*Ecology*, 87(10), 2006, pp. 2614–2625


# VARIATION PARTITIONING OF SPECIES DATA MATRICES: ESTIMATION AND COMPARISON OF FRACTIONS

Pedro R. Peres-Neto,[1] Pierre Legendre, Stéphane Dray, and Daniel Borcard

*Département des sciences biologiques, Université de Montréal, C.P. 6128, succursale Centreville, Montréal, Québec H3C 3J7 Canada*

*Abstract*. Establishing relationships between species distributions and environmental characteristics is a major goal in the search for forces driving species distributions. Canonical ordinations such as redundancy analysis and canonical correspondence analysis are invaluable tools for modeling communities through environmental predictors. They provide the means for conducting direct explanatory analysis in which the association among species can be studied according to their common and unique relationships with the environmental variables and other sets of predictors of interest, such as spatial variables. Variation partitioning can then be used to test and determine the likelihood of these sets of predictors in explaining patterns in community structure. Although variation partitioning in canonical analysis is routinely used in ecological analysis, no effort has been reported in the literature to consider appropriate estimators so that comparisons between fractions or, eventually, between different canonical models are meaningful. In this paper, we show that variation partitioning as currently applied in canonical analysis is biased. We present appropriate unbiased estimators. In addition, we outline a statistical test to compare fractions in canonical analysis. The question addressed by the test is whether two fractions of variation are significantly different from each other. Such assessment provides an important step toward attaining an understanding of the factors patterning community structure. The test is shown to have correct Type I error rates and good power for both redundancy analysis and canonical correspondence analysis.

*Key words: adjusted coefficient of determination; bootstrap; canonical analysis; canonical correspondence analysis (CCA); ecological community; redundancy analysis (RDA); variation partitioning.*

## Introduction

The search for causes dictating patterns in species distributions in natural and disturbed landscapes is of primary importance in ecological science, and establishing relationships between species distributions and environmental characteristics is a widely used approach (e.g., Legendre and Fortin 1989, Jackson and Harvey 1993, Diniz-Filho and Bini 1996, Rodríguez and Lewis 1997, Jenkins and Buikema 1998, Boyce and McDonald 1999, Peres-Neto and Jackson 2001). Habitat models relating habitat characteristics and community structure (species occurrence or abundance) are expected to answer at least two questions. (1) How well is the distribution of a set of species explained by the given set of predictive variables? (2) Which variables are irrelevant or redundant in the sense of failing to strengthen the explanation of patterns after certain other variables have been taken into account? The first question relates to the predictive power of the model that can be used in conservation management, for questions such as estimating habitat suitability, forecasting the effects of habitat change due to human interference, establishing potential locations for species reintroduction, or predicting how community structure may be affected by the invasion of exotic species. The second question is important for heuristic issues such as determining the likelihood of competing hypotheses to explain particular patterns in community structure (Peres-Neto et al. 2001).

Canonical analyses such as redundancy analysis (RDA; Rao 1964), canonical correspondence analysis (CCA; ter Braak 1986), and distance-based redundancy analysis (db-RDA; Legendre and Anderson 1999) are invaluable tools for modeling communities through environmental predictors. They provide the means for conducting direct explanatory analyses in which the association among species can be studied with respect to their common and unique relationships with environmental variables or any other set of predictors of interest. As a demonstration of its success, well over 1500 studies applying CCA or RDA in modeling species–environment relationships have been published (see also Birks et al. [1996] for reviews on ecological studies using these methods). RDA and CCA can be best understood as methods for extending multiple regression that has a single response $y$ and multiple predictors $\mathbf{X}$ (e.g., several environmental predictors), to

Manuscript received 26 August 2005; revised 9 February 2006; accepted 16 February 2006; final version received 21 March 2006. Corresponding Editor: N. G. Yoccoz.

[1] Present address: Department of Biology, University of Regina, Saskatchewan S4S 0A2 Canada.
E-mail: pedro.peres-neto@uregina.ca

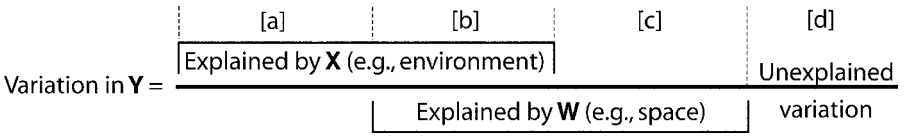


FIG. 1. Variation partitioning scheme of a response variable **Y** between two sets of predictors **X** (e.g., environmental factors) and **W** (e.g., spatial predictors). The total variation in **Y** is partitioned into four fractions as follows: (1) fraction [a + b + c] based on both sets of predictor matrices [**X**,**W**] ([a + b + c] = $R^2_{\mathbf{Y}|[\mathbf{X},\mathbf{W}]}$); (2) fraction [a + b] based on matrix **X** ([a + b] = $R^2_{\mathbf{Y}|\mathbf{X}}$); (3) fraction [b + c] based on matrix **W** ([b + c] = $R^2_{\mathbf{Y}|\mathbf{W}}$); (4) the unique fraction of variation explained by **X**, [a] = [a + b + c] − [b + c]; (5) the unique fraction of variation explained by **W**, [c] = [a + b + c] − [a + b]; (6) the common fraction of variation shared by **X** and **W**, [b] = [a + b + c] − [a] − [c]; and (7) the residual fraction of variation not explained by **X** and **W**, [d] = 1 − [a + b + c].

multiple regression involving multiple response variables **Y** (e.g., several species) and a common matrix of predictors **X**. It follows that the percentage of variation of the response matrix explained by the predictor matrix (hereafter referred to as the redundancy statistic, or simply $R^2_{\mathbf{Y}|\mathbf{X}}$ following Miller and Farr 1971) is the canonical equivalent of the regression coefficient of determination, $R^2$.

In multiple regression analysis, we can apply variation partitioning (also known as commonality analysis; Kerlinger and Pedhazur 1973) to identify common and unique contributions to model prediction and hence better address the question of the relative influences of the groups of independent variables considered in the model (Mood 1969). When partitioning variation is used in regression analysis, independent variables are grouped into sets representing broad factors. In that context, variation partitioning is more suitable than analyzing the individual contributions of regressors via their partial correlation coefficients. In this approach, the total percentage of variation explained by the model ($R^2$) is partitioned into unique and common contributions of the sets of predictors (Fig. 1). For example, variation partitioning for RDA or CCA using two sets of predictors (**X** and **W**) is straightforward as it is based on three canonical analysis (Fig. 1). The first one uses both sets of predictors [**X**,**W**], the second only **X**, and the last one only **W**. All remaining fractions of the partitioning can be obtained by simple subtractions (Fig. 1). Note that the shared variation ([b], Fig. 1) may be negative due to suppressor variables (i.e., a regressor having low, close to zero, correlation with the response variable and a correlation with another regressor, which in turn is correlated with the response variables; see Azen and Budescu [2003] for more details) or due to two strongly correlated predictors with strong effects on *y* of opposite signs (one positive and the other negative; Legendre and Legendre 1998: Section 10.3.5). Variation partitioning based on two sets of predictor matrices was introduced to canonical analysis by Borcard et al. (1992) and Borcard and Legendre (1994), later was extended to three or more sets of predictor matrices (Anderson and Gribble 1998, Cushman and McGarigal 2002, Økland 2003), and is now routinely used in direct gradient analysis.

