilarity

665 coefficient used. For example, in our result, *Stereocerus haematopus* (Sterhaem) is more closely
666 related to *Picea glauca* (Pg) than it is in Bergeron et al. (2012). These authors based the
667 canonical ordination on the Hellinger transformation of the beetle abundance data, which
668 emphasizes the composition nature of the data rather than raw abundance (Anderson et al. 2011).
669 The consensus RDA of Figure 7c, which uses a variety of coefficients along the composition-
670 abundance gradient, indicates that the abundance pattern of *Stereocerus haematopus* is more
671 closely associated with white spruce than previously discovered by Bergeron et al. (2012). To
672 prevent a biased interpretation resulting from the use of a specific dissimilarity coefficient, as
673 was the case for *S*. *haematopus*, consensus RDA is a better option.

674 DISCUSSION

675 This paper presents a new approach to perform canonical ordination using a group of
676 dissimilarity coefficients and proposes a new framework to analyze species communities using
677 abundance and presence-absence data together.

678 A surprising result of this study is that the SAD of a community is not an important
679 criterion for choosing a coefficient (Figures 2, 3 and 6, B1-B8 and D1-D4) in canonical
680 ordinations. This is what prompted us to develop consensus RDA. These results may also bring
681 insight into the comparison of SADs, an important line of research (McGill et al. 2007). Using
682 the result in Figure 6 (and also Figures D1-D4) obtained from abundance data, we can compare
683 SADs because the communities simulated with different SADs were correlated with the same
684 true underlying structure of the data (Equation 6 without the error term). The true underlying
685 structure of the data serves as a reference to know how well a SAD determines the raw
686 community structure because it is the basic information from which all species are constructed in
687 the simulation. From the discussion in McGill et al. (2007) on SAD comparison, it can be

688 expected that SADs defining notably different abundance patterns (e.g. Figure 1b, l, m, q, and u)
689 would correlate differently with the true underlying structure of the data. However, in Figure 6
690 they all correlate equally well with the true underlying structure of the simulated communities.
691 Moreover, the fairly broad range of the RV coefficient 95% confidence intervals for any one of
692 the 25 SADs indicates that the variations in the raw multivariate community data can be
693 surprisingly important even if species have the same abundance structure. Such results may
694 suggest that the SAD of a community may present only a small fraction of the information that
695 characterizes a community matrix. However, further research is still needed to confirm the
696 findings we made that the information lost when constructing SADs may make it difficult to
697 develop a valuable approach to compare communities using SADs.

698 Our study shows that the choice of dissimilarity coefficients in canonical ordinations
699 should primarily be based on the ecological knowledge available for the community under study.
700 The ecological questions and the data type should guide the choice of one or a group of
701 coefficients. Legendre and Legendre (2012, Table 7.4) offer a decision key designed to help
702 ecologists select coefficients for community composition data based on data types (presence-
703 absence or abundance) and type of information to be extracted. If a canonical analysis is
704 performed using only one coefficient when more than one can potentially be used, Legendre and
705 Gallagher (2001) would select the coefficient that explains the largest amount of variance.
706 However, the properties of the selected coefficient may influence the interpretation.

707 If more than one coefficient is chosen, it is important to compare them using an MST
708 based on dissimilarities of pairwise RV coefficients to determine if any of them presents results
709 markedly different from the others. This comparison can be seen as a selection procedure for
710 coefficients. It evaluates the similarities between different RDA models where coefficients are

the only element differentiating the models and finds which model(s) differ(s) notably from the others. This comparison will help decide if any coefficient(s) should be discarded. Comparing RDA models constructed using different coefficients through an MST is a first step in the analysis of a community through more than one coefficient.

When more than one coefficient presents similar information, a consensus RDA allows one to extract the most information out of the data because it focuses on the common information brought out by different coefficients. Using only one coefficient may put too much emphasis on a particular aspect of the data because each coefficient was designed to highlight different particularities of a community matrix. This may lead to a suboptimal ecological interpretation. Consensus RDA prevents this problem from occurring by extracting only the common information generated by a group of coefficients. In that respect, consensus RDA indirectly solves the technical problem of choosing a coefficient by using all the ones that can be suitable to analyse the data. Also, because it diminishes the importance of the information highlighted by one or a few coefficients, it produces a result less influenced by the mathematical properties of a coefficient. For this reason, consensus RDA gives a more accurate representation of a community and will help researchers better understand the factors structuring the species in the community they study.

Conceptually, the new canonical ordination procedure proposed in this paper has similarities with model averaging (Anderson et al. 2008, Chapter 5; Burnham and Anderson 2004). In model averaging, the best models are given more weight than the poor ones. This can be related to the selection procedure we propose where coefficients are considered independently, discarded, or used to construct a consensus model. In consensus RDA, the different RDA models are weighted by the sum of the canonical eigenvalues of their components

included in the construction of the consensus. In that respect, models constructed with different dissimilarity coefficients have different weights, which will influence the consensus; this is another similarity with model averaging. However, model averaging is more flexible because the choice of variables may vary between models, whereas only dissimilarity coefficients vary in consensus RDA.

As Økland (1996) pointed out, unconstrained ordination is a useful method to generate hypotheses when no explanation of the community variation has been proposed, whereas the main purpose of constrained ordination is hypothesis testing. With the development of constrained ordination methods, more complex analyses have been proposed and used by ecologists. For example, one may test a hypothesis by RDA using a set of explanatory variables or experimental factors, then examine the PCA ordination of the non-canonical variation to generate new hypotheses about the origin of the residual variation not explained by the explanatory variables.

Using a different approach, Borcard and Legendre (1994) showed that spatially constrained ordination of community composition data can help ecologists generate hypotheses about the processes that produced the spatial variation of the community. In their 1994 paper, they used a polynomial of the geographic coordinates as constraining factor in CCA. In subsequent papers, they developed spatial eigenfunction analysis based on Moran's Eigenvector Maps (MEM, originally called PCNM, Borcard and Legendre 2002, Borcard et al. 2004, Dray et al. 2006); this is a much more powerful method for modeling fine-scaled spatial variation. Because ecological data are often spatially correlated (Legendre 1993), inclusion of spatial variables such as MEMs in canonical ordination is important to understand and test the

significance of species-environment relationships. Dray et al. (2012) reviewed different ways of considering space in community ecology and including it in canonical ordinations.

Another aspect that researchers need to consider when performing canonical ordinations such as CCA, RDA, or consensus RDA is that these methods compute a linear model of the explanatory variables for each species in the community. In the case of CCA, the species data are chi-square transformed before computing multiple regression and the regression involves the total abundances of the sites as weights (Legendre and Legendre 2012, Section 11.2). Qualitative explanatory variables (factors) can be included in these models, as it is also the case in multiple regression. Because the species-environment relationships in nature are not necessarily linear, it has been proposed to include polynomials of the explanatory variables in the explanatory matrix, instead of the explanatory variables only, to make it possible to model non-linear relationships between the species and the explanatory variables (e.g. Legendre and Legendre 2012, Ecological application 14.1b). Note that dissimilarity coefficients and data transformations do not account by themselves for the non-linearity of the species-environment relationship. They were design to give more (or less) weight to common (or rare) species and to account for the double-zero problem. This approach can be applied to all canonical ordination methods, including consensus RDA.

A problem that we have not approached but warrants further investigation is selection of explanatory variables in consensus RDA. Methods such as forward selection (e.g., Blanchet et al. 2008) assume that an RDA is performed using only one dissimilarity coefficient. Consensus RDA requires all explanatory variables to be the same and that only the coefficient differs between RDAs. If an automatic variable selection procedure is used independently for each RDA, it is likely that different sets of variables will be selected. In this situation, we propose

three variable selection aproaches. (1) A consensus analysis should employ the union of all explanatory variables selected for the various coefficients. That is, if for a coefficient, explanatory variables A and B are selected and with another coefficient it is explanatory variables A and C that are chosen, the union of the explanatory variables for the consensus RDA would be variables A, B, and C. Using this approach, one can at least eliminate the explanatory variables that are totally useless. This idea of using the union of the selected variables is inspired by the selection method of Peres-Neto and Legendre (2010) for Moran's eigenvector maps eigenfunctions. (2) The variable selection is carried out on the consensus RDA result without any variable selection carried out on individual RDAs. (3) Use the union of the selected variables on individual RDAs, as explained in (1), and then carry out a further selection for the consensus RDA. Further studies will need to be carried out to evaluate which of these three approaches yields the models that best define a species community.

Species abundance data contain more information than presence-absence data and often lead to a better understanding of community variations through RDA, although community ecologists generally consider that the single most important information about a species is its presence. However, for certain organisms, abundance data are not reliable. In palynology for example, presence-absence data are often favoured because abundance data are subject to large bias (Davis 2000). Presence-absence data are also more suitable when studying ant communities using pitfall traps because their social behaviour and propensity at creating foraging trails has an enormous influence on abundance data (Higgins and Lindgren 2012). Similarly, in studies of fish biodiversity, variation in size of fish species living in the same area demands that different instruments be used to catch them, and thus the abundance data are not comparable. The only way to consider all species of fish together in a consistent analysis is by using presence-absence

802 data (biomass data can also be used for fish of all sizes caught by electrofishing or recorded
803 during underwater visual census). This is likely to be true for any communities where variations
804 in size between species require that different trapping methods be used to catch enough species
805 to have a representative fraction of the studied species community.

806 When working with presence-absence data, we suggest that one should first draw a
807 species presence distribution, as we did in Figure 7d. The ratio between common and rare
808 species should serve as a general guideline when drawing ecological conclusions. Although it is
809 possible that canonical ordinations performed on presence-absence data show biased results, it is
810 more likely that such ordinations can be complementary to those computed for abundance data.
811 Certain environmental factors may be necessary for a species to occur in an area (e.g., certain
812 plant species are found only in the presence of certain geological formations) while other factors
813 may make species abundances vary (e.g., precipitation). Variation in abundance is efficient in
814 describing how a species is related to a gradient (environmental, physical, or other). However
815 species abundances may conceal the strict relationship a species has with its habitat. This strict
816 relationship is what makes a species occurs or not at a site. In that respect, considering both
817 abundance and presence-absence data may be ecologically valuable to better understand the
818 factors structuring a community. The idea to use both abundance and presence-absence data to
819 better understand an ecological system has been proposed before (see e.g. Van Buskirk 2005).
820 As explained in the previous paragraphs, abundance data may be unreliable when sampling
821 certain groups of organisms. However, for all communities where species abundances can be
822 sampled without diminishing the value of the data, presence-absence data can be easily obtained
823 by transforming all abundances larger than 0 to 1s, allowing ecologists to get a more complete
824 understanding of the data they collected.

In this paper we presented a new approach to perform canonical ordination in community ecology research. This approach has the potential to be used in other fields of research where the structure of the data is similar to that of community ecology. Population and landscape genetics are examples of research areas where consensus RDA could potentially be useful.

ACKNOWLEDGEMENTS

We are greatful to Xianli Wang, John R. Spence, and Dave W. Roberts for insightful comments on an early draft of the manuscipt. This research was supported by GEOIDE Canada and an NSERC to F. He and NSERC grant number 7738 to P. Legendre.

LITERATURE CITED

Anderson, D. R. 2008. Model based inference in the life sciences – A primer on evidence. Springer, New York.

Anderson, M. J. 2006. Distance-based tests for homogeneity of multivariate dispersions. Biometrics **62**:245–253.

Anderson, M. J., T. O. Crist, J. M. Chase, M. Vellend, B. D. Inouye, A. L. Freestone, N. J. Sanders, H. V. Cornell, L. S. Comita, K. F. Davies, S. P. Harrison, N. J. B. Kraft, J. C. Stegen, and N. G. Swenson. 2011. Navigating the multiple meanings of beta diversity: a roadmap for the practicing ecologist. Ecology Letters **14**:19–28.

Bergeron J. A. C., J. R. Spence, and W. J. A. Volney. 2011. Landscape patterns of species-level associations between ground-beetles (Coleoptera: Carabidae) and overstory trees in boreal forests of western Canada (Coleoptera: Carabidae). In Erwin, TL (Ed), *Proceedings of a Symposium honoring the careers of Ross and Joyce Bell and their contributions to scientific work*, Burlington, VT, 12-15 June 2010. ZooKeys 147: 577-600.

847 Bergeron J. A. C., F. G. Blanchet, J. R. Spence, and W. J. A. Volney. 2012. Ecosystem
848 classification and inventory maps as surrogates for ground beetle assemblages in boreal
849 forest. Journal of Plant Ecology **5**:97–108.

850 Blanchet, F. G., J. A. C. Bergeron, J. R. Spence, and F. He. 2013. Landscape effects of
851 disturbance, habitat heterogeneity and spatial autocorrelation for a ground beetle
852 (Carabidae) assemblage in mature boreal forest. Ecography **36**:636–647.

853 Borcard, D., and P. Legendre. 1994. Environmental control and spatial structure in ecological
854 communities: an example using Oribatid mites (*Acari*, *Oribatei*). Environmental and
855 Ecological Statistics **1**:37–61.

856 Borcard, D., and P. Legendre. 2002. All-scale spatial analysis of ecological data by means of
857 principal coordinates of neighbour matrices. Ecological Modelling **153**:51–68.

858 Borcard, D., P. Legendre, C. Avois-Jacquet, and H. Tuomisto. 2004. Dissecting the spatial
859 structure of ecological data at multiple scales. Ecology **85**:1826–1832

860 Borcard, D., P. Legendre, and P. Drapeau. 1992. Partialling out the Spatial Component of
861 Ecological Variation. Ecology **73**:1045–1055.

862 Bray, J. R., and J. T. Curtis. 1957. An Ordination of the upland forest communities of southern
863 Wisconsin. Ecological Monographs **27**:325–349.

864 Burnham, K. P., and D. R. Anderson. 2004. Multimodel inference – understanding AIC and BIC
865 in model selection. Sociological Methods and Research **33**:261–304.

866 Cavalli-Sforza, L. L., and A. W. F. Edwards. 1967. Phylogenetic analysis - models and
867 estimation procedure. Evolution **21**:550–570.

868 Clarke, K. R., and R. H. Green. 1988. Statistical design and analysis for a biological effects
869 study. Marine Ecology-Progress Series **46**:213–226.

870 Davis, M. B. 2000. Palynology after Y2K – understanding the source area of pollen in sediments.
871 Annual Review of Earth and Planetary Sciences **28**:1–18.

872 Dewdney, A. K. 2000. A dynamical model of communities and a new species-abundance
873 distribution. Biological Bulletin **198**:152–165.

874 Dray, S., D. Chessel, and J. Thioulouse. 2003. Co-inertia analysis and the linking of ecological
875 data tables. Ecology **84**:3078–3089.

876 Dray, S., and A.-B. Dufour. 2007. The ade4 package: Implementing the duality diagram for
877 ecologists. Journal of Statistical Software **22**:1–20.

878 Dray S., P. Couteron, M.-J. Fortin, P. Legendre, P. R. Peres-Neto, E. Bellier, R. Bivand, F. G.
879 Blanchet, M. De Cáceres, A.-B. Dufour, E. Heegaard, T. Jombart, F. Munoz, J. Oksanen,
880 R. Pélissier, J. Thioulouse, and H. Wagner. 2012. Community ecology in the age of
881 multivariate multiscale spatial analysis. Ecological Monographs **82**: 257–275.

882 Dray, S., P. Legendre, and P. Peres-Neto. 2006. Spatial modelling: a comprehensive framework
883 for principal coordinate analysis of neighbor matrices (PCNM). Ecological Modelling
884 **196**:483–493.

885 Escoufier, Y. 1973. Le traitement des variables vectorielles. Biometrics **29**:751–760.

886 Fisher, R. A., A. S. Corbet, and C. B. William. 1943. The relation between the number of species
887 and the number of individuals in a random sample of an animal population. Journal of
888 Animal Ecology. **12**:42–58.

889 Gaston, K. J. 2010. Valuing common species. Science **327**:154-155.

890 Gower, J. C. 1966. Some distance properties of latent root and vector methods used in
891 multivariate analysis. Biometrika **53**:325–338.

892 Gower, J. C., and P. Legendre. 1986. Metric and Euclidean properties of dissimilarity
893 coefficients. Journal of Classification **3**:5–48.

894 Gray, J. S., A. Bjorgesaeter, and K. I. Ugland. 2006. On plotting species abundance distributions.
895 Journal of Animal Ecology **75**:752–756.

896 Greenacre, M. 2013. The contributions of rare objects in correspondence analysis. Ecology
897 **94**:241–249.

898 Higgins, R. J., and B. S. Lindgren. 2012. An evaluation of methods for sampling ants
899 (Hymenoptera: Formicidae) in British Columbia, Canada. The Canadian Entomologist
900 **144**: 491-507.

901 Hill, M. O., H. G. Gauch Jr. 1980. Detrended correspondence analysis, an improved ordination
902 technique. Vegetatio **42**: 47–58

903 Hurlbert, S. H. 1984. Pseudoreplication and the Design of Ecological Field Experiments.
904 Ecological Monographs **54**:187–211.

905 Hutchinson, G. 1957. Concluding remarks. Cold Spring Harbor Symposium on Quantitative
906 Biology **22**:415–427.

907 Jaccard, P. 1901. Étude comparative de la distribution florale dans une portion des Alpes et du
908 Jura. Bulletin de la Société Vaudoise des Sciences naturelles **37**:547–579.

909 Lebart, L., and J.-P. Fénelon. 1971. Statistique et Informatique Appliquées. Dunod, Paris.

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

911 Legendre, P., and M. J. Anderson. 1999. Distance-based redundancy analysis: testing
912 multispecies responses in multifactorial ecological experiments. Ecological Monographs
913 **69**:1–24.

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

916 Legendre, P., and E. Gallagher. 2001. Ecologically meaningful transformations for ordination of
917 species data. Oecologia **129**:271–280.

918 Legendre, P., and L. Legendre. 2012. Numerical ecology. Third English edition. Elsevier,
919 Amsterdam.

920 Loreau, M. 2010. From populations to ecosystems: theoretical foundations for a new ecological
921 synthesis. Princeton University Press, New Jersey.

922 Maor, E. 2007. The Pythagorean theorem: a 4,000-year history. Princeton University Press,
923 Princeton.

924 McCoy, E. D., S. S. Bell, and K. Walters. 1986. Identifying biotic boundaries along
925 environmental gradients. Ecology **67**:749–759.