Although canonical analysis and variation partitioning may provide a robust approach for understanding the relative influence of different ecological factors driving community assembly, judging the importance of a factor solely on the basis of its proportional unique contribution is not as straightforward as currently performed. The statistical bias related to estimating a population $\rho^2$ based on a sample $R^2$ is a well-recognized problem (Zar 1999), as sample estimates tend, on average, to be larger than $\rho^2$. The bias is influenced by both the number of independent variables in the model and sample size (Kromrey and Hines 1995). Terms such as adjustment and "shrinkage" refer to the fact that a sample-estimated $R^2$ needs to be reduced in order to provide a more accurate estimate of $\rho^2$. By taking into account the appropriate degrees of freedom, the adjustment provides a way of comparing models with different numbers of predictors (e.g., model selection) and sample sizes. Given that $R^2$ and $R^2_{\mathbf{Y}|\mathbf{X}}$ are intrinsically related, the bias observed in multiple regression also exists in canonical analysis. Although variation partitioning in canonical analysis is routinely used in ecological analysis, no effort has been reported in the literature to consider appropriate estimators so that comparisons between fractions or eventually between different canonical models are meaningful. Specifically, our objective is twofold: (1) to provide adjustments for the bias in sample $R^2_{\mathbf{Y}|\mathbf{X}}$, and (2) to outline a statistical test to contrast partial effects in canonical analysis (i.e., compare fractions of variation).

## REDUNDANCY STATISTIC IN CANONICAL ANALYSIS

Here we present the formulation of the $R^2_{\mathbf{Y}|\mathbf{X}}$ statistic used in canonical analysis applied to species data matrices. In the case of RDA, $R^2_{\mathbf{Y}|\mathbf{X}}$ is calculated as follows:

$$R^2_{\mathbf{Y}|\mathbf{X}} = \frac{\text{trace}(\hat{\mathbf{Y}}'\hat{\mathbf{Y}})}{\text{trace}(\mathbf{Y}'_{\text{cent}}\mathbf{Y}_{\text{cent}})} = 1 - \frac{\text{trace}[(\mathbf{Y}_{\text{cent}} - \hat{\mathbf{Y}})'(\mathbf{Y}_{\text{cent}} - \hat{\mathbf{Y}})]}{\text{trace}(\mathbf{Y}'_{\text{cent}}\mathbf{Y}_{\text{cent}})} \quad (1)$$

where $\hat{\mathbf{Y}} = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_{\text{cent}}$ represents the matrix of predicted values. Note that this is identical to calculating predicted values for individual multiple regressions of each column of **Y** on **X**; $\mathbf{Y}_{\text{cent}} = (\mathbf{I} - \mathbf{P})\mathbf{Y}$ is matrix **Y**

centered by column means (i.e., column means = 0). **I** is an $(n \times n)$ identity matrix and **P** is a $(n \times n)$ matrix with all elements = $1/n$; $n$ refers to the number of sampling units. Matrix **X** can be either centered or standardized (column means = 0 and column variances = 1).

The definition of $R^2_{\mathbf{Y}|\mathbf{X}}$ presented here is the one used in ecological applications; it is called the RDA trace statistic in the Canoco program, Version 4.5 (ter Braak and Smilauer 2002) and the proportion of explained variation in Legendre and Legendre (1998). This definition is different from the one in redundancy analysis as used in behavioral research (Dawson-Saunders 1982, Lambert et al. 1988) where the response variables are standardized rather than centered prior to analysis. In that case, the $R^2_{\mathbf{Y}|\mathbf{X}}$ is simply the mean of the $R^2$ statistics computed for each individual multiple regression of each column of **Y** on matrix **X** (Miller 1975). In ecological analysis the species are centered and not standardized, so $R^2_{\mathbf{Y}|\mathbf{X}}$ is a weighted mean of the $R^2$ of individual models with weights proportional to the species variances divided by the total variance. The same definition based on a weighted mean applies to CCA, and for the sake of brevity we present the $R^2_{\mathbf{Y}|\mathbf{X}}$ used in CCA in Appendix A.

## An Adjusted Redundancy Statistic for Canonical Analysis—The Continuous Case

Our first task was to determine whether adjustments for the multiple coefficient of determination ($R^2_{\text{adj}}$) developed for a single response variable could also be applied to the canonical $R^2_{\mathbf{Y}|\mathbf{X}}$. Dawson-Saunders (1982) has shown that Ezekiel's adjustment (1930), commonly used in the case of multiple regressions (Legendre and Legendre 1998, Zar 1999), is appropriate for the case where response variables are standardized prior to analysis. Ezekiel's formulation applied to canonical analysis based on centered values is as follows:

$$R^2_{(\mathbf{Y}|\mathbf{X})\text{adj}} = 1 - \frac{n-1}{n-p-1}(1 - R^2_{\mathbf{Y}|\mathbf{X}}) = 1 - \frac{\text{trace}[(\mathbf{Y}_{\text{cent}} - \hat{\mathbf{Y}})(\mathbf{Y}_{\text{cent}} - \hat{\mathbf{Y}})]/(n-p-1)}{\text{trace}(\mathbf{Y}'_{\text{cent}}\mathbf{Y}_{\text{cent}})/(n-1)} \quad (2)$$

where $n$ is the sample size, $p$ is the number of predictors, and $R^2_{\mathbf{Y}|\mathbf{X}}$is the sample estimation of the $\rho^2_{\mathbf{Y}|\mathbf{X}}$.

Since fractions of variation represent redundancy statistics, they also need to be adjusted. Fractions [a + b + c], [b + c], and [a + b] can be adjusted directly, leading to $[a + b + c]_{\text{adj}}$, $[b + c]_{\text{adj}}$, and $[a + b]_{\text{adj}}$. The individual fractions $[a]_{\text{adj}}$, $[b]_{\text{adj}}$, $[c]_{\text{adj}}$, and $[d]_{\text{adj}}$ have to be calculated by appropriate subtractions based on $[a + b + c]_{\text{adj}}$, $[b + c]_{\text{adj}}$, and $[a + b]_{\text{adj}}$.

We conducted a Monte Carlo study equivalent to the one used by Kromrey and Hines (1995) who assessed the accuracy of different methods for adjusting sample $R^2$ in the univariate multiple regression case. The first step was to generate large population matrices (200 000 individuals) with known $\rho^2_{\mathbf{Y}|\mathbf{X}}$ and then draw a large number of samples with replacement from these populations and calculate $R^2_{\mathbf{Y}|\mathbf{X}}$and $R^2_{(\mathbf{Y}|\mathbf{X})\text{adj}}$ for each sample. We decided to use large generated populations instead of standard protocols such as generating samples using established correlation matrices (see Peres-Neto et al. 2003 for an example) or by defining the $\rho^2$ as in the method introduced by Cramer (1987). The reason is that these previously used methods are capable of generating population values only for continuous variables; hence we cannot generate species-like data (e.g., abundance) where some sites are occupied (values > 0) and others are not (values = 0).

We started with a real data set comprised of stream fish communities of a watershed in eastern Brazil (Peres-Neto 2004). A total of 27 species and six environmental variables were considered. The first step was to calculate individual slopes between the species and the environmental variables (slopes are presented in Appendix B: Table B1). Slopes were calculated on the centered species and environmental matrices and used as the basis for our simulation study. The next step was to generate a matrix **X** containing six random normally distributed variables $\mathcal{N}(0,1)$ with 200 000 observations (rows). The columns of **X** were then standardized (i.e., mean = 0 and variance = 1). Then, a data matrix **Y** was generated as: $\mathbf{Y} = \mathbf{XB}\text{mlt} + \mathbf{E}$, where **B** (Appendix B: Table B1) is a $(6 \times 27)$ matrix containing the slopes for each species on each environmental variable; mlt is a multiplication factor used to reduce the slopes so that we can manipulate them to attain the desirable $R^2_{\mathbf{Y}|\mathbf{X}}$ values. The multiplication factor will be given for each simulated population. **E** represents a $(200\,000 \times 27)$ matrix containing $\mathcal{N}(0,1)$ deviates. The last step was to calculate the $\rho^2_{\mathbf{Y}|\mathbf{X}}$ based on the generated **X** $(200\,000 \times 6)$ and **Y** $(200\,000 \times 27)$ matrices. Since all slopes were different from zero, all predictors were active in the sense that they all contributed to the explanation of matrix **Y**.

### *Assessing the accuracy of $R^2_{(\mathbf{Y}|\mathbf{X})\text{adj}}$ using a single set of predictors (canonical analysis)*

The first set of simulations considered the simplest case of matrices **X** and **Y** made of continuous data; abundance-like data will be considered later. In the two sets of simulations, we considered the influence of random predictors by manipulating the number of random $\mathcal{N}(0,1)$ variables added to the set of true predictors **X**, as well as the sample size $n$. Two populations with $\rho^2_{\mathbf{Y}|\mathbf{X}} = 0.2007$ (mlt = 0.0004) and $\rho^2_{\mathbf{Y}|\mathbf{X}} = 0.6105$ (mlt = 0.001) were considered. In the first set of simulations, 1000 samples of 100 observations each were randomly drawn from the population [**Y**,**X**] and a certain number of random $\mathcal{N}(0,1)$ variables were added to the sample **X**. In the second experiment, 1000 samples with varying numbers of observations were randomly drawn and no random predictors were added to the model. Fig. 2 presents the results of the two

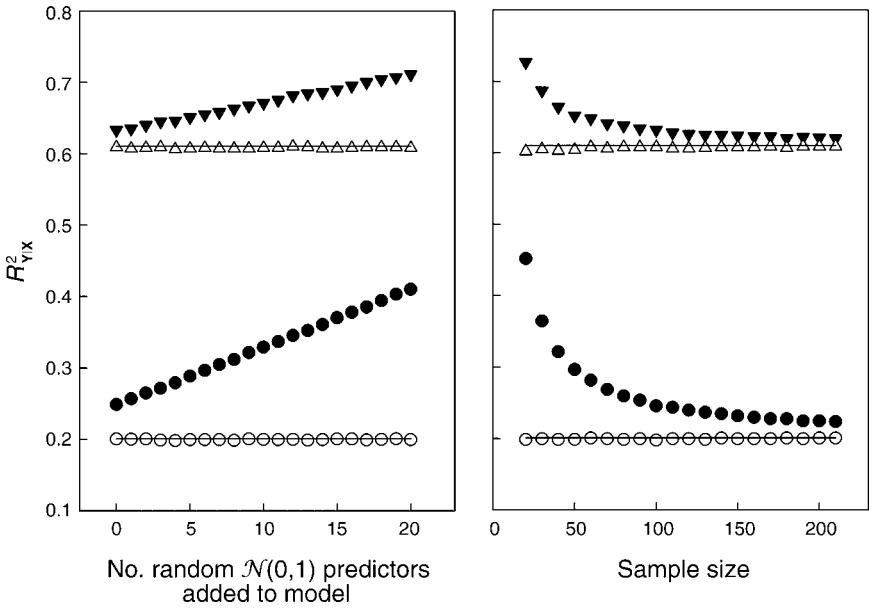


FIG. 2. The influence of null predictors and sample size on the sample mean $R^2_{\mathbf{Y}|\mathbf{X}}$ (solid symbols) and the mean adjusted $R^2_{(\mathbf{Y}|\mathbf{X})\text{adj}}$ (open symbols) considering two RDA populations with normally distributed data. Triangles represent samples from a population $R^2_{\mathbf{Y}|\mathbf{X}} = 0.608$, whereas circles represent samples from a population $R^2_{\mathbf{Y}|\mathbf{X}} = 0.201$. Horizontal lines represent population values. In the case of the influence of the number of null predictors (left panel), samples were based on 100 observations, whereas in the case of the influence of sample size (right panel), no null predictor was added to the model.

simulations; each point in the graphs represents mean values of sample $R^2_{\mathbf{Y}|\mathbf{X}}$ and $R^2_{(\mathbf{Y}|\mathbf{X})\text{adj}}$ based on 1000 random pairs of matrices [**Y**,**X**]. It is evident from these two simulations that the sample $\rho^2_{\mathbf{Y}|\mathbf{X}}$ statistic is highly biased and that $R^2_{(\mathbf{Y}|\mathbf{X})\text{adj}}$ is a much more accurate estimator of the $R^2_{\mathbf{Y}|\mathbf{X}}$. Results (Fig. 2) show that adjusted statistics are always to be preferred to $R^2_{\mathbf{Y}|\mathbf{X}}$ sample values.

### *Assessing the accuracy of $R^2_{(\mathbf{Y}|\mathbf{X})\text{adj}}$ on variation partitioning (partial canonical analysis)*

We conducted a second set of simulations to test whether $R^2_{(\mathbf{Y}|\mathbf{X})\text{adj}}$ also performed well in variation partitioning in canonical analysis involving three data matrices [**Y**,**X**,**W**]. The method used for generating population matrices was basically the same, except that we considered the case where predictors have a certain level of correlation between matrices **X** and **W**. In order to generate populations having predictors **X** and **W** with a given level of correlation (i.e., [b] $> 0$], the following procedure was applied. (1) Generate a matrix **XW** containing 12 random normally distributed variables $\mathcal{N}(0,1)$ with 200 000 observations. As before, the columns of matrix **XW** were standardized. (2) Generate a $(12 \times 12)$ correlation matrix where all cross-correlation values were 0.1. Next, decompose the correlation matrix using Cholesky decomposition. Finally, post-multiply the upper-triangular matrix resulting from the matrix factorization by the matrix **XW** of step 1. (3) The matrix of population slopes $\mathbf{B_{XW}}$ $(12 \times 27)$ is constructed by assembling two slope matrices **B** $(6 \times 27)$ (Appendix B: Table B1). Each one was multiplied by a different multiplication factor, mlt (i.e., $\mathbf{B_{XW}} = [\mathbf{B}\text{mlt}_\mathbf{X}\ \mathbf{B}\text{mlt}_\mathbf{W}]$), so that the relative contributions of **X** and **W** to the generated **Y** are different. (4) As before, a data matrix **Y** was generated as: $\mathbf{Y} = \mathbf{XWB_{XW}} + \mathbf{E}$. The first six columns of matrix **XW** were used to represent matrix **X** and the last six columns became matrix **W**. The population was then represented by juxtaposing the three pertinent matrices as [**Y**,**X**,**W**].

Simulations were also done to evaluate the influence of random predictors and the sample size. Population fractions based on $\text{mlt}_\mathbf{X} = 0.0004$ and $\text{mlt}_\mathbf{W} = 0.0003$ were as follows: [a + b + c] = 0.3968, [a + b] = 0.3001, [b + c] = 0.2243, [a] = 0.1725, [b] = 0.1276, [c] = 0.0967, and [d] = 0.6032. Matrix **X** contributed to **Y** with a larger portion of the variation in the population (30.01%) than **W** (22.43%), and fraction [b] explained 12.76% of the variation in the population. In the third series of simulations, 1000 samples of 100 observations each were drawn from the population [**Y**,**X**,**W**], and random $\mathcal{N}(0,1)$ variables were added to the sample matrix **W**. In the fourth set of simulations, 1000 samples of a varying numbers of observations were drawn. No random predictors were added to the model. Fig. 3 presents the variation partitioning results from these two last simulations. Each fraction represents mean values of sample fractions and adjusted sample fractions based on the 1000 samples. It is obvious from these results that sample fraction estimates (left panels) are highly biased whereas adjusted fractions (right panels) are much more accurate estimators of the