926 McGill, B. J., R. S. Etienne, J. S. Gray, D. Alonso, M. J. Anderson, H. K. Benecha, M. Dornelas,
927 B. J. Enquist, J. L. Green, F. L. He, A. H. Hurlbert, A. E. Magurran, P. A. Marquet, B. A.
928 Maurer, A. Ostling, C. U. Soykan, K. I. Ugland, and E. P. White. 2007. Species
929 abundance distributions: moving beyond single prediction theories to integration within
930 an ecological framework. Ecology Letters **10**:995–1015.

931 Motyka, J., 1947. O zadaniach i metodach badan geobotanicznych. Sur les buts et les méthodes
932 des recherches géobotaniques. Pages viii+168 in Annales Universitatis Mariae Curie-
933 Sklodowska (Lublin, Polonia), Sectio C, Supplementum I.

934 Niemelä, J. 1993. Mystery of the missing species: species-abundance distribution of boreal
935 ground-beetles. Annales Zoologici Fennici. **30**:169–172.

936 Ochiai, A. 1957. Zoogeographic studies on the soleoid fishes found in Japan and its
937 neighbouring regions. Bulletin of the Japanese Society of Scientific Fisheries **22**:526–
938 530.

939 Odum, E. P. 1950. Bird Populations of the Highlands (North Carolina) Plateau in Relation to
940 Plant Succession and Avian Invasion. Ecology **31**:587–605.

941 Økland, R. H. 1996. Are Ordination and Constrained Ordination Alternative or Complementary
942 Strategies in General Ecological Studies? Journal of Vegetation Science **7**:289–292.

943 Oksanen, J., F. G. Blanchet, R. Kindt, P. Legendre, P. R. Minchin, R. B. O'Hara, G. L. Simpson,
944 P. Sólymos, M. H. H. Stevens, and H. Wagner, 2012. vegan: Community Ecology
945 Package. URL `http://CRAN.R-project.org/package=vegan`.

946 Orlóci, L. 1967. An agglomerative method for classification of plant communities. Journal of
947 Ecology **55**:193–206.

948 Ouellette, M.-H., and P. Legendre, 2011. RsimSSDCOMPAS: Simulation of environment and
949 species composition in a deterministique environment.

950 Pearson, K. 1901. On lines and planes of closest fit to systems of points in space. Philosophical
951 Magazine **2**:559–572.

952 Peres-Neto, P. R. and P. Legendre, 2010. Estimating and controlling for spatial structure in the
953 study of ecological communities. Global Ecology and Biogeography **19**:174–184.

954 R Development Core Team, 2012. R: A Language and Environment for Statistical Computing. R
955 Foundation for Statistical Computing, Vienna, Austria. URL `http://www.R-`
956 `project.org/`.

957 Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research.
958 Sankhya: The Indian journal of statistic **26**:329–358.

959 Rao, C. R. 1995. A review of canonical coordinates and an alternative to correspondence
960 analysis using Hellinger distance. Qüestiió **19**:23–63.

961 Raup, D. M., and R. E. Crick. 1979. Measurement of faunal similarity in paleontology. Journal
962 of Paleontology **53**:1213–1227.

963 Robert, P., and Y. Escoufier. 1976. A unifying tool for linear multivariate statistical methods: the
964 RV-coefficient. Applied Statistics **25**:257–265.

965 Roux, G., and M. Roux. 1967. À propos de quelques méthodes de classification en
966 phytosociologie. Revue de Statistique Appliquée **15**:59–72.

967 Shepard, R. N. 1962. The analysis of proximities: multidimentional scaling with an unknown
968 distance. I. Psychometrika **27**:125–140.

969 Sokal, R., and C. Michener. 1958. A statistical method for evaluating systematic relationships.
970 University of Kansas Scientific Bulletin **38**:1409–1438.

971 Sørensen, T. 1948. A method of establishing groups of equal amplitude in plant sociology based
972 on similarity of species content and its application to analysis of vegetation on Danish
973 commons. Biologiske skrifter **5**:1–34.

974 Spence J. R., J. K. Niemelä 1994. Sampling carabid assemblages with pitfall traps: the madness
975 and the method. Canadian Entomologist **126**:881–894

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

978 ter Braak, C. J. F. 1987. The analysis of vegetation-environment relationships by canonical
979 correspondence analysis. Vegetatio **69**: 69-77.

980 ter Braak, C. J. F. 1994. Canonical community ordination – Part I: Basic theory and linear
981 methods. Écoscience **1**:127–140.

982 ter Braak, C. J. F., and P. F. M. Verdonschot. 1995. Canonical correspondence analysis and
983 related multivariate methods in aquatic ecology. Aquatic Sciences-Research Across
984 Boundaries **57**:255–289.

985 Van Burskirk, J. 2005. Local and landscape influence on amphibian occurrence and abundance.
986 Ecology **86**:1936–1947.

987 Whittaker, R. H. 1967. Gradient analysis of vegetation. Biological Review **49**:207–264.

988

APPENDIX A

Explanation of the construction of the explanatory variables and how they were combined for the simulation. One figure (Fig. A1), one table (Table A1) and R code

APPENDIX B

Comparison of association coefficients using a coefficient of determination (R^2). Eight figures (Figs. B1, B2, B3, B4, B5, B6, B7, and B8)

APPENDIX C

Comparison of consensus RDA constructed using only significant canonical axes with consensus RDA constructed with all canonical axes. Nine figures (Figs. C1, C2, C3, C4, C5, C6, C7, C8, and C9)

APPENDIX D

Comparison of canonical ordination models for abundance and presence-absence data using simulations. Four figures (Figs. D1, D2, D3, and D4)

APPENDIX E

Species code and names for Carabidae and trees species Two tables (Tables E1 and E2)

SUPPLEMENT

The R package ordiconsensus compiled for all plateforms.

1012 TABLE 1. List of dissimilarity coefficients compared. All coefficients are presented in a
1013 dissimilarity (distance) format.

Dissimilarity coefficients[a]	Equation	Reference	Comment
Binary symmetrical			
Simple-matching	$\sqrt{1 - \frac{a+d}{a+b+c+d}}^{\,b}$	Sokal and Michener (1958)	Directly related to Euclidean (see details in section *RDA and dissimilarity coefficients*)
Binary probabilistic			
Raup-Crick	$1 - p(a_{hi})^{\,bc}$	Raup and Crick (1979) McCoy et al. (1986)	
Binary asymmetrical			
Jaccard	$\sqrt{1 - \frac{a}{a+b+c}}^{\,b}$	Jaccard (1901)	Binary equivalent of any variation of the modified Gower dissimilarity

1014

1015 TABLE 1. Continue

Sørensen	$\sqrt{1-\frac{2a}{2a+b+c}}^{\,b}$	Sørensen (1948)	Binary equivalent of percentage difference
Ochiai	$\sqrt{1-\frac{a}{(a+b)+(a+c)}}^{\,b}$	Ochiai (1957)	
Distance between species profiles	$\sqrt{\frac{b+c}{(a+b)(a+c)}}^{\,b}$	Legendre and De Cáceres (2013)	A species' contribution is directly related to their abundance.
Abundance symmetrical			
Euclidean	$\sqrt{\sum_{j=1}^{p}(y_{1j}-y_{2j})^2}$	Mesopotamia ~1800 BC (Maor 2007)	Distance preserved in RDA
Abundance asymmetrical			
Chord	$\sqrt{\sum_{j=1}^{p}\left(\frac{y_{1j}}{\sqrt{\sum_{j=1}^{p}y_{1j}^2}}-\frac{y_{2j}}{\sqrt{\sum_{j=1}^{p}y_{2j}^2}}\right)^2}$	Orlòci (1967) Cavalli-Sforza and Edwards (1967)	On presence-absence data chord becomes $\sqrt{2(1-\text{Ochiai})}$

1016

1017 TABLE 1. Continue

Hellinger	$\sqrt{\sum_{j=1}^{p}\left(\sqrt{\frac{y_{1j}}{y_{1+}}}-\sqrt{\frac{y_{2j}}{y_{2+}}}\right)^2}^{\mathrm{d}}$	Rao (1995)	On presence-absence data Hellinger becomes $\sqrt{2(1-\mathrm{Ochiai})}$
Abundance asymmetrical			
χ^2	$\sqrt{y_{++}}\sqrt{\sum_{j=1}^{p}\frac{1}{y_{+j}}\left(\frac{y_{1j}}{y_{1+}}-\frac{y_{2j}}{y_{2+}}\right)^2}^{\mathrm{d}}$	Lebart and Fénelon (1971)	Dissimilarity preserved in CCA. Can also be used with presence-absence data.
Distance between species profiles	$\sqrt{\sum_{j=1}^{p}\left(\frac{y_{1j}}{y_{1+}}-\frac{y_{2j}}{y_{2+}}\right)^2}^{\mathrm{d}}$	Legendre and Gallagher (2001)	A species' contribution is directly related to their abundance.
Percentage difference	$\frac{\sum_{j=1}^{p}\lvert y_{1j}-y_{2j}\rvert}{\sum_{j=1}^{p}(y_{1j}+y_{2j})}$	Odum (1950)	This dissimilarity is often wrongfully referred to as the Bray-Curtis index[f]

1018

1019 TABLE 1. Continue

$\sqrt{\text{Percentage difference}}$	$\frac{\sum_{j=1}^{p} \lvert\sqrt{y_{1j}} - \sqrt{y_{2j}}\rvert}{\sum_{j=1}^{p} (\sqrt{y_{1j}} + \sqrt{y_{2j}})}$	Clarke and Green (1988)	Square or fourth rooting the raw data prior to calculating Percentage difference is often used when there is marked variation in abundance between species
$\sqrt[4]{\text{Percentage difference}}$	$\frac{\sum_{j=1}^{p} \lvert\sqrt[4]{y_{1j}} - \sqrt[4]{y_{2j}}\rvert}{\sum_{j=1}^{p} (\sqrt[4]{y_{1j}} + \sqrt[4]{y_{2j}})}$	Clarke and Green (1988)	
Modified Gower $\log_2$	$\frac{\sum_{j=1}^{p} w_j \lvert \log_2(y_{1j}) - \log_2(y_{1j})\rvert}{\sum_{j=1}^{p} w_j}$ [e]	Anderson et al. (2006)	Different log base are often used when there is marked variation in abundance between species. A high log base will generally reduce the emphasis of very abundant species more than a smaller one
Modified Gower $\log_5$	$\frac{\sum_{j=1}^{p} w_j \lvert \log_5(y_{1j}) - \log_5(y_{1j})\rvert}{\sum_{j=1}^{p} w_j}$ [e]	Anderson et al. (2006)	
Modified Gower $\log_{10}$	$\frac{\sum_{j=1}^{p} w_j \lvert \log_{10}(y_{1j}) - \log_{10}(y_{1j})\rvert}{\sum_{j=1}^{p} w_j}$ [e]	Anderson et al. (2006)	

1020 [a] All binary dissimilarities are presented in the form: $\mathbf{D} = \sqrt{1 - \mathbf{S}}$.

[b] The letters a, b, c, and d are defined in Table 2.

[c] *h* and *i* defined two different sites.

[d] y_{++} is the total sum of table **Y**, y_{+j} is the abundance of species *j*, and y_{i+} is the sum of all abundance of site *i*.

[e] w_j is used to exclude double-zeros by setting $w_j = 0$ whenever $y_{1j} = y_{1j} = 0$ and $w_j = 1$ elsewhere.

[f] Bray and Curtis (1957) did not design this coefficient nor was it their purpose. They used a transformed version of Steinhaus coefficient (Motyka 1947) in their paper, which is equivalent to the coefficient proposed by Odum (1950) described above (Legendre and Legendre, 2012).

TABLE 2. Contingency table describing the similarity between two sites where species presence or absence were observed. *a* is the number of species present at sites 1 and 2, *b* is the number of species present at site 1 but absent at site 2, *c* is the number of species found at site 2 but not at site 1, and *d* is the number of species absent at both sites. The mathematical formulas explain how to calculate *a*, *b*, *c* and *d* from a community matrix **Y** comprising of *p* species, where y_{1j} and y_{2j} represent the occurrence (0 for absence, 1 for presence) of species *j* at sites 1 and 2 respectively.

		Site 2	
		1 (species present)	0 (species absent)
Site 1	1 (species present)	a $\sum_{j=1}^{p} y_{1j}y_{2j}$	b $\sum_{j=1}^{p} y_{1j} - \sum_{j=1}^{p} y_{1j}y_{2j}$
	0 (species absent)	c $\sum_{j=1}^{p} y_{2j} - \sum_{j=1}^{p} y_{1j}y_{2j}$	d $p - a - b - c$

TABLE 3. Variance explained (R^2) by RDA models constructed independently with each dissimilarity coefficient using data from the ecological illustration, where the tree relative basal area was used to model a ground beetle (Carabidae) assemblage. The abundance data are the abundances of carabids divided by the number of days the traps were active at each sites, while the presence-absence data are the occurrence of species at each site. Results are given for all but the double-zero symmetrical coefficients. The coefficients are described in Table 1.

Dissimilarity coefficient	R^2
Abundance data	
Species profiles	0.303
Chord	0.321
Hellinger	0.340
χ^2	0.094
Percentage difference	0.203
$\sqrt{\text{Percentage difference}}$	0.238
$\sqrt[4]{\text{Percentage difference}}$	0.249
Modified Gower $\log_2$	0.297
Modified Gower $\log_5$	0.304
Modified Gower $\log_{10}$	0.302
Presence-absence	
Species profiles	0.225
Ochiai	0.244

1048 TABLE 3. Continue.

Raup-Crick	0.190
χ^2	0.048
Jaccard	0.188
Sørensen	0.244

1049
1050

FIGURE CAPTIONS

FIGURE 1. Species-abundance distributions (SAD) used in the simulations. These SADs are presented using Preston (1948) graphs where the abundance classes in the abscissa increase according to a geometric progression whose lower bound is made of the values 2^k with k being the successive integers from 0 and up and the ordinate indicates the number of species in each abundance class. These SADs were used as a basis for the simulations to generate site-by-species data table. Each SAD represents a community of 20 species. They were constructed to encompass a wide range of variation in abundance patterns.

FIGURE 2. Comparison of explained variance (R^2) between 11 dissimilarity coefficients calculated from simulated communities following different species abundance distributions (SAD) using abundance data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved to calculate R^2. Points are R^2 means of all simulations and error bars represent 95% confidence intervals. Coefficients are presented in different panels for visual clarity. Letters along the abscissa refer to the SADs presented in Figure 1. A line was drawn along the R^2 results of each coefficient to facilitate comparisons between coefficients. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 0.001). A thousand simulations were run for each SAD.

FIGURE 3 Comparison of explained variance (R^2) between 7 dissimilarity coefficients calculated from simulated communities following different species abundance distributions (SAD) using presence-absence data. Details as in Fig. 2.

FIGURE 4 Schematic representation of consensus RDA. (a) The first step of the procedure is to perform a series of RDAs (tb-RDA or db-RDA) to model the community data **Y** using explanatory variables **X**. Each RDA is computed with a different dissimilarity coefficient using

scaling type 1 (distance triplot, **Z** matrices). In the figure, *K* different dissimilarity coefficients are used. (b) For each of the *K* dissimilarity coefficients, the significant axes within each **Z** matrix are grouped in a large matrix. (c) An RDA is then performed on this large matrix using **X** as the explanatory variables. (d) This RDA yields the site scores consensus matrix $\mathbf{Z}^*$, a diagonal matrix of eigenvalues $\mathbf{\Lambda}^*$, and the consensus canonical coefficients $\mathbf{C}^*$. (e) Equation 5 is then used to obtain the consensus species scores $\mathbf{U}^*$. (f) $\mathbf{Z}^*$, $\mathbf{U}^*$, and $\mathbf{C}^*$ can be used to draw a consensus RDA triplot; the eigenvalues in $\mathbf{\Lambda}^*$ show the importance of each axis in the consensus triplot.

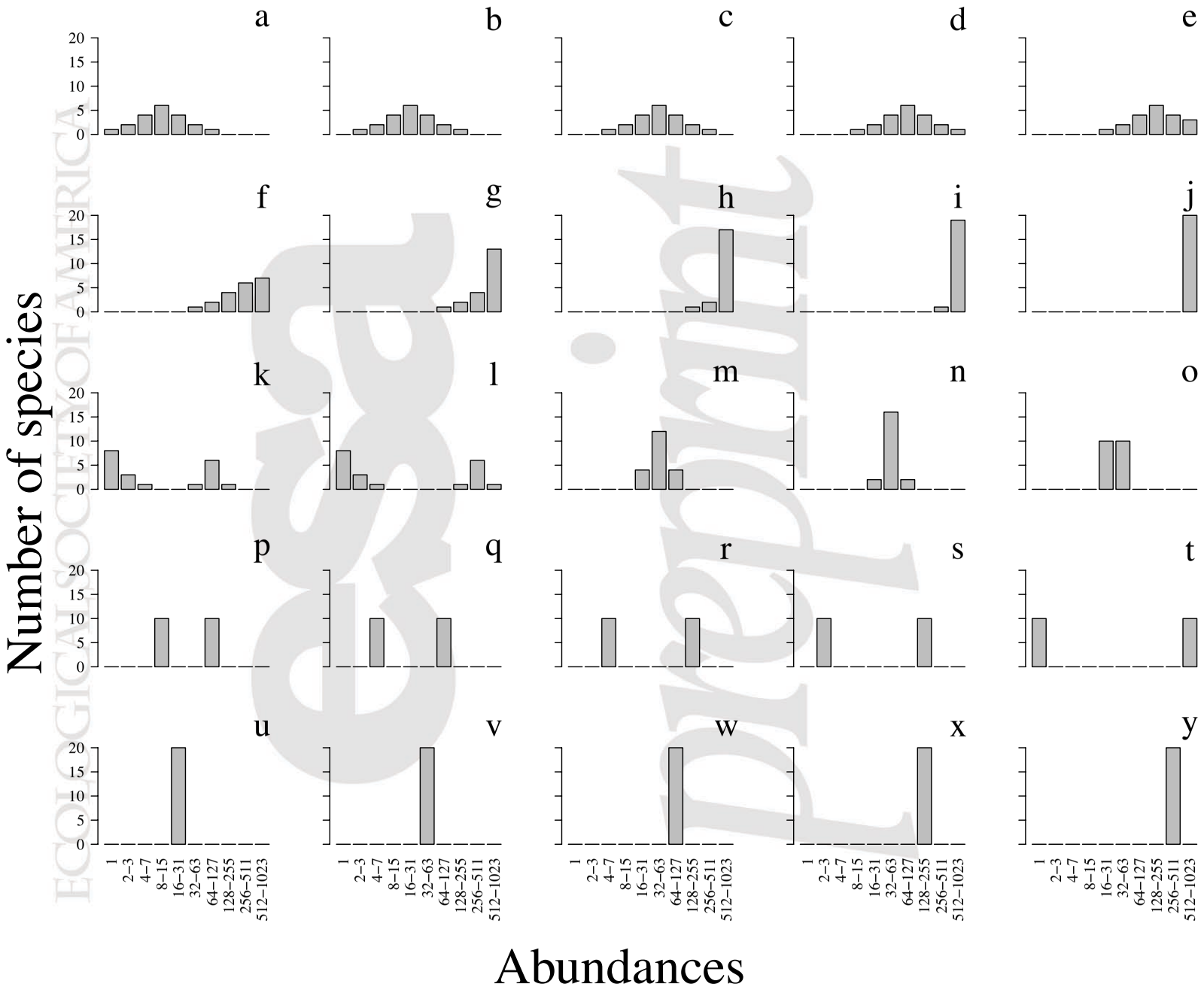
FIGURE 5. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only the significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from the abundance data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (SAD) presented in Figure 1. The ordinate presents the difference between RV coefficients calculated using all canonical axes and RV coefficients calculated using only the significant axes; all difference values were in the interval [–0.03, 0.09]. The results are presented using boxplots. The upper and lower sections of the box define the first (25%) and third (75%) quartiles of the data, and the line in the middle of the box is the median (50%). The lower whiskers describe the 1.5 interquartile range of the first quartile, the upper whisker stands for the 1.5 interquartile range of the third quartile, and the points indicate outliers. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 2). A thousand simulations were run for each SAD.

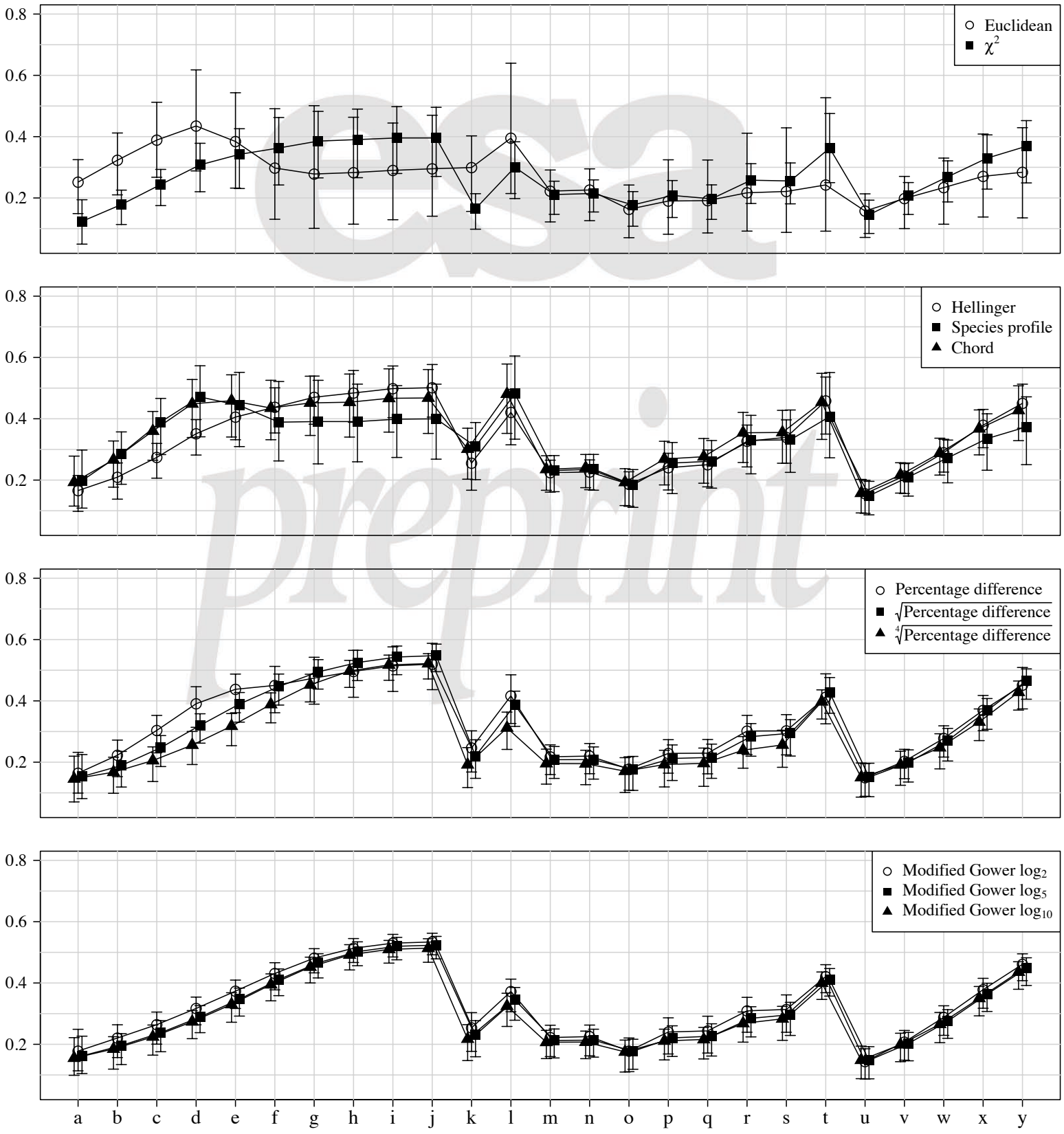
FIGURE 6. Comparison between abundance and presence-absence data, showing how much of the true species structure (Equation 6 without the error term) is modelled by the canonical ordination models. For each data type (abundance and presence-absence), the significant

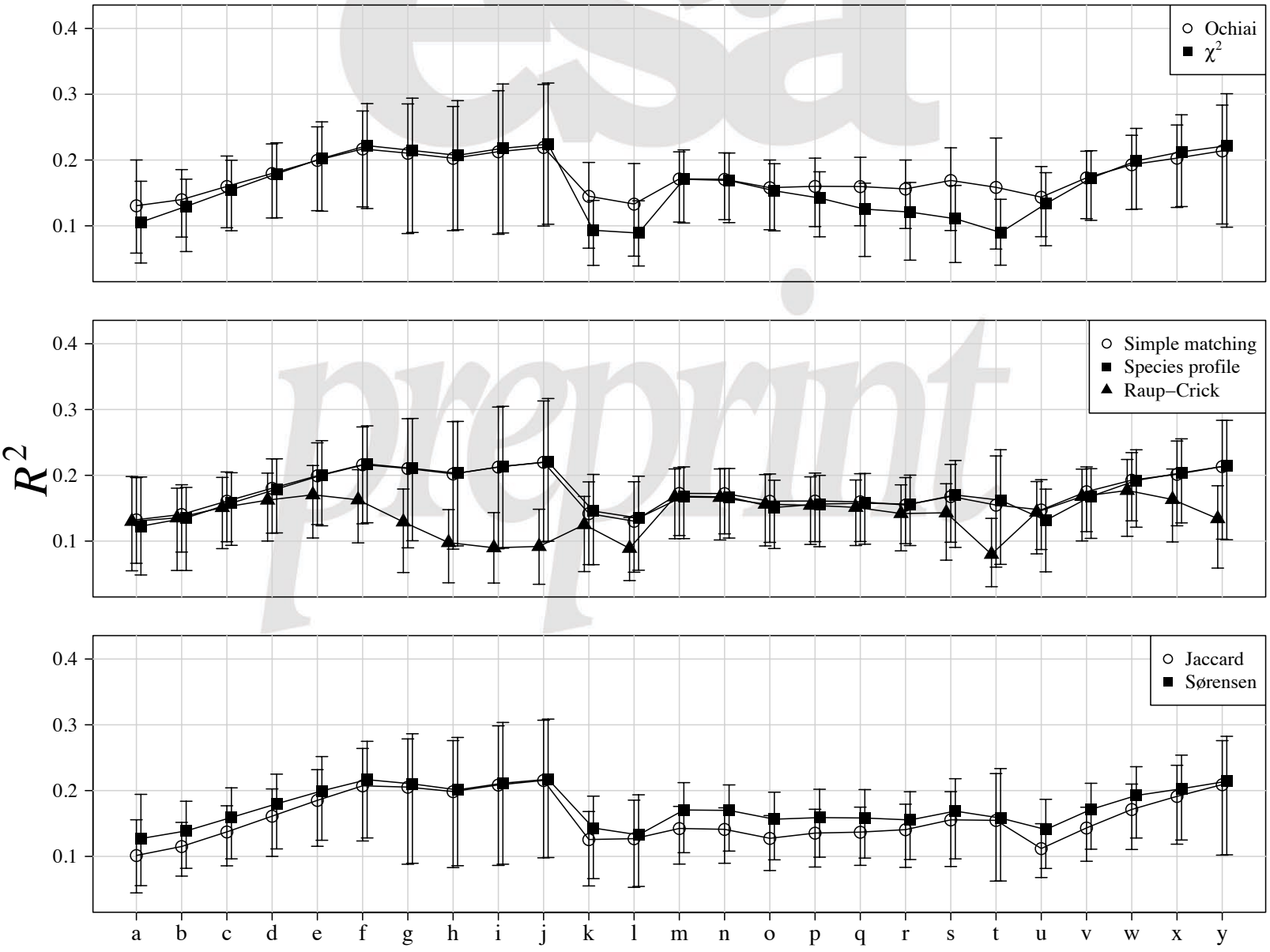
canonical axes computed using all dissimilarity coefficients (excluding the double-zero symmetrical coefficients) were grouped. RV coefficients were then used to correlate the true species structure with the grouped significant canonical axes. Error bars represent 95% confidence intervals. Letters along the abscissa refer to the species-abundance distribution (SAD) presented in Figure 1. A line was drawn along the R^2 results of each dissimilarity coefficient to facilitate comparison between the two data types. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 0.001). A thousand simulations were run for each SAD.

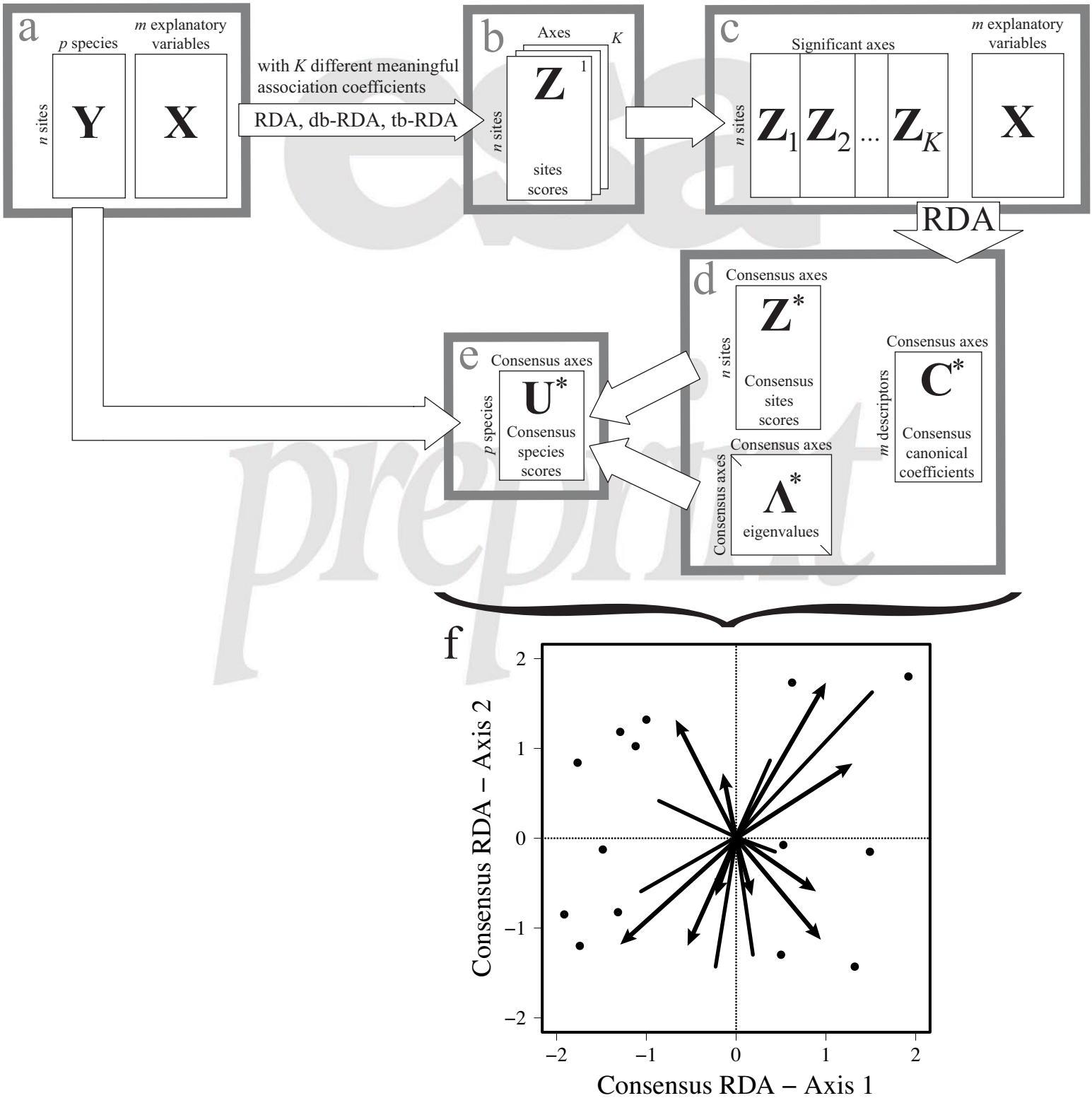
FIGURE 7. Comparison of (a) species-abundance distributions (SAD) and (d) species-presence distributions (SPD), and consensus RDA results for (c) abundance and (f) presence-absence data, using Carabidae data sampled at the *Ecosystem Management Emulating Natural Disturbances* (EMEND) experimental area in Alberta, Canada. The minimum spanning trees (MST) comparing coefficients (b) for abundance data and (e) for presence-absence data show that the χ^2 distance produced RDAs very different from the other coefficients. For both data types, the χ^2 distance was the only coefficient not used to compute the consensus RDA in (c) and (f). The SAD and SPD were constructed in the same way, with the exception that for SPD it is the occurrence of species that is considered to construct bins, not their abundance. The SAD (a) and SPD (d) were used as references to relate the results presented in this figure to the simulation results presented in Figures 2, 3 and 6. The consensus RDA triplots (scaling 2, correlation triplots) (c, f) describe the relationship between ground beetle species (arrows) and the relative basal area of trees by species (lines) using all coefficients except the double-zero symmetrical coefficients and the χ^2 distance. The species codes for the Carabidae and trees are

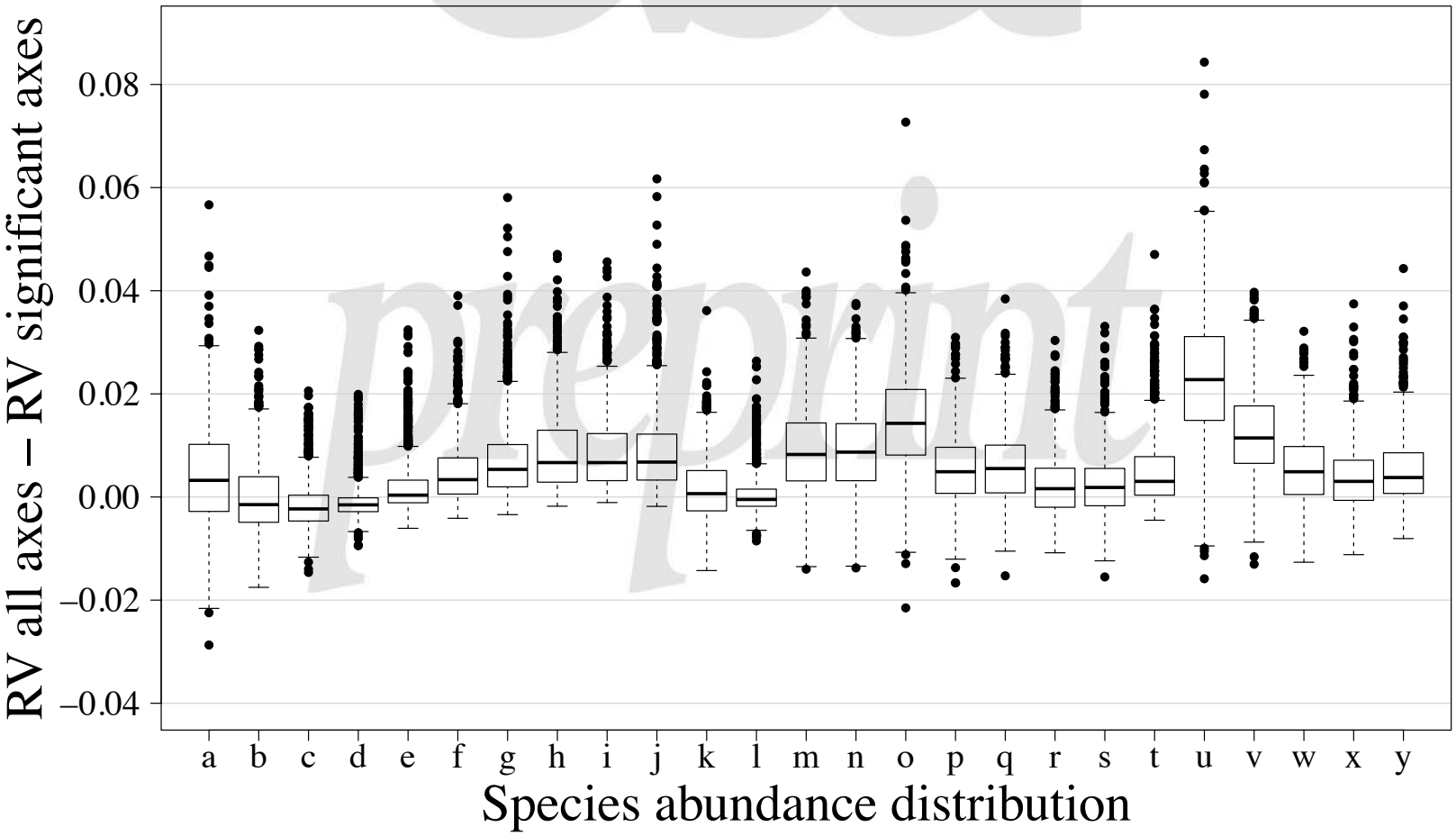
provided in Tables E1-E2. In (b) and (e), MG stands for modified Gower and PD for percentage difference; the names of all other coefficients are written fully.