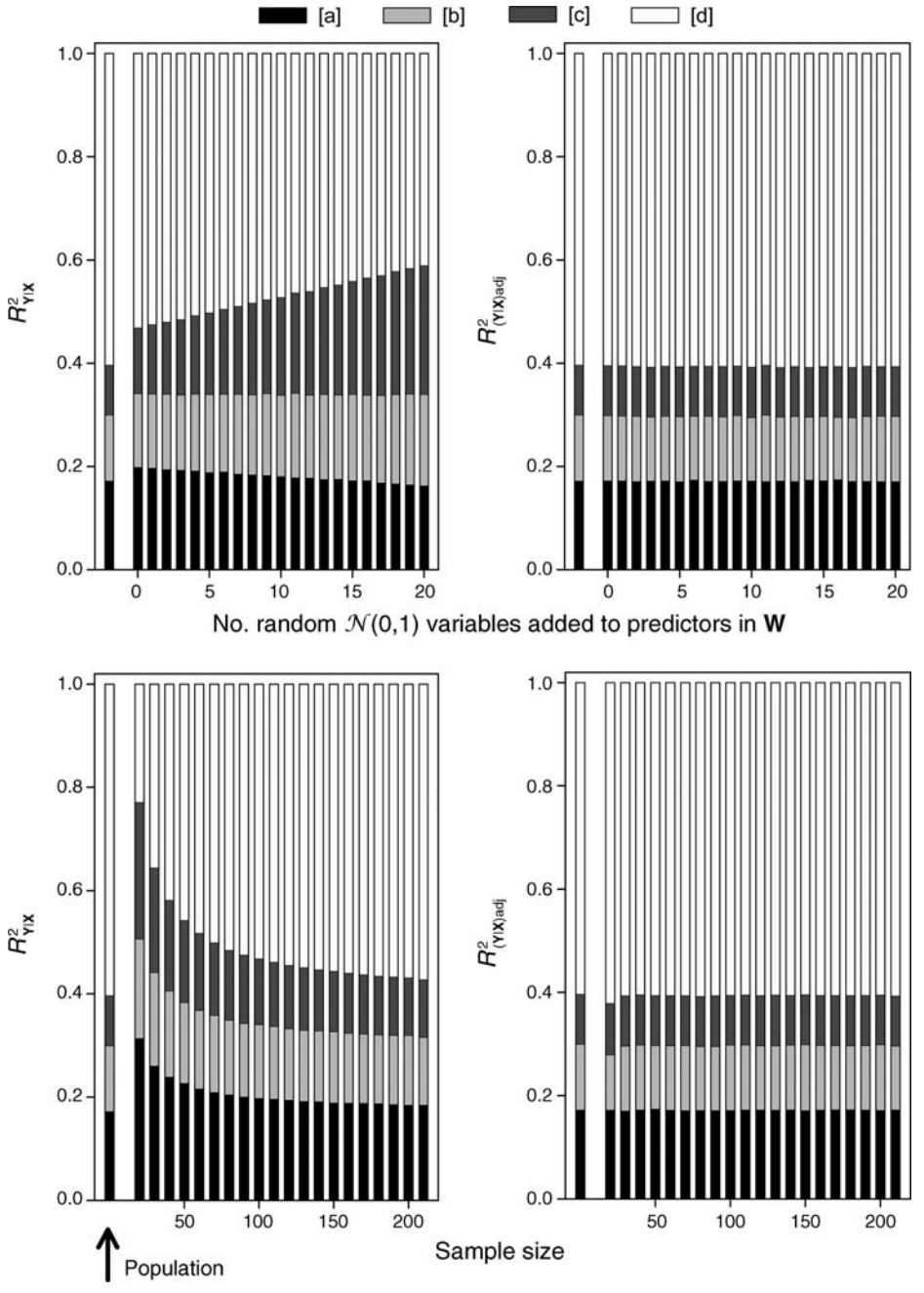


FIG. 3. The influence of null predictors and sample size on fraction estimation in RDA variance partitioning, considering normally distributed data. Left panels, sample $R^2_{\mathbf{Y}|\mathbf{X}}$; right panels, adjusted $R^2_{(\mathbf{Y}|\mathbf{X})\text{adj}}$. In the case of the influence of the number of null predictors (upper panels), samples were based on 100 observations, whereas in the case of the influence of sample size (lower panels), only active predictors were used (i.e., no null predictor was added to the model). In that case, all predictors in **X** were active, whereas in **W** a mix of active and null predictors was found. The sequence is [a], [b], [c], [d] from bottom to top of each panel.

population values. Although the population fraction [a] is larger than [c], the addition of random variables to matrix **W** offset the differences in sample fractions without adjustment, demonstrating clearly the need for an adjustment to obtain correct estimates of the importance of each matrix to the model.

## ADJUSTED REDUNDANCY STATISTIC FOR THE DISCRETE ABUNDANCE CASE

In the previous section, we considered the case of continuous variables as they represented the case where we expected a direct match between the multiple regression $R^2$ and the redundancy statistic $R^2_{\mathbf{Y}|\mathbf{X}}$. How-

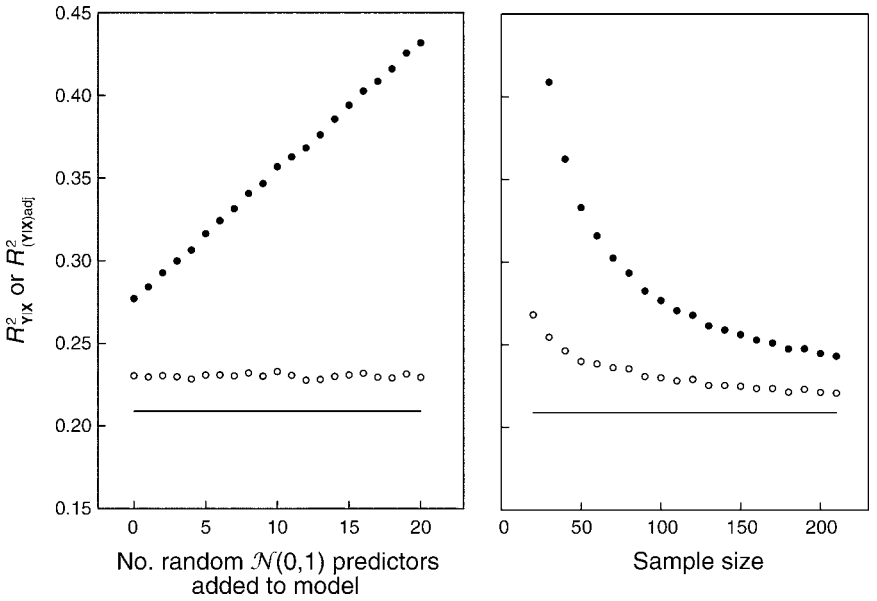


FIG. 4. The influence of null predictors and sample size on the sample mean $R^2_{\mathbf{Y}|\mathbf{X}}$ and the adjusted mean $R^2_{(\mathbf{Y}|\mathbf{X})\text{adj}}$ considering an RDA population ($R^2_{\mathbf{Y}|\mathbf{X}} = 0.2089$) with abundance-like dependent variables. Solid circles represent sample values, and open circles represent adjusted values. Horizontal lines represent population values.

ever, ecologists are most interested in the case where response variables are counts of species abundances, which are discrete and also generally overdispersed and zero inflated (Martin et al. 2005). In order to generate discrete and zero inflated data, we transformed a simulated population matrix $\mathbf{Y}$ as follows: $\mathbf{Y}' = [y'_{ij}] = \exp(1.2(y_{\text{std}ij} - 0.5))$. Once generated, matrix $\mathbf{Y}$ was first standardized into $\mathbf{Y}_{\text{std}}$ so that a variance value of 1.2 could be applied; subtracting 0.5 from the data provided a greater number of zeros (absences; the generated matrices contained roughly 47% zeros). The values $y'_{ij}$ were then rounded to the lower integer to generate $\mathbf{Y}'$. A similar protocol was applied in Legendre et al. (2005). We performed simulations similar to those reported in the previous section, this time using population matrices $\mathbf{Y}'$. We started by investigating the case of RDA. In order to generate $\mathbf{Y}$, a value mlt = 0.0011 was necessary to generate a $\rho^2_{\mathbf{Y}|\mathbf{X}} = 0.2089$, which is close to the value generated in the first set of simulations (see *An adjusted redundancy statistic for canonical analysis—the continuous case*). Fig. 4 shows the results as a function of the number of random variables added to sample predictor matrices $\mathbf{X}$ and the sample size. As before, in the case of random predictors, 100 observations were considered. Although the adjusted values provided better estimates than the sample $R^2_{\mathbf{Y}|\mathbf{X}}$, the estimates were quite biased when compared to the $\rho^2_{\mathbf{Y}|\mathbf{X}}$.