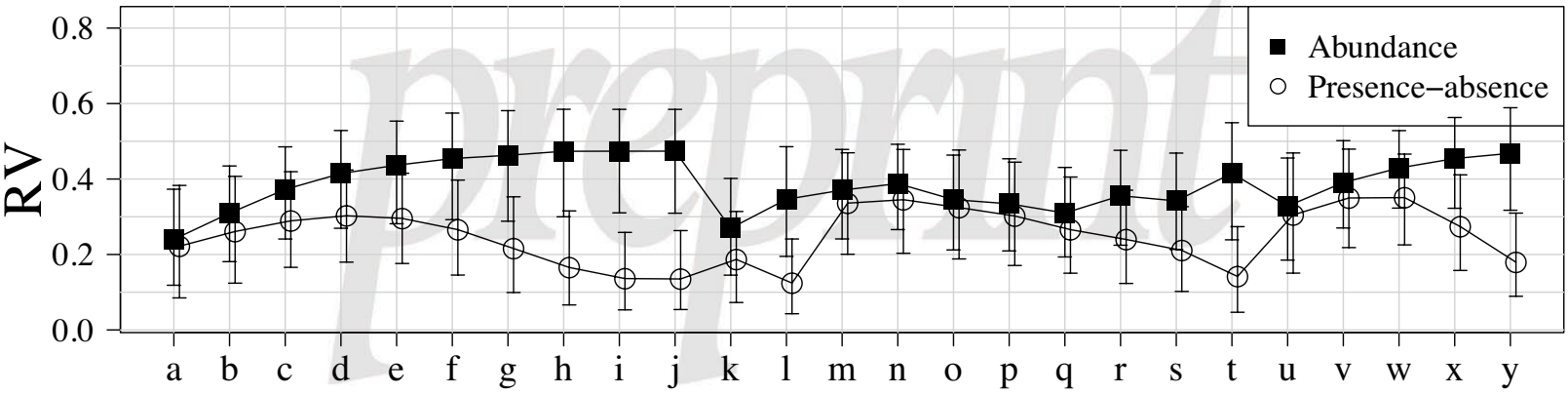


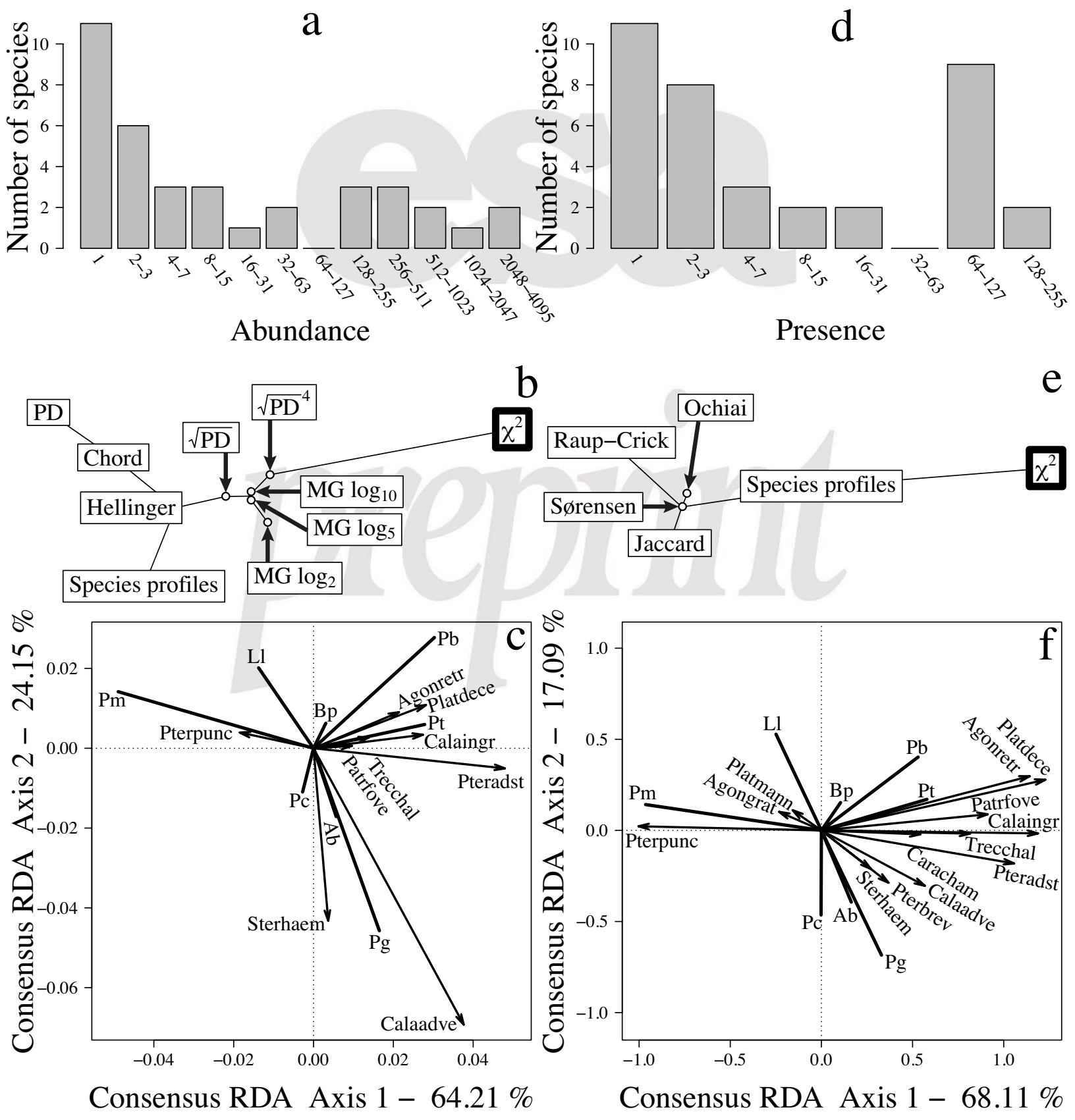












APPENDIX A

Ecological Archives EXXX-XXX-A1

EXPLANATION OF THE CONSTRUCTION OF THE EXPLANATORY VARIABLES AND HOW THEY WERE COMBINED FOR THE SIMULATION. ONE FIGURE (FIG. A1), ONE TABLE (TABLE A1) AND R CODE.

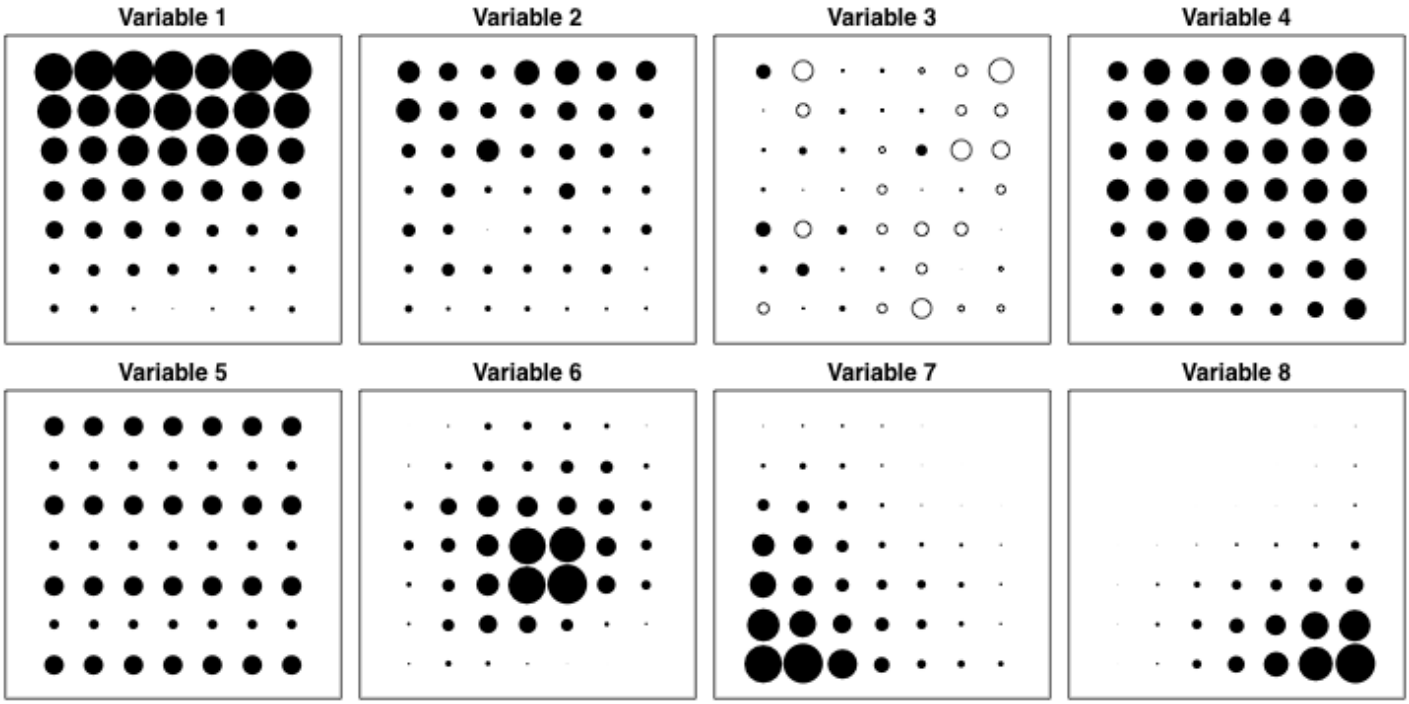


Fig. A1. Eight variables used in the construction of the simulated species

These variables were constructed using the RsimSSDCOMPAS package through the R statistical language using the following R code:

```
variable1<-SimSSDR(7,7,1,10,range11=5,range12=5,range21=1,range22=1,
                   nsp1=1,nsp2=1,varnor=list(rep(0,3)),SAE=TRUE,
                   SAR=FALSE)$E

variable2<-SimSSDR(7,7,1,5,range11=1,range12=1,range21=1,range22=1,
                   nsp1=1,nsp2=1,varnor=list(rep(0,3)),SAE=TRUE,
                   SAR=FALSE)$E

variable3<-SimSSDR(7,7,0,range11=5,range12=5)$E

variable4<-SimSSDR(7,7,2,10,range11=5,range12=5,range21=1,range22=1,
                   nsp1=1,nsp2=1,varnor=list(rep(0,3)),SAE=TRUE,
                   SAR=FALSE)$E
```

```
variable5<-SimSSDR(7,7,4,5,range11=2,range12=2,range21=1,range22=1,
                   nsp1=1,nsp2=1,varnor=list(rep(0,3)),SAE=FALSE,
                   SAR=FALSE)$E

variable6<-SimSSDR(7,7,3,10,range11=10,range12=10,range21=1,range22=1,
                   nsp1=1,nsp2=1,varnor=list(rep(0,3)),SAE=FALSE,
                   SAR=FALSE,centroide=list(c(0,0)))$E

variable7<-SimSSDR(7,7,3,10,range11=10,range12=10,range21=1,range22=1,
                   nsp1=1,nsp2=1,varnor=list(rep(0,3)),SAE=FALSE,
                   SAR=FALSE,centroide=list(c(1,1)))$E

variable8<-SimSSDR(7,7,3,10,range11=10,range12=10,range21=1,range22=1,
                   nsp1=1,nsp2=1,varnor=list(rep(0,3)),SAE=FALSE,
                   SAR=FALSE,centroide=list(c(10,0)))$E
```

Table A1: Combinations of explanatory variables and weight (regression coefficient) used to construct each species. The number associated to each species is the order given in the site-by-species table

Species	Explanatory variables combined	Weight given to (regression coefficient of) each species
1	1 and 4	2
2	1 and 5	0.1
3	1 and 6	-2
4	1 and 7	-0.1
5	1 and 8	2
6	2 and 3	0.5
7	2 and 5	-2
8	2 and 6	-0.5
9	2 and 7	2
10	2 and 8	1
11	3 and 5	-2
12	3 and 6	-1
13	3 and 7	2
14	3 and 8	0.5
15	4 and 5	-2
16	4 and 6	-0.5
17	4 and 7	2
18	4 and 8	0.1
19	5 and 8	-2
20	6 and 7	-0.1

APPENDIX B

Ecological Archives EXXX-XXX-A2

COMPARISON OF ASSOCIATION COEFFICIENTS USING A COEFFICIENT OF DETERMINATION (R^2). EIGHT FIGURES (FIGS. B1, B2, B3, B4, B5, B6, B7, AND B8)

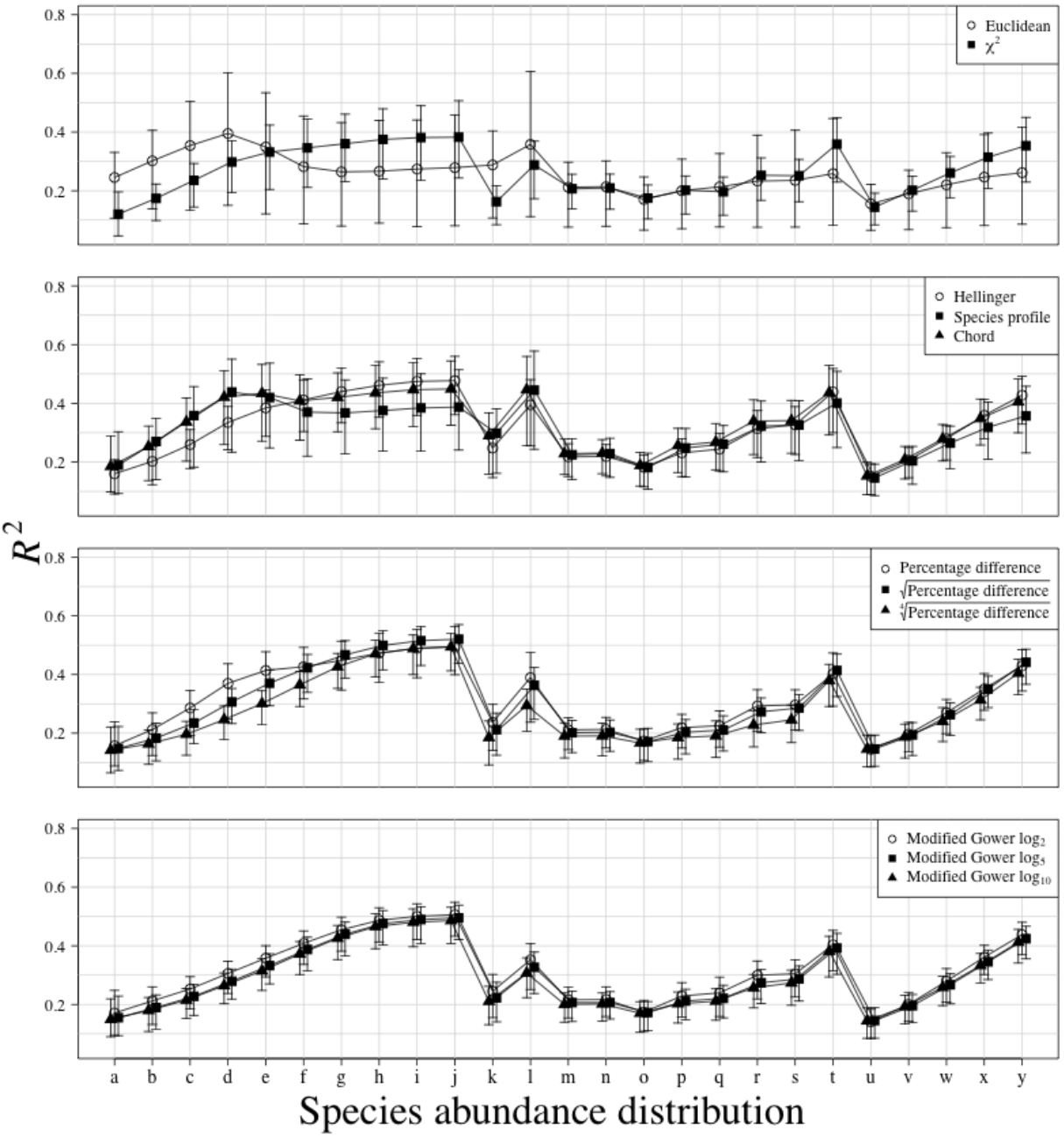


Fig. B1. Comparison of explained variance (R^2) between 11 association coefficients calculated on abundance data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved to calculate R^2. Points are R^2 means of all simulations and error bars represent 95% confidence intervals. Association coefficients are presented in different panels for visual clarity. Letters on the x-axis refer to the species-abundance distribution (SAD) presented in Fig. 1. A line was drawn between each SAD of each association coefficient to ease comparisons between coefficients. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 0.25). A thousand simulations were run for each SAD.

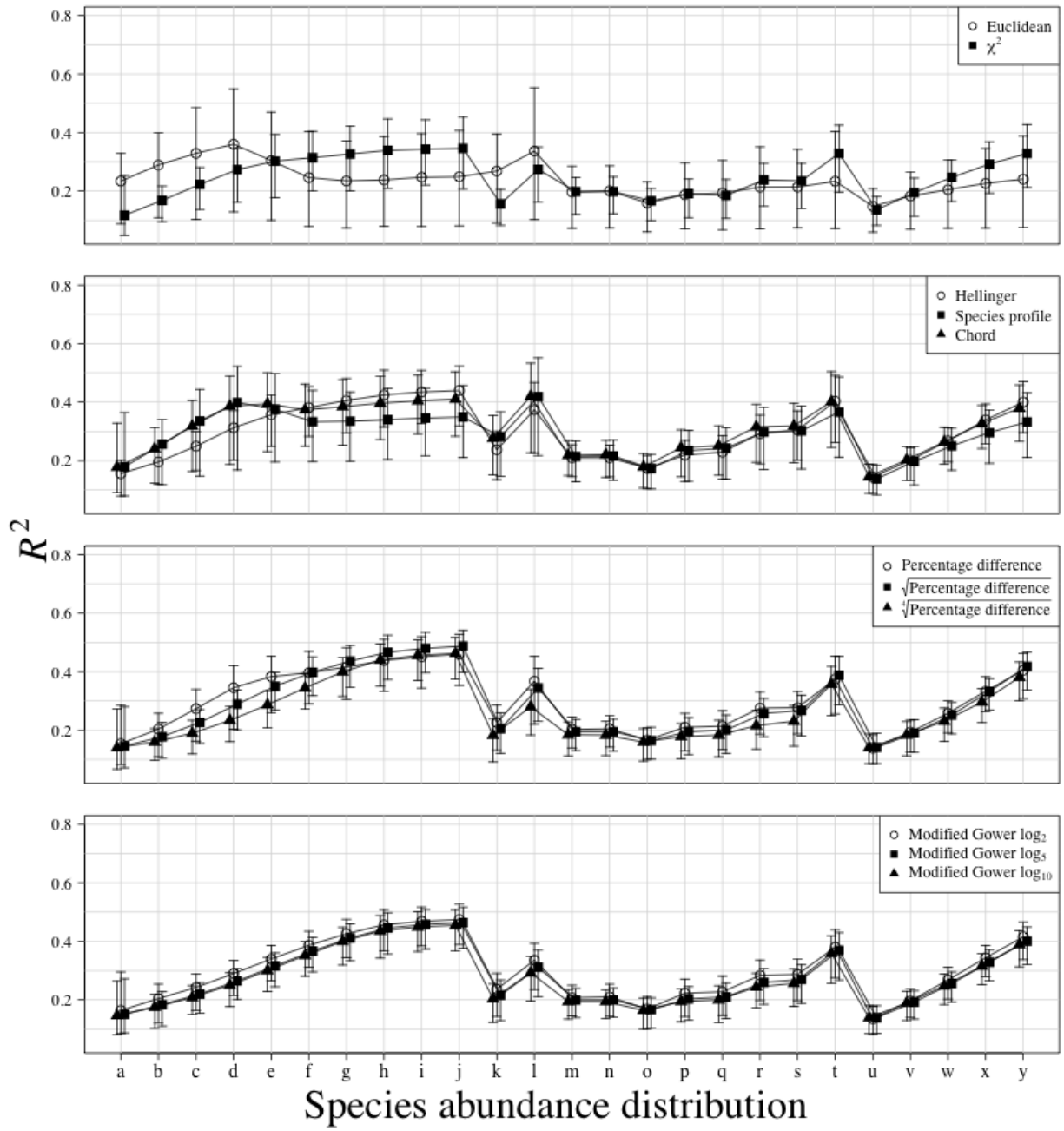


Fig. B2. Comparison of explained variance (R^2) between 11 association coefficients calculated on abundance data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved to calculate R^2. Points are R^2 means of all simulations and error bars represent 95% confidence intervals. Association coefficients are presented in different panels for visual clarity. Letters on the x-axis refer to the species-abundance distribution (SAD) presented in Fig. 1. A line was drawn between each SAD of each association coefficient to ease comparisons between coefficients. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 0.5). A thousand simulations were run for each SAD.

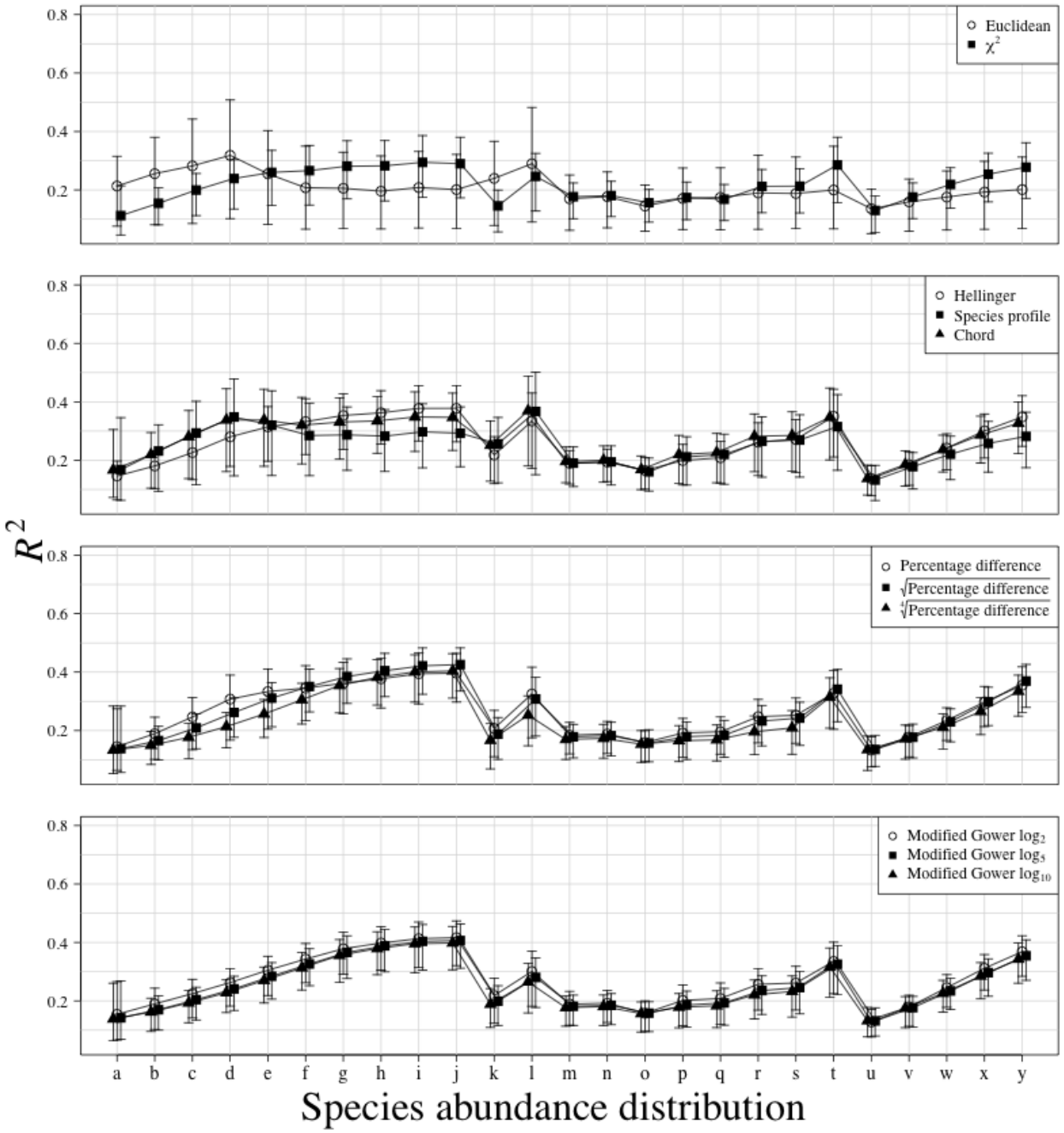


Fig. B3. Comparison of explained variance (R^2) between 11 association coefficients calculated on abundance data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved to calculate R^2. Points are R^2 means of all simulations and error bars represent 95% confidence intervals. Association coefficients are presented in different panels for visual clarity. Letters on the x-axis refer to the species-abundance distribution (SAD) presented in Fig. 1. A line was drawn between each SAD of each association coefficient to ease comparisons between coefficients. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 1). A thousand simulations were run for each SAD.

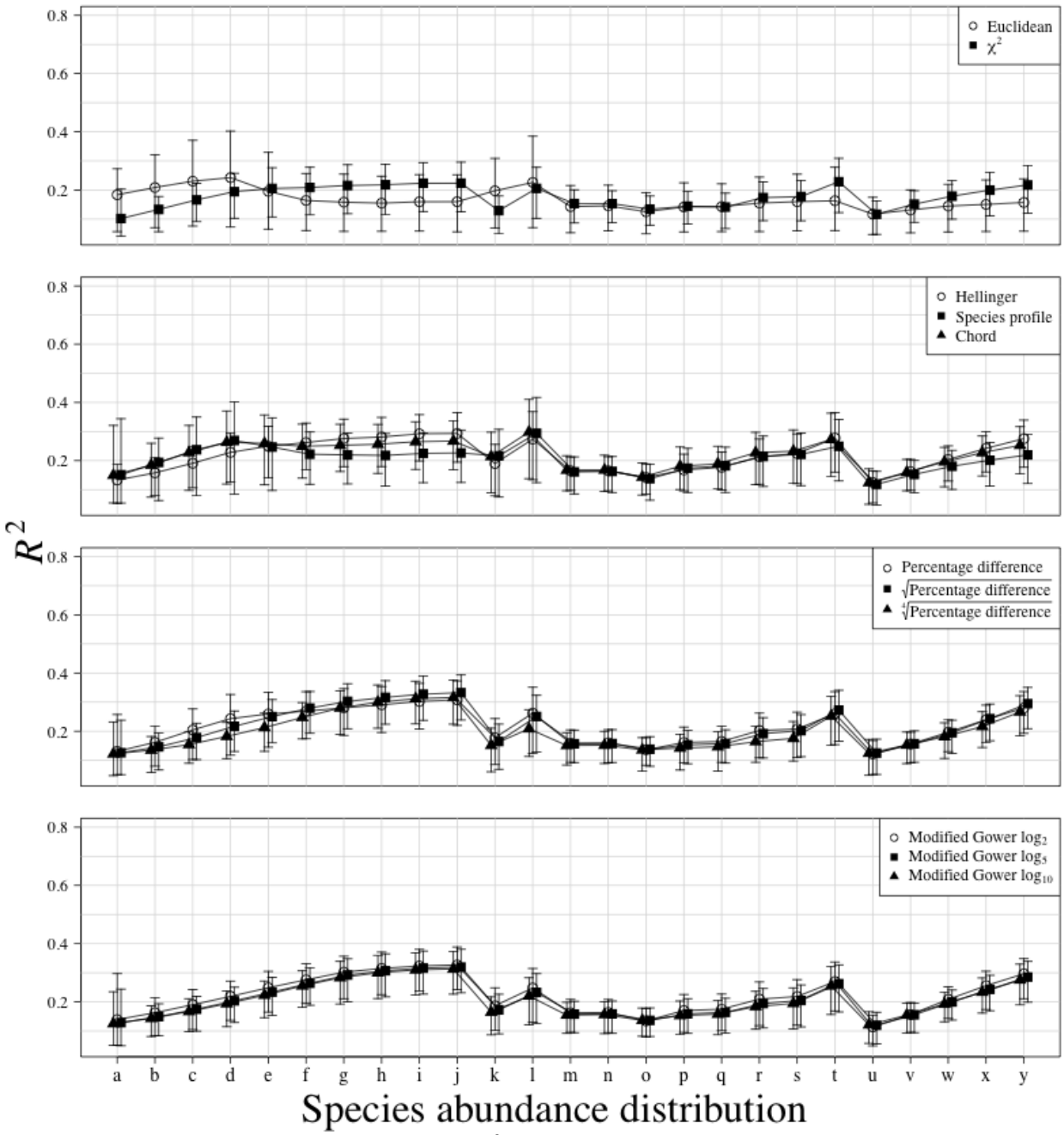


Fig. B4. Comparison of explained variance (R^2) between 11 association coefficients calculated on abundance data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved to calculate R^2. Points are R^2 means of all simulations and error bars represent 95% confidence intervals. Association coefficients are presented in different panels for visual clarity. Letters on the x-axis refer to the species-abundance distribution (SAD) presented in Fig. 1. A line was drawn between each SAD of each association coefficient to ease comparisons between coefficients. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 2). A thousand simulations were run for each SAD.

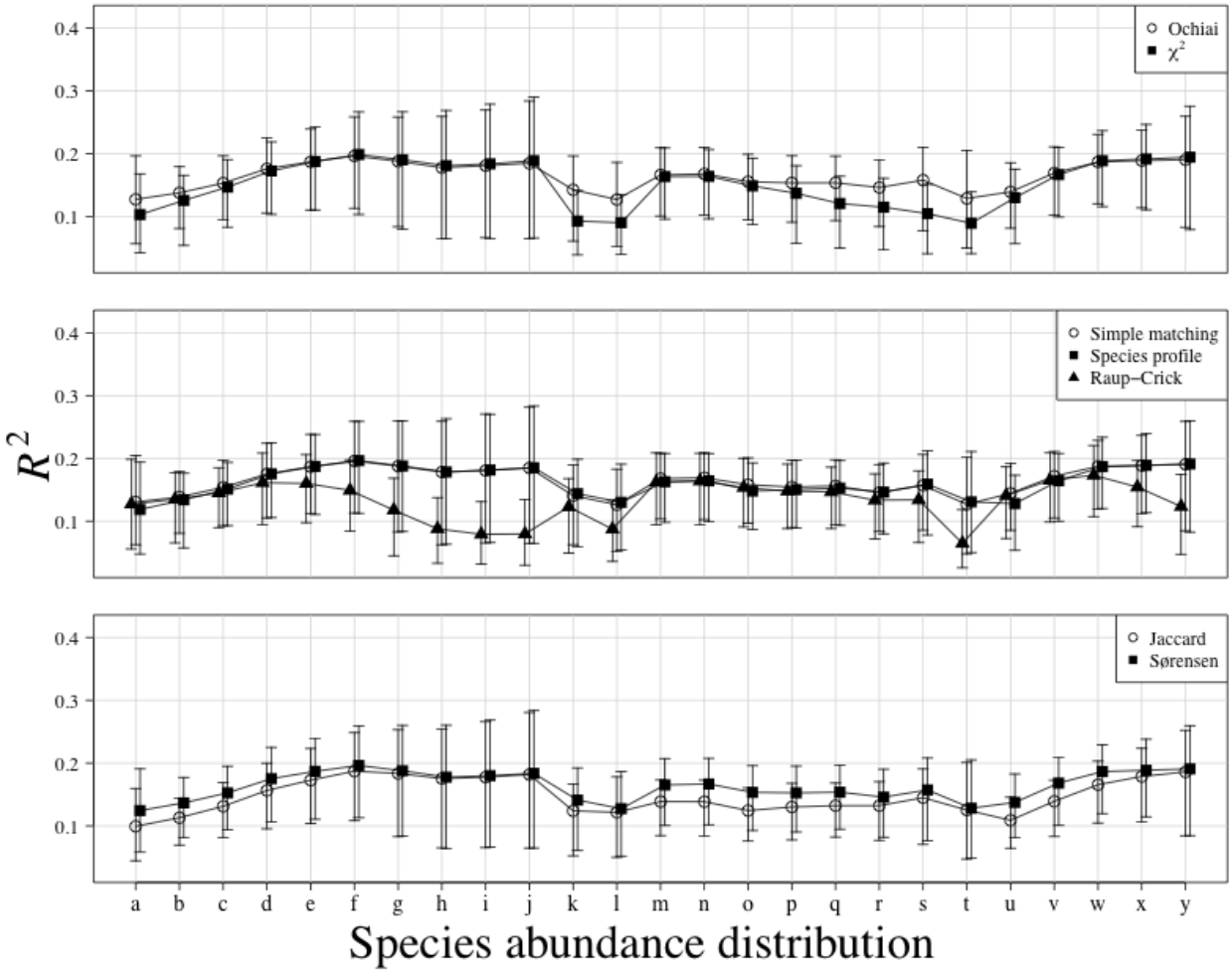


Fig. B5. Comparison of explained variance (R^2) between 6 association coefficients calculated on presence-absence data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved to calculate R^2. Points are R^2 means of all simulations and error bars represent 95% confidence intervals. Association coefficients are presented in different panels for visual clarity. Letters on the x-axis refer to the species-abundance distribution (SAD) presented in Fig. 1. A line was drawn between each SAD of each association coefficient to ease comparisons between coefficients. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 0.25). A thousand simulations were run for each SAD.

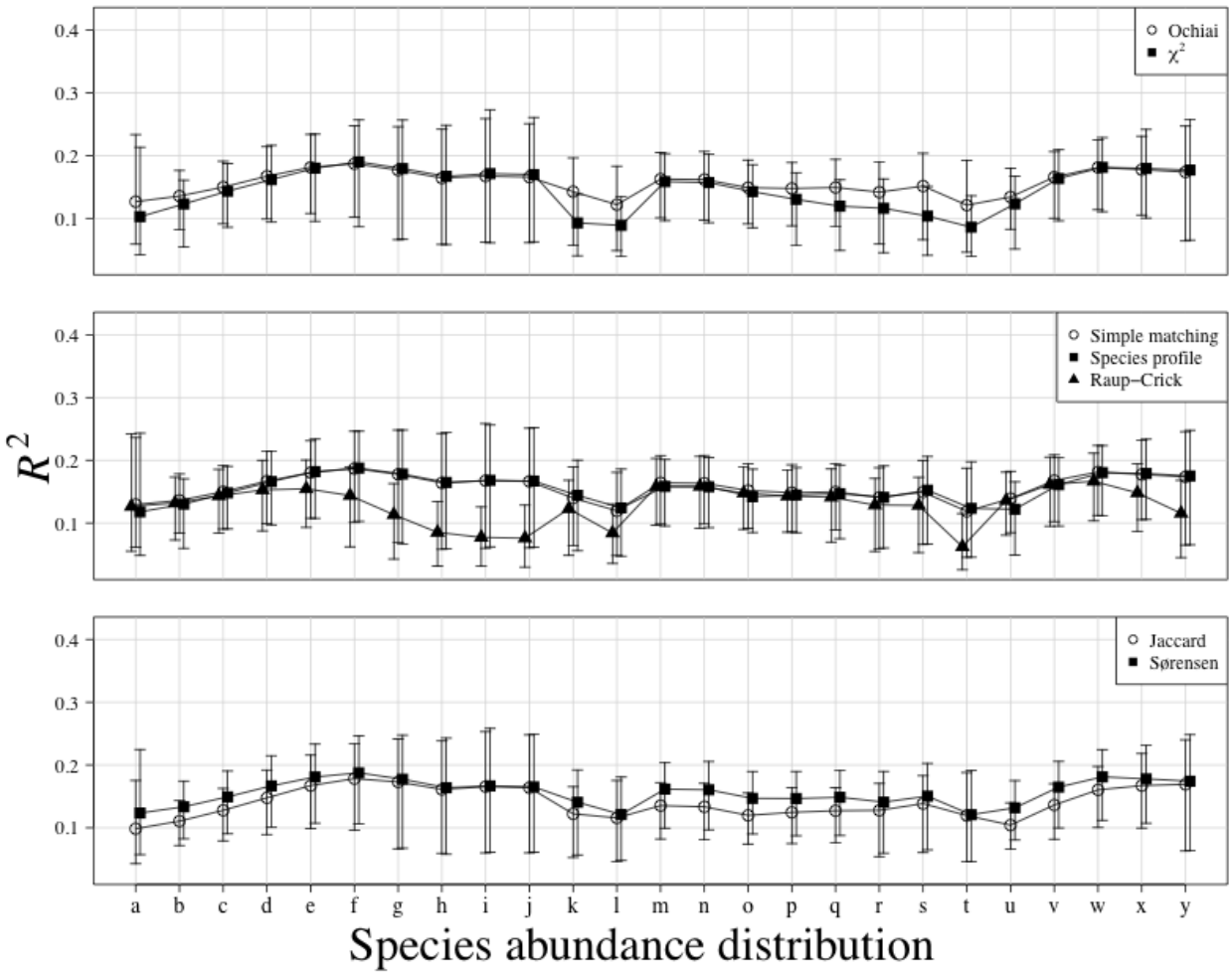


Fig. B6. Comparison of explained variance (R^2) between 6 association coefficients calculated on presence-absence data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved to calculate R^2. Points are R^2 means of all simulations and error bars represent 95% confidence intervals. Association coefficients are presented in different panels for visual clarity. Letters on the x-axis refer to the species-abundance distribution (SAD) presented in Fig. 1. A line was drawn between each SAD of each association coefficient to ease comparisons between coefficients. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 0.5). A thousand simulations were run for each SAD.