Next we investigated the effect of species transformations in the estimates prior to RDA. We used the Hellinger transformation (Legendre and Gallagher 2001) as follows:

$$\mathbf{H} = [h_{ij}] = \left[\sqrt{\frac{y'_{ij}}{\sum_{j=1}^{k} y'_{ij}}}\right] \tag{3}$$

where $h_{ij}$ is the transformed abundance $y'_{ij}$ of species $j$ at site $i$, and $k$ is the total number of species across all replicates. The use of Hellinger-transformed data prior to performing RDA is equivalent to a distance-based redundancy analysis (db-RDA; Legendre and Anderson 1999) based on Hellinger distance. In the present simulations, the Hellinger-transformed population matrix $\mathbf{H}$ was used instead of $\mathbf{Y}'$. Two Hellinger-transformed population data sets with $\rho^2_{\mathbf{H}|\mathbf{X}} = 0.2071$ (mlt = 0.0007) and $\rho^2_{\mathbf{H}|\mathbf{X}} = 0.6099$ (mlt = 0.0500) were generated. Contrary to RDA on raw species abundance-like data (Fig. 4), the Hellinger transformation produced much more accurate estimates of $R^2_{\mathbf{H}|\mathbf{X}}$ regarding both sample size and number of random predictors added to the sample matrices (see Appendix C: Fig. C1). The same conclusions were attained when considering variation partitioning as results showed identical patterns as the ones depicted in Fig. 3 for the continuous case (see Appendix C: Fig. C2). Thus, for continuous response variables, Ezekiel's adjustment to RDA $R^2$ gives quite accurate values, and for the discrete zero inflated data with many zeros (as simulated here), Ezekiel's adjustment using Hellinger-transformed data (i.e., db-RDA on Hellinger distance) is appropriate.

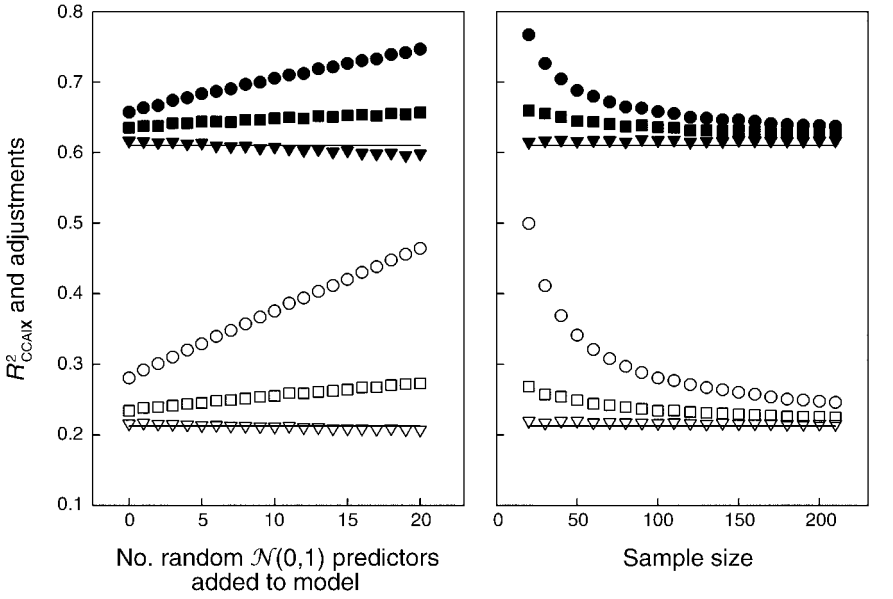


FIG. 5. The influence of null predictors and sample size on the sample mean $R^2_{CCA|X}$ (circles), the mean adjusted $R^2_{adj}$ (squares), and the permutation-based adjusted $R^2_{adj-perm}$ (triangles) considering two CCA populations. Solid symbols in each panel represent samples from a population with $R^2_{CCA|X} = 0.610$, while open symbols represent samples from a population with $R^2_{CCA|X} = 0.213$. Horizontal lines represent population values. In the case of the influence of the number of null predictors (left panels), samples were based on 100 observations, whereas in the case of the influence of sample size (right panels), only active predictors were used (i.e., no null predictor was added to the model).

## The Case of CCA and a Permutational Form of Adjustment

Here we present the results of simulations using CCA. As in the two previous sections, we assessed the influence of sample size and number of random predictors added to the model. In the case of random predictors, sample sizes of 100 observations were drawn. Two CCA populations with $\rho^2_{CCA|X} = 0.2128$ (mlt = 0.0011) and $\rho^2_{CCA|X} = 0.6101$ (mlt = 0.0490) were considered. As in the preceding section, values in matrix **Y** were transformed into abundance-like values (**Y**′) after generation and prior to CCA. Ezekiel's adjustments did improve the large bias obtained when using unadjusted $R^2$ with CCA; however they were still quite inaccurate, especially for the population having $\rho^2_{CCA|X} = 0.2128$ (Fig. 5). We propose a new type of adjustment based on the definition of Ezekiel's formulation. Ezekiel's adjustment is based on the idea that, in regression models, a random predictor explains, on average, under random sampling variation, $1/(n - 1)$ of the variation in the response variable. Hence $p$ random predictors explain, on average, under random sampling variation, $p/(n - 1)$ of the variation. The adjustment can be rewritten as

$$R^2_{(Y|X)adj} = 1 - \frac{n-1}{n-p-1}(1 - R^2_{Y|X}) = 1 - \frac{1}{1 - \frac{p}{n-1}}(1 - R^2_{Y|X}). \quad (4)$$

Note that when $R^2_{Y|X}$ equals $p/(n - 1)$, $R^2_{(Y|X)adj}$ equals zero. In CCA, because a weighted multiple regression is used (see Appendix A), the average $R^2_{CCA|X}$ expected by chance is unknown. We propose that this may be estimated for a given situation using a permutation procedure. The new adjustment hereafter referred to as $R^2_{perm}$ substitutes $p/(n - 1)$ in Eq. 4 by an empirical estimate of the value expected under chance alone estimated as follows: (1) randomly permute entire rows of data matrix **X** (i.e., no substantial difference was found if regressors were permuted separately), leading to $\mathbf{X}_{perm}$; (2) Calculate $R^2_{CCA|X}$ for a CCA based on $\mathbf{X}_{perm}$; (3) repeat steps 1 and 2 $m$ times (in this study we used $m = 1000$); (4) calculate the mean $\overline{\mathbf{X}}_{perm}$ across all 1000 $R^2_{CCA|X}$ obtained under permutation in step 3. The proposed adjustment is then simply

$$R^2_{perm} = 1 - \frac{1}{1 - \overline{\mathbf{X}}_{perm}}(1 - R^2_{CCA|X}). \quad (5)$$

The correction provides improved estimations in comparison to Ezekiel's adjustment (Fig. 5). The CCA code used here can be used on matrices containing huge number of observations (200 000 observations or more; see Supplement 1 code for CCA) and may be useful for researchers interested in analyzing large data sets.

## Testing the Difference between Fractions

Although the adjustment of fractions is an important step toward achieving unbiased canonical models, there

remains the question of whether the influences of two or more groups of predictors (factors), after adjustments, are significantly different. Methods for testing the significance of fractions are well established in the literature as in Legendre and Legendre (1998:608–612). However, one aspect that has not been addressed yet is whether two fractions, say [a] and [c], come from the same statistical population of explained fractions of variation and that they only differ by sampling variation. Such an assessment would provide an important step toward attaining an understanding of the factors patterning community structure. For instance, do environmental factors explain more variation than spatial patterning? In this section, we propose a method for testing for the difference between fractions in canonical variation partitioning. The test proposed here is based on a bootstrap procedure for empirically constructing sampling distributions reflecting the differences between adjusted $R^2_{\mathbf{Y}|\mathbf{X}}$. The use of bootstrapping as means of adjustment in multiple regressions has been advocated by Kromrey and Hines (1995), but our proposed bootstrap procedure provided a much better estimate than the one proposed by them. The proposed procedure for RDA is as follows.