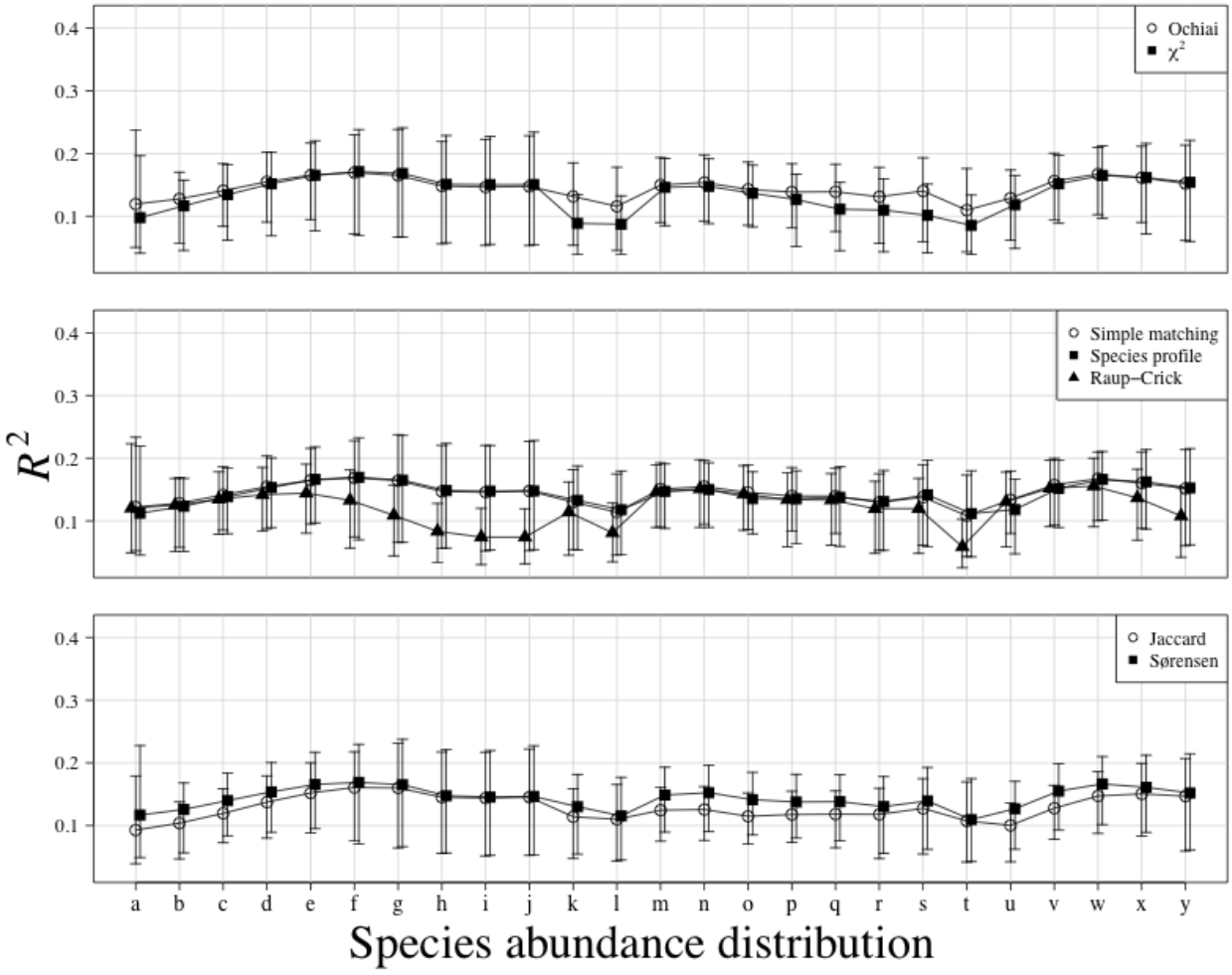


Fig. B7. Comparison of explained variance (R^2) between 6 association coefficients calculated on presence-absence data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved to calculate R^2. Points are R^2 means of all simulations and error bars represent 95% confidence intervals. Association coefficients are presented in different panels for visual clarity. Letters on the x-axis refer to the species-abundance distribution (SAD) presented in Fig. 1. A line was drawn between each SAD of each association coefficient to ease comparisons between coefficients. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 1). A thousand simulations were run for each SAD.

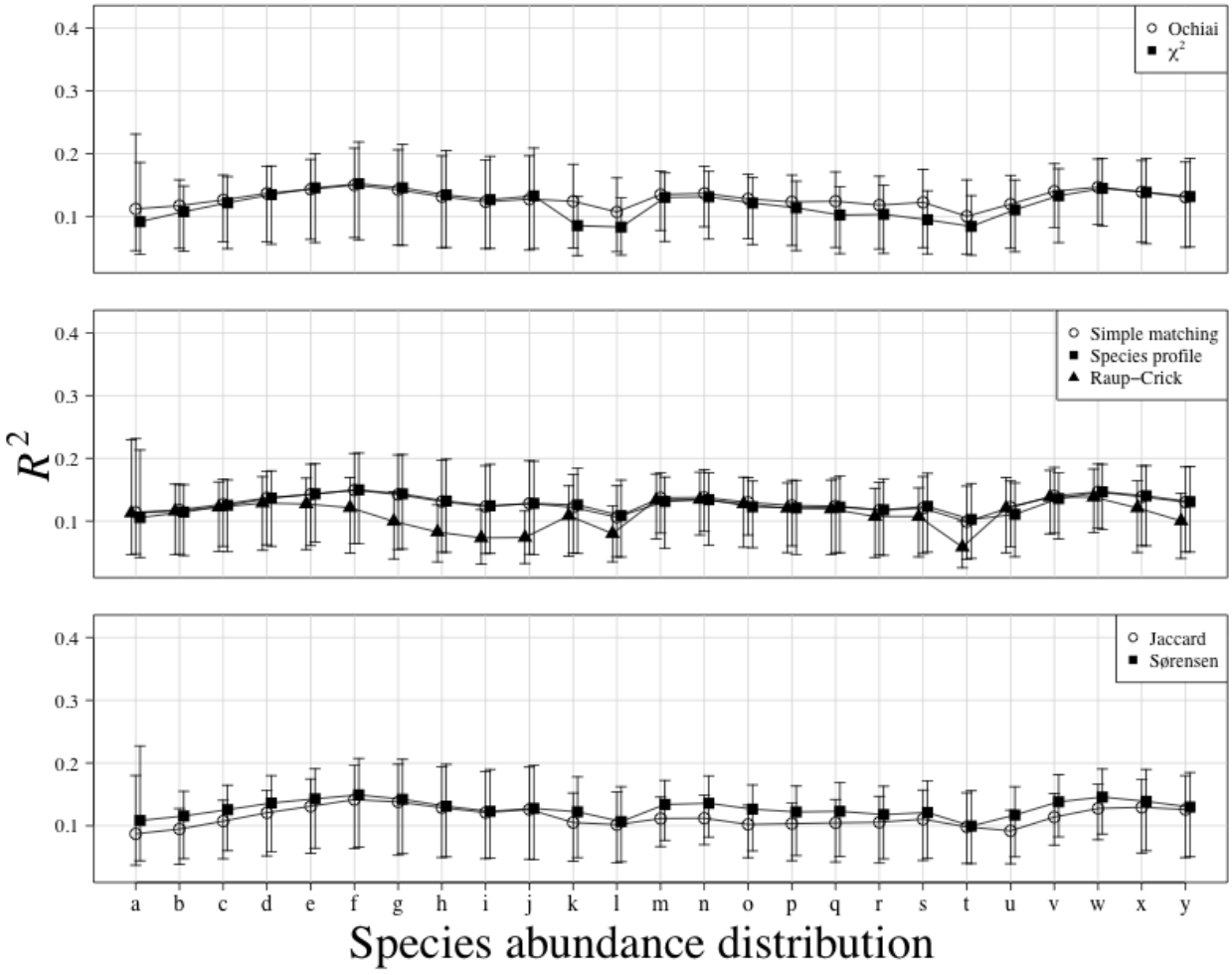


Fig. B8. Comparison of explained variance (R^2) between 6 association coefficients calculated on presence-absence data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved to calculate R^2. Points are R^2 means of all simulations and error bars represent 95% confidence intervals. Association coefficients are presented in different panels for visual clarity. Letters on the x-axis refer to the species-abundance distribution (SAD) presented in Fig. 1. A line was drawn between each SAD of each association coefficient to ease comparisons between coefficients. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 2). A thousand simulations were run for each SAD.

APPENDIX C

Ecological Archives EXXX-XXX-A3

COMPARISON OF CONSENSUS RDA CONSTRUCTED USING ONLY SIGNIFICANT CANONICAL AXES WITH CONSENSUS RDA CONSTRUCTED WITH ALL CANONICAL AXES. NINE FIGURES (FIGS. C1, C2, C3, C4, C5, C6, C7, C8, AND C9)

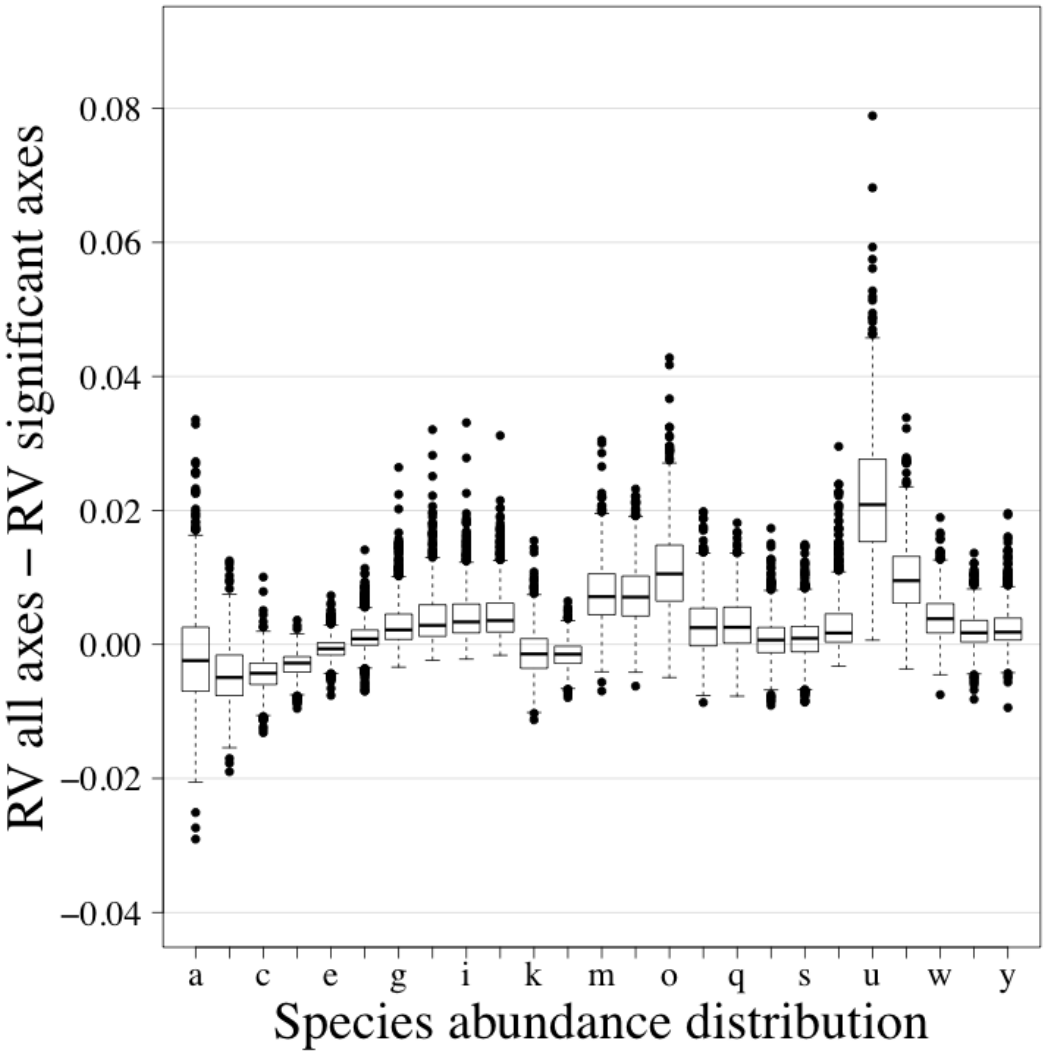


FIGURE C1. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from abundance data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (SAD) as presented in Figure 3.1. The ordinate presents the difference between RV coefficients calculated using all canonical axes and RV coefficients calculated using only the significant axes. The results are presented using boxplots. The upper and lower sections of the box define the first (25%) and third (75%) quartiles of the data, and the line in the middle of the box the median (50%). The lower whiskers describe the 1.5 interquartile range of the first quartile, the upper whisker stands for the 1.5 interquartile range of the third quartile, and the points indicate outliers. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 0.001). A thousand simulations were run for each SAD.

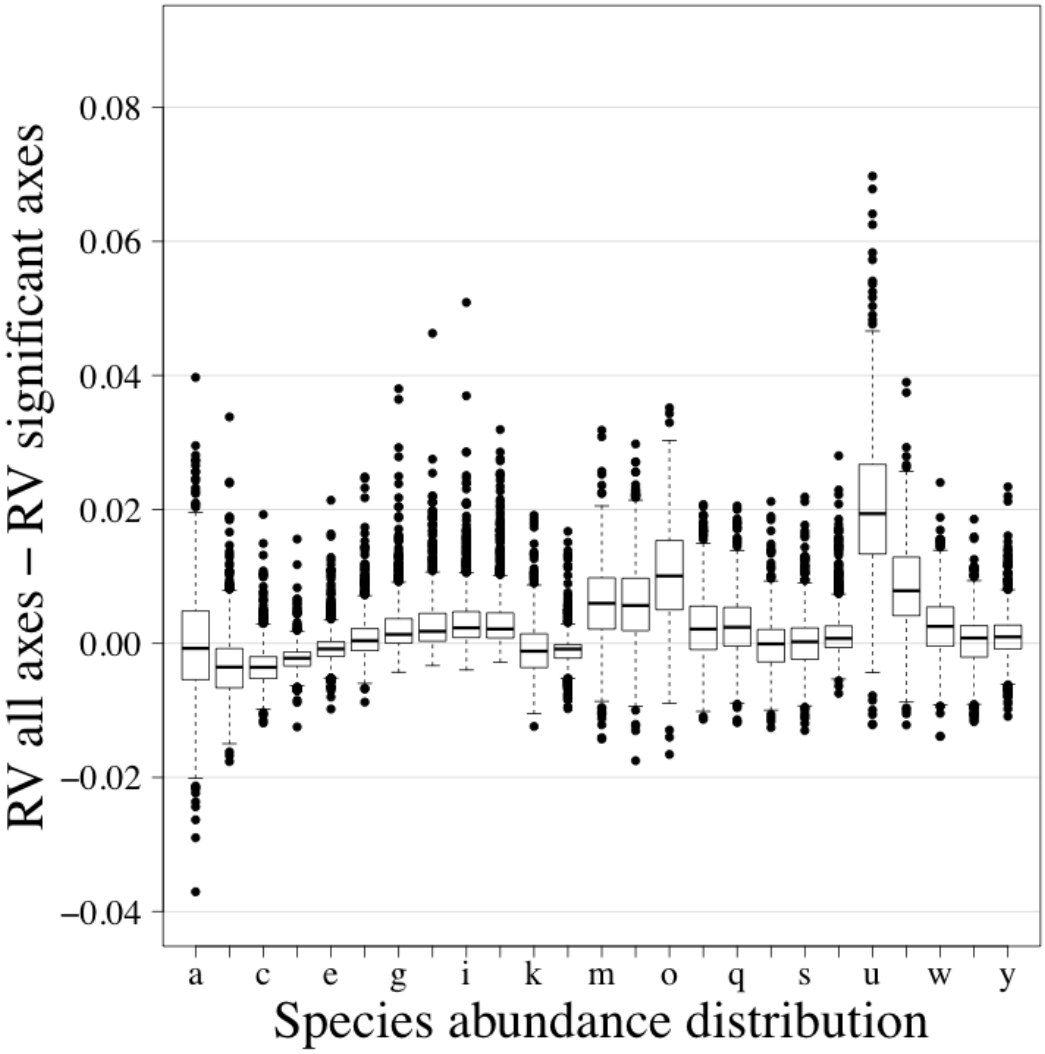


FIGURE C2. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from abundance data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (SAD) as presented in Figure 3.1. The ordinate presents the difference between RV coefficients calculated using all canonical axes and RV coefficients calculated using only the significant axes. The results are presented using boxplots. The upper and lower sections of the box define the first (25%) and third (75%) quartiles of the data, and the line in the middle of the box the median (50%). The lower whiskers describe the 1.5 interquartile range of the first quartile, the upper whisker stands for the 1.5 interquartile range of the third quartile, and the points indicate outliers. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 0.25). A thousand simulations were run for each SAD.

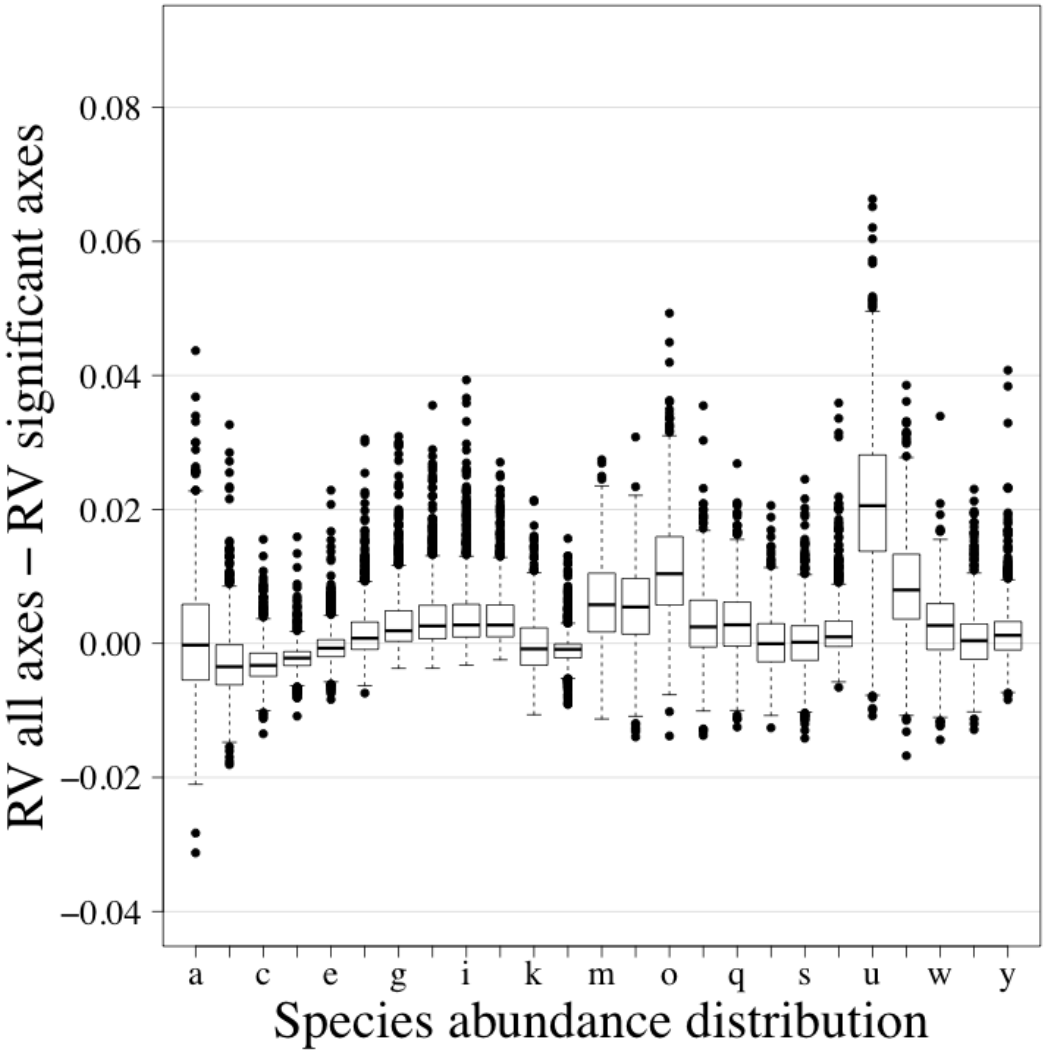


FIGURE C3. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from abundance data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (SAD) as presented in Figure 3.1. The ordinate presents the difference between RV coefficients calculated using all canonical axes and RV coefficients calculated using only the significant axes. The results are presented using boxplots. The upper and lower sections of the box define the first (25%) and third (75%) quartiles of the data, and the line in the middle of the box the median (50%). The lower whiskers describe the 1.5 interquartile range of the first quartile, the upper whisker stands for the 1.5 interquartile range of the third quartile, and the points indicate outliers. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 0.5). A thousand simulations were run for each SAD.

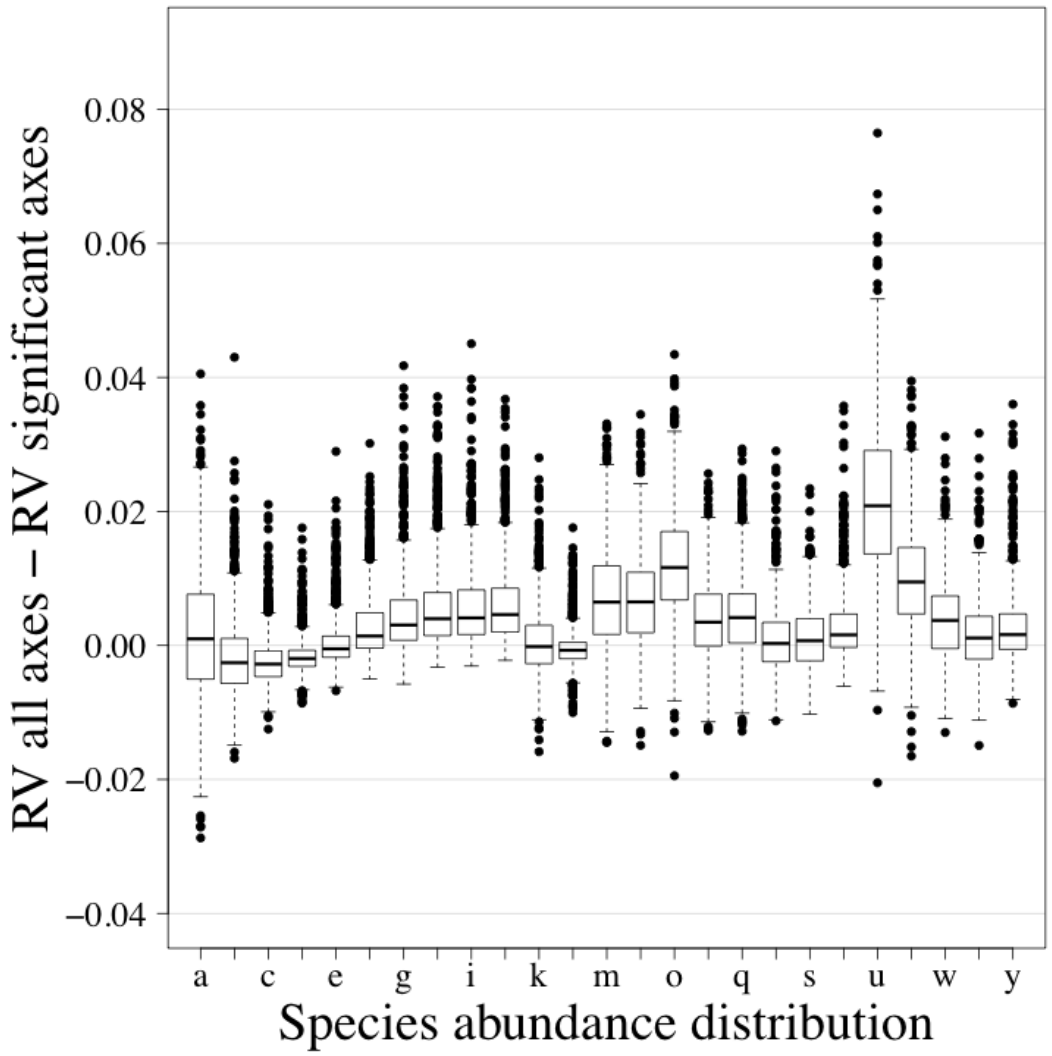


FIGURE C4. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from abundance data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (SAD) as presented in Figure 3.1. The ordinate presents the difference between RV coefficients calculated using all canonical axes and RV coefficients calculated using only the significant axes. The results are presented using boxplots. The upper and lower sections of the box define the first (25%) and third (75%) quartiles of the data, and the line in the middle of the box the median (50%). The lower whiskers describe the 1.5 interquartile range of the first quartile, the upper whisker stands for the 1.5 interquartile range of the third quartile, and the points indicate outliers. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 1). A thousand simulations were run for each SAD.

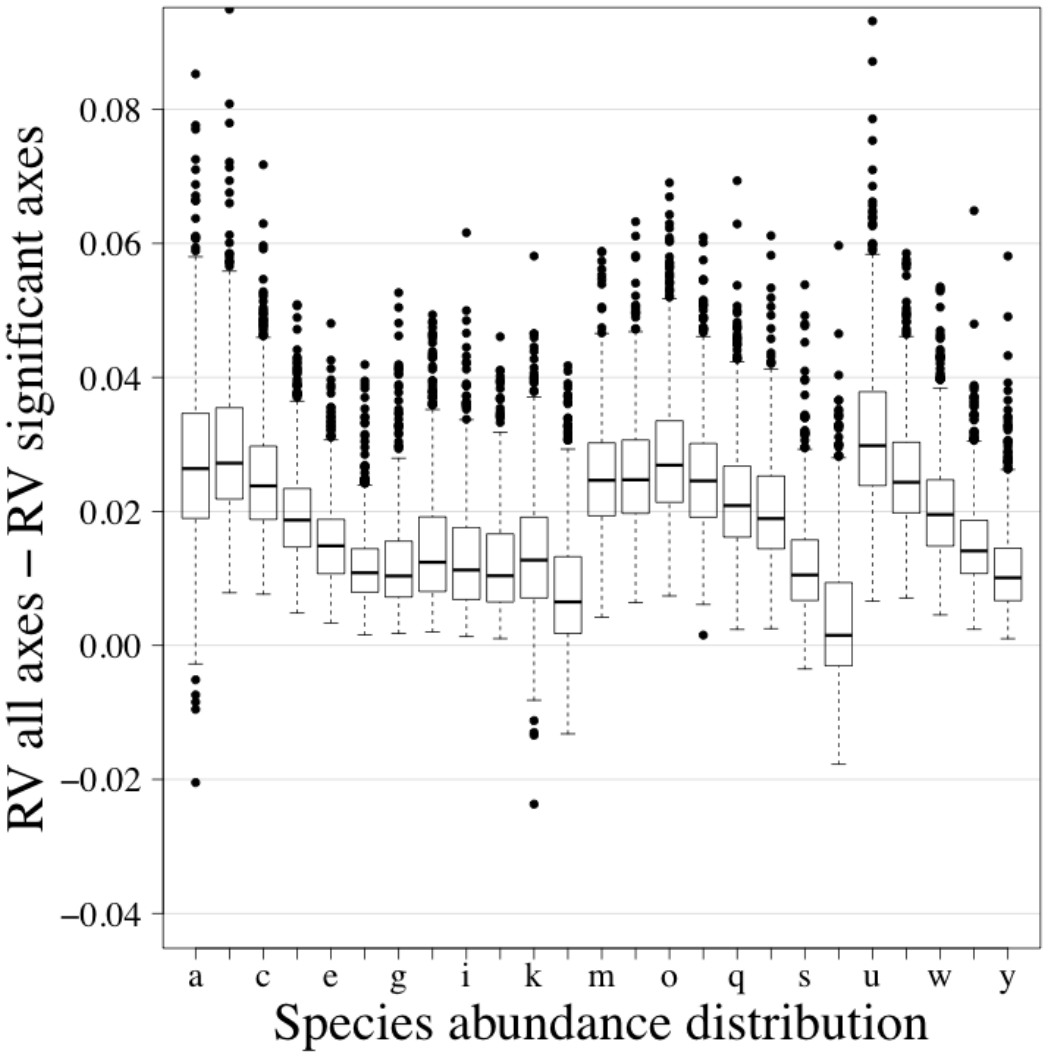


FIGURE C5. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from presence-absence data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (SAD) as presented in Figure 3.1. The ordinate presents the difference between RV coefficients calculated using all canonical axes and RV coefficients calculated using only the significant axes. The results are presented using boxplots. The upper and lower sections of the box define the first (25%) and third (75%) quartiles of the data, and the line in the middle of the box the median (50%). The lower whiskers describe the 1.5 interquartile range of the first quartile, the upper whisker stands for the 1.5 interquartile range of the third quartile, and the points indicate outliers. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 0.001). A thousand simulations were run for each SAD.

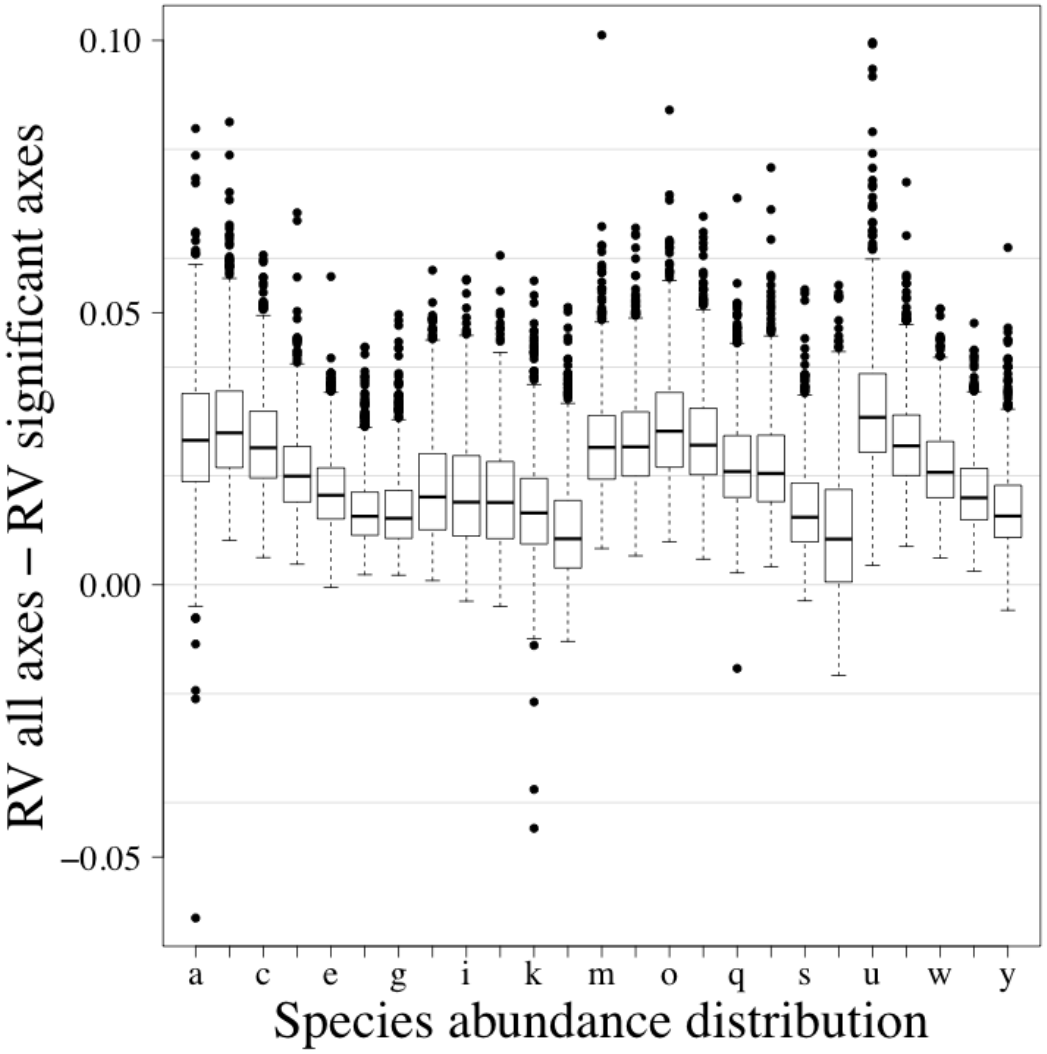


FIGURE C6. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from presence-absence data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (SAD) as presented in Figure 3.1. The ordinate presents the difference between RV coefficients calculated using all canonical axes and RV coefficients calculated using only the significant axes. The results are presented using boxplots. The upper and lower sections of the box define the first (25%) and third (75%) quartiles of the data, and the line in the middle of the box the median (50%). The lower whiskers describe the 1.5 interquartile range of the first quartile, the upper whisker stands for the 1.5 interquartile range of the third quartile, and the points indicate outliers. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 0.25). A thousand simulations were run for each SAD.

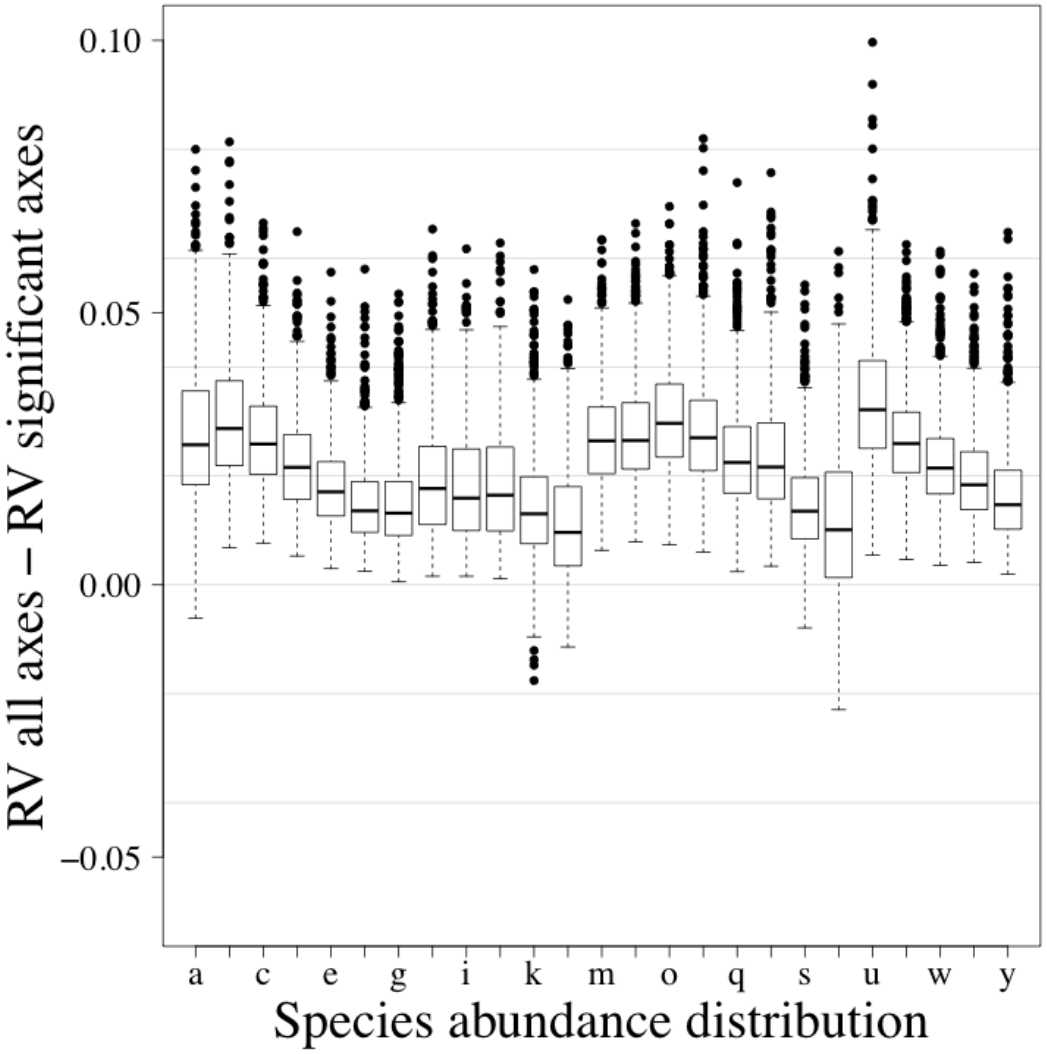


FIGURE C7. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from presence-absence data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (SAD) as presented in Figure 3.1. The ordinate presents the difference between RV coefficients calculated using all canonical axes and RV coefficients calculated using only the significant axes. The results are presented using boxplots. The upper and lower sections of the box define the first (25%) and third (75%) quartiles of the data, and the line in the middle of the box the median (50%). The lower whiskers describe the 1.5 interquartile range of the first quartile, the upper whisker stands for the 1.5 interquartile range of the third quartile, and the points indicate outliers. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 0.5). A thousand simulations were run for each SAD.