(1) Compute the RDA residuals based on the original sample for fractions [a + b] and [b + c]:

$$\mathbf{E}_{\mathbf{X}} = \mathbf{Y}_{\text{cent}} - \mathbf{X}\mathbf{B}_{\mathbf{X}} \qquad \mathbf{E}_{\mathbf{W}} = \mathbf{Y}_{\text{cent}} - \mathbf{W}\mathbf{B}_{\mathbf{W}} \quad (6)$$

where $\mathbf{E}_{\mathbf{X}}$ and $\mathbf{E}_{\mathbf{W}}$ are $(n \times k)$ matrices of residuals related to the matrix of predictors $\mathbf{X}$ (i.e., [a + b]) and $\mathbf{W}$ (i.e., [b + c]), respectively; $n$ is the number of sites and $k$ is the number of species. $\mathbf{B}_{\mathbf{X}}$ and $\mathbf{B}_{\mathbf{W}}$ are the matrices of slopes for the predictors in $\mathbf{X}$ and $\mathbf{W}$, respectively, and can be calculated simply as: $\mathbf{B}_{\mathbf{X}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{Y}_{\text{cent}}$; $\mathbf{B}_{\mathbf{W}} = (\mathbf{W}'\mathbf{W})^{-1}\mathbf{W}'\mathbf{Y}_{\text{cent}}$. Note that $\mathbf{X}$ and $\mathbf{W}$ can be either centered or standardized.

(2) Rescale each column of matrices $\mathbf{E}_{\mathbf{X}}$ and $\mathbf{E}_{\mathbf{W}}$ as

$$\mathbf{E}_{\mathbf{X}\text{-scl}} = \sqrt{n/(n - p_{\mathbf{X}})}\mathbf{E}_{\mathbf{X}}$$
$$\mathbf{E}_{\mathbf{W}\text{-scl}} = \sqrt{n/(n - p_{\mathbf{W}})}\mathbf{E}_{\mathbf{W}} \quad (7)$$

where $p_{\mathbf{X}}$ and $p_{\mathbf{W}}$ are the number of predictors in matrices $\mathbf{X}$ and $\mathbf{W}$, respectively. Residuals are scaled in this fashion so that the average squared residuals in the bootstrap sample has expectation $\sigma^2$ (Wu 1986; i.e., $\mathbf{E} = [\boldsymbol{e}_j] \sim \mathbf{IID}(0, \sigma^2\mathbf{I})$) where $\mathbf{I}$ is a $(n \times 1)$ vector of 1's).

(3) Resample entire rows from the matrix of joint residuals [$\mathbf{E}_{\mathbf{X}\text{-scl}}$, $\mathbf{E}_{\mathbf{W}\text{-scl}}$] with replacement, so that the bootstrapped sample is consistent with the original dimensions of the data matrices leading to $\mathbf{E}_{\mathbf{X}\text{-boot}}$ and $\mathbf{E}_{\mathbf{W}\text{-boot}}$. Matrices of residuals were juxtaposed during the bootstrap sampling to make sure that the same rows were randomly chosen in $\mathbf{E}_{\mathbf{X}\text{-scl}}$ and $\mathbf{E}_{\mathbf{W}\text{-scl}}$, hence maintaining the covariance among predictors in $\mathbf{X}$ and $\mathbf{W}$ during resampling.

(4) Calculate bootstrapped $\mathbf{Y}$ data tables based on matrices $\mathbf{X}$ and $\mathbf{W}$ as the sum of original fitted values plus bootstrapped residuals:

$$\mathbf{Y}_{\mathbf{X}\text{-boot}} = \mathbf{X}\mathbf{B}_{\mathbf{X}} + \mathbf{E}_{\mathbf{X}\text{-boot}}$$
$$\mathbf{Y}_{\mathbf{W}\text{-boot}} = \mathbf{W}\mathbf{B}_{\mathbf{W}} + \mathbf{E}_{\mathbf{W}\text{-boot}}. \quad (8)$$

(5) Calculate fitted values based on the centered values of each set:

$$\hat{\mathbf{Y}}_{\mathbf{X}\text{-boot}} = \mathbf{X}\mathbf{B}_{\mathbf{X}}(\mathbf{I} - \mathbf{P})\mathbf{Y}_{\mathbf{X}\text{-boot}}$$
$$\hat{\mathbf{Y}}_{\mathbf{W}\text{-boot}} = \mathbf{W}\mathbf{B}_{\mathbf{W}}(\mathbf{I} - \mathbf{P})\mathbf{Y}_{\mathbf{W}\text{-boot}}. \quad (9)$$

Bootstrapped adjusted $R^2_{\mathbf{Y}|\mathbf{X}}$ based on each set of predictors were calculated as shown in Eq. 10 (at bottom of page). $R^2_{(\mathbf{Y}|\mathbf{X})\text{adj-boot}^i}$ was calculated in the same way by replacing $\mathbf{Y}_{\mathbf{X}\text{-boot}} - \hat{\mathbf{Y}}_{\mathbf{X}\text{-boot}}$ by $\mathbf{Y}_{\mathbf{W}\text{-boot}} - \hat{\mathbf{Y}}_{\mathbf{W}\text{-boot}}$. Note that the total sum of squares (TSS = $\text{trace}(\mathbf{Y}'_{\text{cent}}\mathbf{Y}_{\text{cent}})$) in Eq. 10 was divided by $n$ instead of $(n - 1)$ as in Eq. 2. Our decision was based on the fact that the maximum likelihood estimator of TSS provided better adjusted estimates under bootstrap (see Appendix D).

(6) Repeat steps 3–5 $m$ times ($m = 1000$ in this study). For each bootstrap replicate, calculate the difference between the two adjusted estimates as $D_i = R^2_{(\mathbf{Y}|\mathbf{X})\text{adj-boot}^i} - R^2_{(\mathbf{Y}|\mathbf{W})\text{adj-boot}^i}$. Then, using all bootstrapped $D_i$ values, build a confidence interval for the differences between adjusted fraction values. There are a number of procedures for estimating confidence intervals (Manly 1997); we used the percentile method (Manly 1997:39). First, $D_i$ values were ordered in ascending order, then we identified the $D_i$ values that occupied the $\alpha m/2$-th and $(1 - \alpha/2)m$-th values in the sorted list, which were then used as confidence limits estimates. Both $\alpha m/2$ and $(1 - \alpha/2)m$ were rounded to the nearest integer. Note that if confidence intervals are reported they can be helpful in comparing differences between fractions based on different data sets (e.g., two or more landscapes). We considered an $\alpha = 0.05$ significance level throughout this study. If the estimated interval did not encompass zero, then the null hypothesis was rejected. Alternatively, a $P$ value can also be estimated. First, calculate the median of $D_i$ values. If the median is positive, then calculate the number of $D_i$ smaller than zero; or alternatively if $D_{\text{obs}}$ is negative, calculate the number of $D_i$ larger than zero,

$$R^2_{(\mathbf{Y}|\mathbf{X})\text{adj-boot}^i} = [\text{a} + \text{b}]_{\text{adj}_i}$$
$$= 1 - \frac{\text{trace}\left(\left[(\mathbf{I} - \mathbf{P})\mathbf{Y}_{\mathbf{X}\text{-boot}} - \hat{\mathbf{Y}}_{\mathbf{X}\text{-boot}}\right]'\left[(\mathbf{I} - \mathbf{P})\mathbf{Y}_{\mathbf{X}\text{-boot}} - \hat{\mathbf{Y}}_{\mathbf{X}\text{-boot}}\right]\right)/(n - p_x - 1)}{\text{trace}\left(\mathbf{Y}'_{\text{cent}}\mathbf{Y}_{\text{cent}}\right)/n} \quad (10)$$

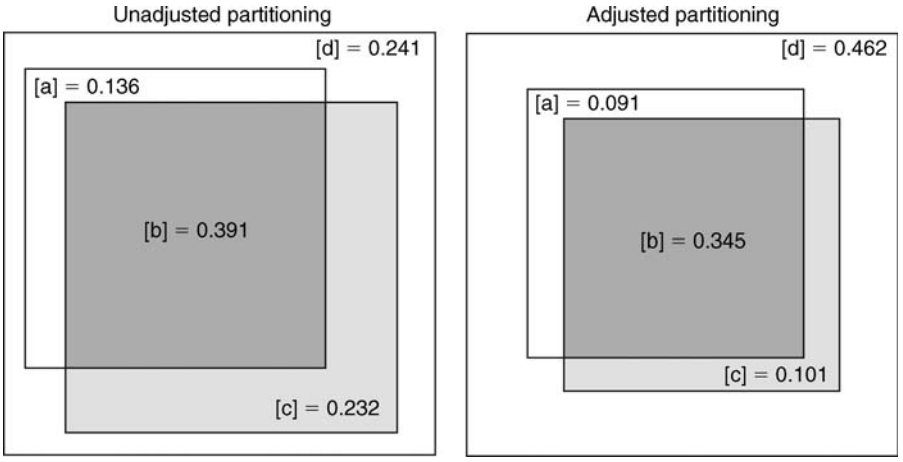


FIG. 6. Variation partitioning Venn diagrams representing the unadjusted (left) and adjusted (right) percentages of unique contribution of [c] spatial and [a] environmental components to the oribatid mite distribution. Fraction [b] represents the shared variation between the environmental and spatial components and [d] the residual variation left unexplained by the canonical model. Distance-based eigenvector maps were used to analyze the spatial component of the mite variation. Although fractions [a] and [c] were both significant in terms of explaining the variation in oribatid mite distribution, the proportions of variation explained by each do not differ significantly from each other (bootstrap test for fractions, $P = 0.4865$).

divided by the number of bootstrap samples, and then multiplied by two in order to make the test two tailed.

In order to assess the statistical robustness of the proposed test, we empirically estimated Type I error rates and power based on Monte Carlo simulations (Peres-Neto and Olden 2001). The complete description of the method involved in this assessment and the detailed results are presented in Appendix E. Type I error rates were either equal to or smaller than the established significance level for both the RDA based on the Hellinger transformation and CCA, hence validating the test (Appendix E).

## Variation Partitioning—An Example of Fraction Adjustment and Testing

In this section we present a complete example of the approaches introduced here, where we contrast results based on unadjusted and adjusted fractions in variation partitioning. Data were comprised of the abundances of 35 species of oribatid mites and five habitat variables from 70 soil cores $10 \times 2.6$ m in area in the peat blanket surrounding a bog lake originally presented by Borcard et al. (1992). Three of the five environmental variables are qualitative and were transformed into Helmert orthogonal contrasts as used in ANOVA (i.e., each $k$-level categorical variable was represented by $k - 1$ contrasts). A total of 11 environmental predictors were then used (nine Helmert contrasts representing qualitative variables and two quantitative variables). In order to generate spatial descriptors, we applied a distance-based eigenvector map (Dray et al. 2006) to the spatial coordinates of the cores. Here, the 22 eigenvectors with positive eigenvalues were retained as spatial descriptors to be used in the variation partitioning of the oribatid data. We applied variation partitioning by RDA to the matrix of Hellinger-transformed data. Results of variation partitioning based on adjusted and unadjusted fractions are presented in Fig. 6. Differences between adjusted and unadjusted fractions were quite noticeable, especially given the large number of spatial descriptors considered and when comparing the residual fraction [d] between the unadjusted and adjusted partitioning. Based on permutation tests (see Legendre and Legendre 1998:608–610), both fractions [a] and [c] explained a significant portion of the variation ($P[a] = 0.001$; $P[c] = 0.001$; 1000 permutations were applied). Based on our bootstrap procedure, however, the spatial and environmental components explained similar proportions of variation ($P = 0.4865$) of the species distribution.

## Discussion

Canonical analyses have become standard tools to analyze ecological community data in order to search for patterns and test hypotheses regarding species distributions and structuring factors. Although the issue of adjustment of the explained variation has been greatly stressed in univariate multiple regression modeling, it had not been discussed for canonical analysis applied to ecological data, especially for the case of variation partitioning. The present study brought attention to the importance of adjusting model explanation in canonical analysis and assessed the appropriateness of standard procedures used in regression analysis; it also offered a novel adjustment in the case of CCA. Our study demonstrated that sample $R^2_{\mathbf{Y}|\mathbf{X}}$ used in canonical analysis and variation partitioning is biased and that adjustments are not only preferable but necessary to provide more accurate estimations and valid comparisons between sets of factors in explaining community structure. Discrepancies between nonadjusted and ad-

justed fractions in variation partitioning will depend on the differences between the numbers of variables in each set of predictors. For instance, spatial regressors are always simpler to generate in comparison to environmental variables. As a result, the spatial fraction could appear more important due to a large number of unimportant spatial predictors considered in the canonical analysis. Indeed, in our example (Fig. 6), we found that the fraction due to the spatial component contributed with almost twice as much variation as the fraction due to the environmental component. Once adjusted, both fractions were quite similar and indeed the bootstrap test of fractions did not detect a significant difference between the environmental and spatial contributions. Adjustments will also permit comparisons between different canonical analyses as they adjust not only for the number of predictors but also for the number of samples. Cottenie (2005), comparing the contribution of environmental and spatial fractions based on 158 published data sets using canonical analysis, was forced to control for the potential bias in the estimation by using subsets of 30 sites and four spatial and environmental predictors in each analysis since appropriate adjustments were not yet available.

Our assessment did not consider the case of presence–absence data, other species transformations (Legendre and Gallagher 2001), and different types of distances such as Bray-Curtis that can be considered in distance-based RDA (Legendre and Anderson 1999). However, the simulation protocols applied here could be easily adapted to assess the accuracy of the adjustments considered here. The permutation approach used in the case of CCA can also be easily implemented in the case of other transformations prior to RDA and to different distances in the case of db-RDA.

The testing procedure for assessing the difference between two fractions of variation in canonical analysis is analogous to the comparison of nonnested models. There is a large body of literature dealing with tests for nonnested models for multiple regressions (Royston and Simpson 1995, Watnik et al. 2001), where sets of predictors are compared. The existing methods for comparing nonnested models, however, determine which set of predictors (model) is the most important in explaining the response variable. In the bootstrap test of fractions proposed here, we consider that the combination of the two sets of regressors is important in explaining the response variable, but that one set may be more important than the other in patterning the response variables (e.g., species). Current implementations for comparing nonnested models may identify whether two sets of predictors are significant or not, but not if the contribution of one set is significantly greater than the other. In addition, these implementations are only applicable to the case of multiple regressors having one response variable and are not implemented for the case of canonical analysis. Overall, we hope that the methods suggested will aid ecologists in attaining an understanding of the factors driving community structure.

A Matlab library and an executable version for conducting variation partitioning based upon adjusted $R^2$ values in RDA or CCA, and testing fractions, are available in Supplement 2 and Supplement 3, respectively. The R-language function "varpart," available in the vegan library (Version 1.7-81 or higher; Oksanen et al. 2005), automatically conducts variation partitioning of a response table with respect to two, three, or four tables of explanatory variables, using RDA-adjusted $R^2$ values.

Acknowledgments

This research was supported by NSERC Grant No. OGP0007738 to P. Legendre. This manuscript was greatly improved by the comments provided by Marti Anderson, Helene Wagner, and an anonymous reviewer. We thank Einar Heegaard for valuable discussions.

Literature Cited

Anderson, M. J., and N. A. Gribble. 1998. Partitioning the variation among spatial, temporal and environmental components in a multivariate data set. Australian Journal of Ecology **23**:158–167.