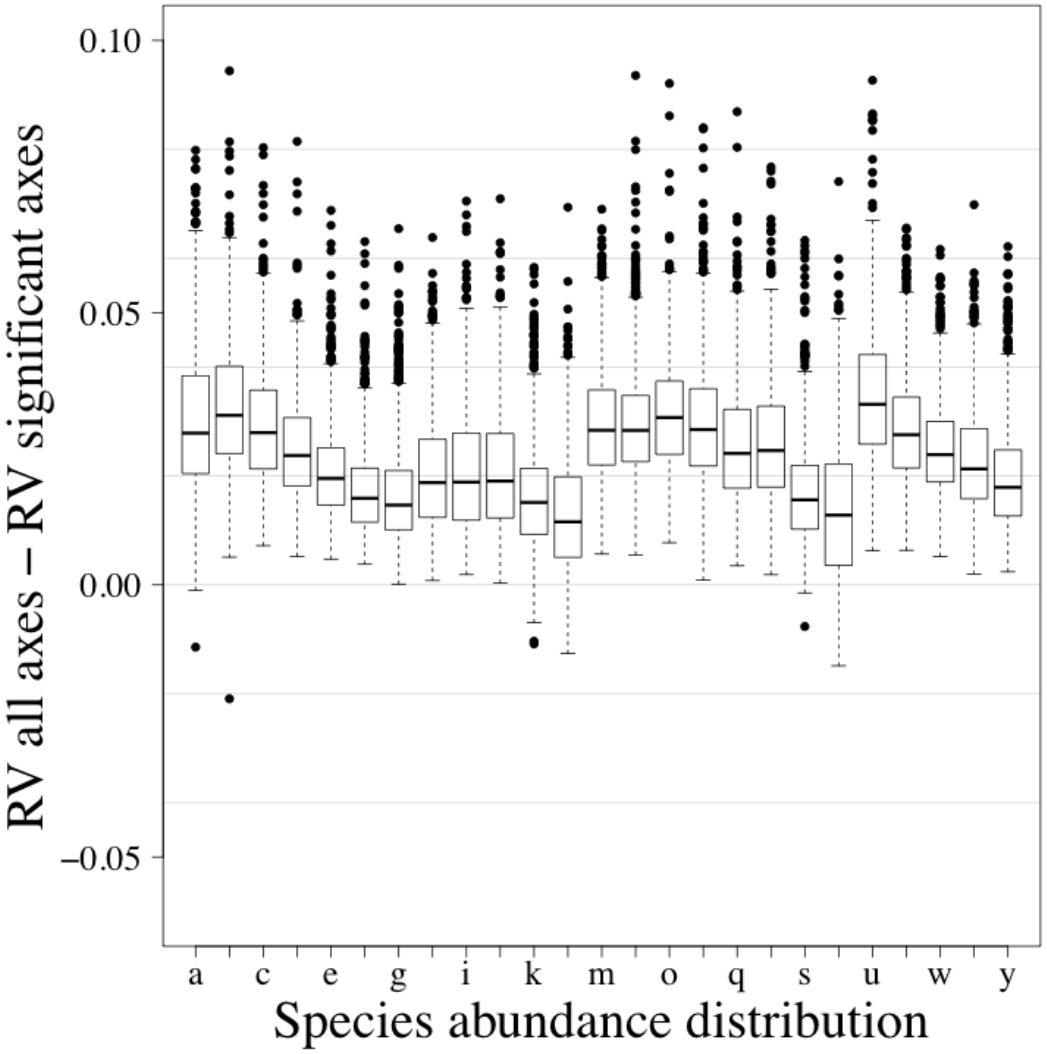


FIGURE C8. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from presence-absence data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (SAD) as presented in Figure 3.1. The ordinate presents the difference between RV coefficients calculated using all canonical axes and RV coefficients calculated using only the significant axes. The results are presented using boxplots. The upper and lower sections of the box define the first (25%) and third (75%) quartiles of the data, and the line in the middle of the box the median (50%). The lower whiskers describe the 1.5 interquartile range of the first quartile, the upper whisker stands for the 1.5 interquartile range of the third quartile, and the points indicate outliers. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 1). A thousand simulations were run for each SAD.

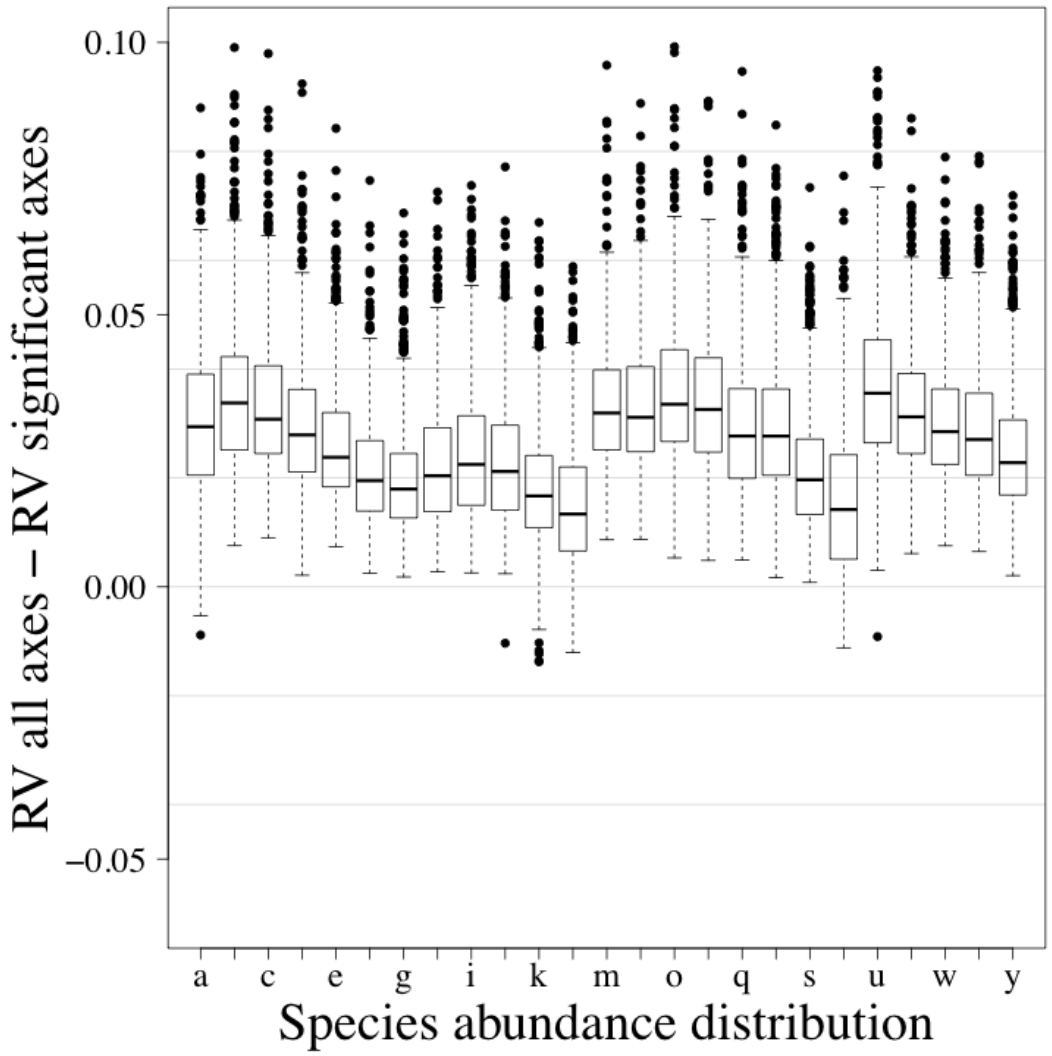


FIGURE C9. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from presence-absence data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (SAD) as presented in Figure 3.1. The ordinate presents the difference between RV coefficients calculated using all canonical axes and RV coefficients calculated using only the significant axes. The results are presented using boxplots. The upper and lower sections of the box define the first (25%) and third (75%) quartiles of the data, and the line in the middle of the box the median (50%). The lower whiskers describe the 1.5 interquartile range of the first quartile, the upper whisker stands for the 1.5 interquartile range of the third quartile, and the points indicate outliers. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 2). A thousand simulations were run for each SAD.

APPENDIX D

Ecological Archives EXXX-XXX-A4

COMPARISON OF CANONICAL ORDINATION MODELS FOR ABUNDANCE AND PRESENCE-ABSENCE DATA USING SIMULATIONS. FOUR FIGURES (FIGS. D1, D2, D3, AND D4)

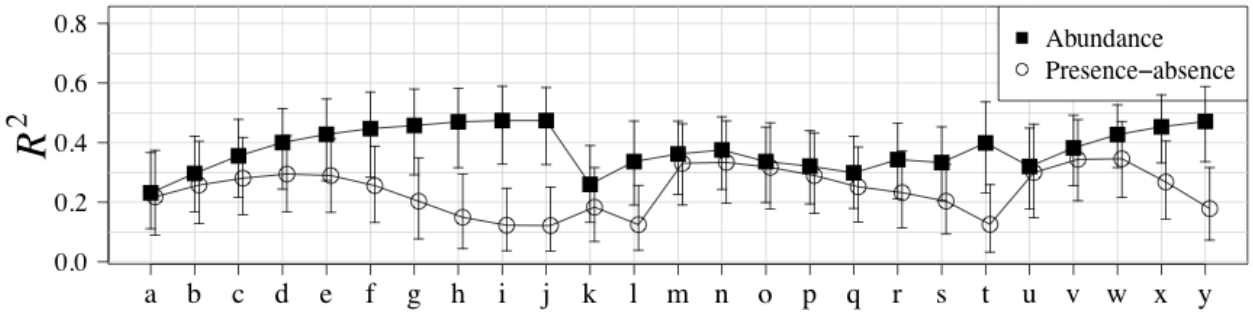


Fig. D1. RV coefficients (points) between canonical ordination model and the true species structure (equation 6 without the error term). For each data type (abundance and presence-absence), the significant canonical axes for all association coefficients (with the exception of the symmetric coefficient) were grouped. Error bars represent 95% confidence intervals. Letters on the x-axis refer to the species-abundance distribution (SAD) presented in Fig. 1. A line was drawn between each SAD of each association coefficient to ease comparisons between coefficients. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 0.25). A thousand simulations were run for each SAD.

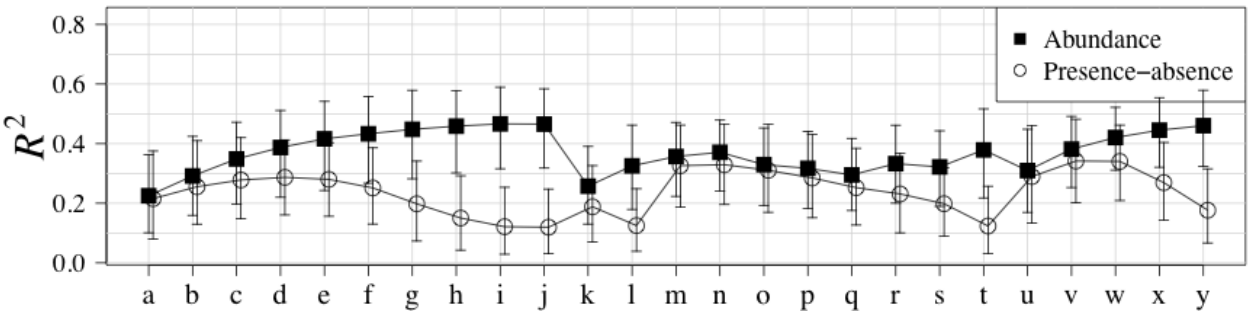


Fig. D2. RV coefficients (points) between canonical ordination model and the true species structure (equation 6 without the error term). For each data type (abundance and presence-absence), the significant canonical axes for all association coefficients (with the exception of the symmetric coefficient) were grouped. Error bars represent 95% confidence intervals. Letters on the x-axis refer to the species-abundance distribution (SAD) presented in Fig. 1. A line was drawn between each SAD of each association coefficient to ease comparisons between coefficients. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 0.5). A thousand simulations were run for each SAD.

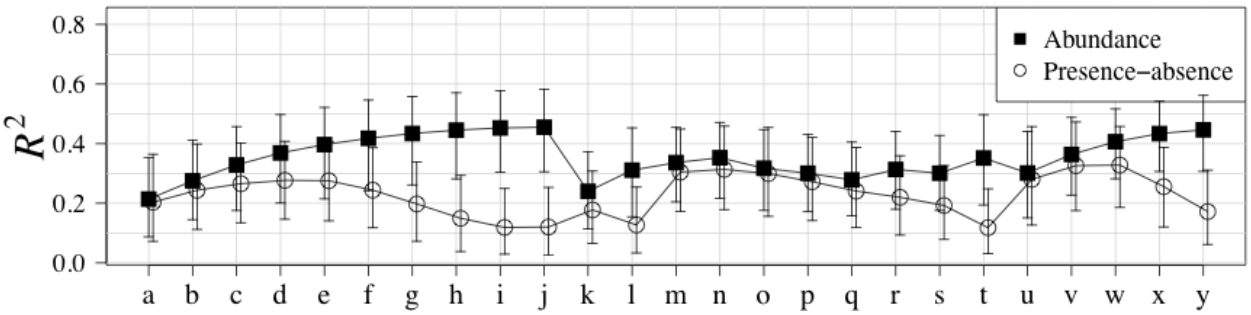


Fig. D3. RV coefficients (points) between canonical ordination model and the true species structure (equation 6 without the error term). For each data type (abundance and presence-absence), the significant canonical axes for all association coefficients (with the exception of the symmetric coefficient) were grouped. Error bars represent 95% confidence intervals. Letters on the x-axis refer to the species-abundance distribution (SAD) presented in Fig. 1. A line was drawn between each SAD of each association coefficient to ease comparisons between coefficients. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 1). A thousand simulations were run for each SAD.

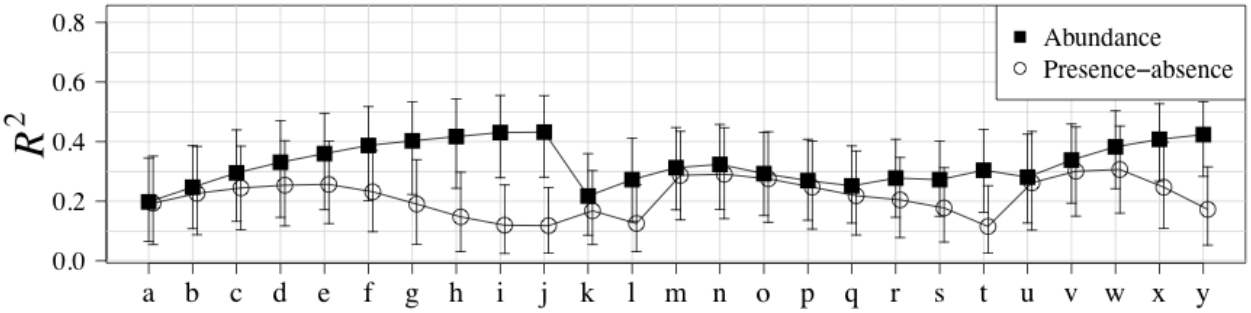


Fig. D4. RV coefficients (points) between canonical ordination model and the true species structure (equation 6 without the error term). For each data type (abundance and presence-absence), the significant canonical axes for all association coefficients (with the exception of the symmetric coefficient) were grouped. Error bars represent 95% confidence intervals. Letters on the x-axis refer to the species-abundance distribution (SAD) presented in Fig. 1. A line was drawn between each SAD of each association coefficient to ease comparisons between coefficients. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 2). A thousand simulations were run for each SAD.

APPENDIX E

Ecological Archives EXXX-XXX-A5

SPECIES CODE AND NAMES FOR CARABIDAE AND TREES SPECIES TWO TABLES (TABLES E1 AND E2)

Table D1: Species code and Latin name for Carabidae.

Code	Latin name
Agongrat	*Agonum gratiosum*
Agonplac	*Agonum placidum*
Agonretr	*Agonum retractum*
Agonsord	*Agonum sordens*
Agonsupe	*Agonum superioris*
Amarlitt	*Amara littoralis*
Amarluni	*Amara lunicollis*
Badiobtu	*Badister obtusus*
Bembgrap	*Bembidion grapii*
Bembrupi	*Bembidion rupicola*
Calaadve	*Calathus advena*
Calaingr	*Calathus ingratus*
Calofrig	*Calosoma frigidum*
Caracham	*Carabus chamissonis*
Dichcogn	*Dicheirotrichus cognatus*
Elapamer	*Elaphrus americanus*
Elaplapp	*Elaphrus lapponicus*
Harpfulv	*Harpalus fulvilabris*
Loripili	*Loricera pilicornis*
Miscarct	*Miscodera arctica*
Nebrgyll	*Nebria gyllenhali*
Notibore	*Notiophilus borealis*
Notidire	*Notiophilus directus*
Patrfove	*Patrobus foveocollis*
Patrsept	*Patrobus septentrionis*
Platdece	*Platynus decentis*
Platmann	*Platynus mannerheimii*
Pteradst	*Pterostichus adstrictus*
Pterbrev	*Pterostichus brevicornis*
Pterpens	*Pterostichus pensylvanicus*
Pterpunc	*Pterostichus punctatissimus*
Pterripa	*Pterostichus riparius*
Seriquad	*Sericoda quadripunctata*
Sterhaem	*Stereocerus haematopus*
Synuimpu	*Synuchus impunctatus*
Trecapic	*Trechus apicalis*
Trecchal	*Trechus chalybeus*

Table D2: Species code, common and Latin name of trees species.

Code	Common name	Latin name
Pt	Aspen	*Populus tremuloides*
Bp	White birch	*Betula papyrifera*
Ab	Balsam fir	*Abie balsamea*
Ll	Tamarack	*Larix laricina*
Pb	Balsam poplar	*Populus balsamifera*
Pc	Lodgepole pine	*Pinus contorta*
Pm	Black spruce	*Picea mariana*
Pg	White spruce	*Picea glauca*