Azen, R., and D. V. Budescu. 2003. The dominance analysis approach for comparing predictors in multiple regression. Psychological Methods **8**:129–148.

Birks, H. J. B., S. M. Peglar, and H. A. Austin. 1996. An annotated bibliography of canonical correspondence analysis and related constrained ordination methods 1986–1993. Abstracta Botanica **20**:17–36.

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

Borcard, D., P. Legendre, and P. Drapeau. 1992. Partialling out the spatial component of ecological variation. Ecology **73**: 1045–1055.

Boyce, M. S., and L. L. McDonald. 1999. Relating populations to habitats using resource selection functions. Trends in Ecology and Evolution **14**:268–272.

Cottenie, K. 2005. Integrating environmental and spatial processes in ecological community analysis. Ecology Letters **8**:1175–1182.

Cramer, J. S. 1987. Mean and variance of $R^2$ in small and moderate samples. Journal of Econometrics **35**:253–266.

Cushman, S. A., and K. McGarigal. 2002. Hierarchical, multiscale decomposition of species–environment relationships. Landscape Ecology **17**:637–646.

Dawson-Saunders, B. T. 1982. Correcting for the bias in the canonical redundancy statistic. Educational and Psychological Measurement **42**:131–143.

Diniz, J. A. F., and L. M. Bini. 1996. Assessing the relationship between multivariate community structure and environmental variables. Marine Ecology Progress Series **143**:303–306.

Dray, S., P. Legendre, and P. R. Peres-Neto. 2006. Spatial modelling: a comprehensive framework for principal coordinate analysis of neighbor matrices (PCNM). Ecological Modelling, in press.

Ezekiel, M. 1930. Methods of correlational analysis. Wiley, New York, New York, USA.

Jackson, D. A., and H. H. Harvey. 1993. Fish and benthic invertebrates: community concordance and community–environment relationships. Canadian Journal of Fisheries and Aquatic Sciences **50**:2641–2651.

Jenkins, D. G., and A. L. Buikema. 1998. Do similar communities develop in similar sites? A test with zooplank-

ton structure and function. Ecological Monographs **68**:421–443.

Kerlinger, F. N., and E. J. Pedhazur. 1973. Multiple regression in behavioral research. Holt, Rinehart and Winston, New York, New York, USA.

Kromrey, J. D., and C. V. Hines. 1995. Use of empirical estimates of shrinkage in multiple regression: a caution. Educational and Psychological Measurement **55**:901–925.

Lambert, Z. V., A. R. Wildt, and R. M. Durand. 1988. Redundancy analysis—an alternative to canonical correlation and multivariate multiple-regression in exploring interset associations. Psychological Bulletin **104**:282–289.

Legendre, P., and M. J. Anderson. 1999. Distance-based redundancy analysis: testing multi-species responses in multi-factorial ecological experiments. Ecological Monographs **69**:1–24.

Legendre, P., D. Borcard, and P. R. Peres-Neto. 2005. Analyzing beta diversity: partitioning the spatial variation of community composition data. Ecological Monographs **75**: 435–450.

Legendre, P., and M.-J. Fortin. 1989. Spatial pattern and ecological analysis. Vegetatio **80**:107–138.

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

Legendre, P., and L. Legendre. 1998. Numerical ecology. Second English edition. Elsevier Science BV, Amsterdam, The Netherlands.

Manly, B. F. J. 1997. Randomization, bootstrap, and Monte Carlo methods in biology. Second edition. Chapman and Hall, London, UK.

Martin, T. G., B. A. Wintle, J. R. Rhodes, P. M. Kuhnert, S. A. Field, S. J. Low-Choy, A. J. Tyre, and H. P. Possingham. 2005. Zero tolerance ecology: improving ecology inference by modeling the source of zero observations. Ecology Letters **8**: 1235–1246.

Miller, J. K. 1975. The sampling distribution and a test for the significance of the bimultivariate redundancy statistic: a Monte Carlo study. Multivariate Behavioral Research **10**: 233–244.

Miller, J. K., and S. D. Farr. 1971. Bimultivariate redundancy: a comprehensive measure of interbattery relationship. Multivariate Behavioral Research **6**:313–324.

Mood, A. M. 1969. Macro-analysis of the American educational system. Operations Research **17**:770–784.

Økland, R. H. 2003. Partitioning the variation in a plot-by-species data matrix that is related to *n* sets of explanatory variables. Journal of Vegetation Science **14**:693–700.

Oksanen, J., R. Kindt, P. Legendre, and R. B. O'Hara. 2005. vegan: community ecology package. Version 1.7–81. ⟨http://cran.r-project.org/⟩

Peres-Neto, P. R. 2004. Patterns in the co-occurrence of stream fish metacommunties: the role of site suitability, morphology and phylogeny versus species interactions. Oecologia **140**: 352–360.

Peres-Neto, P. R., and D. A. Jackson. 2001. How well do multivariate data sets match? The robustness and flexibility of a Procrustean superimposition approach over the Mantel test. Oecologia **129**:169–178.

Peres-Neto, P. R., D. A. Jackson, and K. M. Somers. 2003. Giving meaningful interpretation to ordination axes: assessing the significance of eigenvector coefficients in principal component analysis. Ecology **84**:2347–2363.

Peres-Neto, P. R., and J. D. Olden. 2001. Assessing the robustness of randomization tests: examples from behavioural studies. Animal Behaviour **61**:79–86.

Peres-Neto, P. R., J. D. Olden, and D. A. Jackson. 2001. Environmentally constrained null models: site suitability as occupancy criterion. Oikos **93**:110–120.

Rao, C. R. 1964. The use and interpretation of principal component analysis in applied research. Sankhyaá, Series A **26**:329–358.

Rodríguez, M. A., and W. M. Lewis, Jr. 1997. Structure of fish assemblages along environmental gradients in floodplain lakes of the Orinoco River. Ecological Monographs **67**:109–128.

Royston, P., and S. G. Simpson. 1995. Comparing non-nested regression models. Biometrics **51**:114–127.

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

ter Braak, C. J. F., and P. Smilauer. 2002. Canoco reference manual and CanoDraw for Windows user's guide: software for canonical community ordination (Version 4.5). Microcomputer Power, Ithaca, New York, USA.

Watnik, M., W. Johnson, and E. J. Bedrick. 2001. Nonnested linear model selection revisited. Communications in statistics—theory and methods **30**:1–20.

Wu, C. F. J. 1986. Jackknife, bootstrap and other resampling methods in regression-analysis. Annals of Statistics **14**:1261–1295.

Zar, J. H. 1999. Biostatistical analysis. Third edition. Prentice Hall, London, UK.

## APPENDIX A

A detailed description of the steps involved in calculation of the redundancy statistics in canonical correspondence analysis (CCA) (*Ecological Archives* E087-158-A1).

## APPENDIX B

A table presenting the slopes relating the species to the environmental variables considered in the simulations (*Ecological Archives* E087-158-A2).

## APPENDIX C

Results for the Hellinger-transformed species data (*Ecological Archives* E087-158-A3).

## APPENDIX D

A simulation study showing the accuracy of the suggested bootstrapped adjusted $R^2$ (*Ecological Archives* E087-158-A4).

## APPENDIX E

Results of simulations to assess the Type I error and power of the proposed bootstrap test of the difference between fractions in variation partitioning, including results for redundancy analysis (RDA) and canonical correspondence analysis (CCA) (*Ecological Archives* E087-158-A5).

## SUPPLEMENT 1

Matlab function for conducting canonical correspondence analysis for very large data sets (*Ecological Archives* E087-158-S1).

## SUPPLEMENT 2

Matlab library for conducting variation partitioning and test of fractions in multiple regression and canonical models (*Ecological Archives* E087-158-S2).

## SUPPLEMENT 3

An executable program for conduction variation partitioning with adjustments, test of fractions in redundancy analysis (RDA) and canonical correspondence analysis (CCA) (*Ecological Archives* E087-158-S3).