University of Alberta

Analytical developments in the use of resemblance measures in community ecology and applications to boreal forest Carabidae

by

Guillaume Blanchet

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfilment of the requirements for the degree of

Doctor of Philosophy
in
Conservation Biology

Department of Renewable Resources

Fall 2012
Edmonton, Alberta

Abstract

Understanding the factors influencing the distribution of species is one of the main goals of ecology. This thesis presents three contributions to better and more efficiently understand the factors defining the composition of ecological communities. First, I studied the impact of anthropogenic disturbances, habitat heterogeneity, and spatial autocorrelation on Carabidae in a mature boreal forest. I showed that carabids were influenced mainly by forest floor cover, soil drainage, and tree composition. Moderate levels of anthropogenic disturbance only mildly influenced the spatial distribution of the carabid assemblages. I concluded that, carabid diversity would be best conserved in boreal forests if a network of large forest patches were left after harvest. Second, I considered the difficulty of analysing multivariate data, the main challenge in analysing species communities. Canonical redundancy analysis (RDA) is a flexible approach to relating a species community to environmental constraints. Although flexibility flows from the fact that any resemblance measure can be used within this framework, there is little guidance for how to select from the large number of existing resemblance measures. Using communities simulated from 25 different species abundance distributions (SAD), I compared results from 16 different resemblance measures within the RDA framework. The results showed that, independent of SADs, all resemblance measures gave equivalent results whether the communities were recorded as abundance or presence-absence data. In light of these results, I proposed a new canonical ordination to make a consensus of RDAs across resemblance measures. In my simulations presence-absence data were directly

derived from abundance data, and so I also evaluated if the information in presence-absence and abundance data gave equivalent result. I found that the data formats may be complementary. Lastly, in ecological applications, either abundance/cover or presence-absence data are collected when species communities are sampled. With the help of resemblance measures, I propose a new way to survey ecological communities that is intermediate between presence-absence and abundance data. This approach is more cost-effective than counting abundance yet more informative than recording presence-absence. Overall, this thesis contributes to understanding spatial distribution of carabids in boreal forests and provides new methods to analyse multivariate ecological data.

Acknowledgements

First and foremost I am indebt to my research supervisor Fangliang He. His contributions to many aspects of my work is the base on which I now rely to pursue a career in science. I am also grateful to John Spence and his research group who have included me as part of their own and have introduced me to the wonderful world of arthropods.

This work would have not been possible without the help of Colin Bergeron who gathered the Carabidae data I used throughout this thesis. The advices he gave me and what he thought me of these amazing insects has been of great help and will not be forgotten.

I am also thankful to Ellen Macdonald, Subhash Lele, Mark Lewis, and David W. Roberts for their interest in my work.

Through this thesis I also received comments from friends and colleagues often around a beer or a glass of wine. For these comments I am very thankful. This work would not have been the same without you. Thank you Pierre Legendre, Daniel Borcard, Xianli Wang, Anne Oxbrough, Benoît Gendreau-Bertiaume, Stéphane Bourassa, Charlene Wood, Jaime Pinzon, Jian Zhang, Xiaofeng Ruan, and Valentin Reyes-Hernendez.

The support I received from my parents, Diane and Richard, and from my family Elisa and Jérôme Xavier has made the completion of this thesis possible.

Merci à vous tous !

Table of Contents

List of Tables

List of Figures

FIGURE 2.4 Partial canonical redundancy analysis (partial RDA) of the carabid species with spatial structure (Moran's eigenvector maps and the sites spatial coordinates) as covariate. Results are presented in two triplots to make them easier to read. Top plot – Forest floor cover is represented by empty squares and topography by the black arrows. Floor covers followed by 1, 2 and 3 are the 1st, 2nd and 3rd most dominant, respectively. Bottom plot – Drainage is represented by full circles and vegetation structure variables are the black arrows. Drainage is defined following Appendix A. Forest productivity is the mean tree diameter at breast height (DBH) and forest structure is the coefficient of variation of the trees DBH. Carabid species are represented in both plots by segments. Ground beetles species close to the center of the triplot were removed. The remaining species were sampled at a minimum of 18 sites and contributed to at least 29% of the variance of all canonical axes. Agongrati = *Agonum gratiosum* (Mannerheim), Agonretra = *Agonum retractum* LeConte, Calaadven = *Calathus advena* (LeConte), Calaingra = *Calathus ingratus* Dejean, Patrfoveo = *Patrobus foveocollis* (Eschscholtz), Platdecen = *Platynus decentis* (Say), Platmann = *Platynus mannerheimi* Dejean, Pteradstr = *Pterostichus adstrictus* Eschscholtz, Pterpunct = *Pterostichus punctatissimus* (Randall), Sterhaema = *Stereocerus haematopus (*Dejean), Trecchaly = *Trechus chalybeus* Dejean. Each axis

FIGURE 3.3. 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 performed using a distance scaling (scaling 1) but with a different association coefficient. In the figure, *K* different association coefficients are used. An RDA result is composed of four principal matrices: the species scores **U**, the sites scores **Z** calculated in the space of **X**, the site scores **F** calculated in the space of **Y** and the canonical coefficients **C**. Because *K* different meaningful association coefficients are used and thus *K* RDAs are performed, *K* sets of matrices **U**, **Z**, **F**, and **C** are calculated. (b) For each of the *K* association coefficients, the significant axes within each **Z** matrix are grouped in a large matrix. A PCA is then performed on this large matrix yielding the site scores consensus matrix $\mathbf{Z}^*$ and a diagonal matrix of eigenvalues $\mathbf{\Lambda}^*$. (c) Using $\mathbf{Z}^*$ as a reference, a matrix of orthogonal rotation **H** is calculated for each **Z** matrix. The construction of **H** matrices are carried out using the scaled $\mathbf{Z}^*$ and $\mathbf{Z}_k$. By that we mean that $\mathbf{Z}^*$ and $\mathbf{Z}_k$ were divided by their sum of squares before being square root transformed. (d) The consensus species scores $\mathbf{U}^*$ can then be computed by multiplying each $\mathbf{U}_k$ by its respective orthogonal rotation matrix and averaging all the rotated $\mathbf{U}_k$ matrices. The same calculation is performed to obtain the consensus canonical coefficients $\mathbf{C}^*$ and if necessary $\mathbf{F}^*$. For $\mathbf{U}^*$, $\mathbf{C}^*$ and $\mathbf{F}^*$ to be optimal, they need to be calculated from the independent matrices with a sum of squares equal to 1. (e) $\mathbf{Z}^*$, $\mathbf{U}^*$, $\mathbf{C}^*$ and $\mathbf{F}^*$ can than be used to draw a consensus RDA triplot. The

List of Abbreviations

ANOVA ..... Analysis of variance

CA ..... Correspondance Analysis

CCA ..... Canonical Correspondance Analysis

CV ..... Coefficient of Variance

DBH ..... Diameter at Breast Height

EMEND ..... Ecosystem Management Emulating Natural Disturbance

MEM ..... Moran's Eigenvector Maps

MST ..... Minimum Spanning Tree

NMDS ..... Non-metric Multidimensional Scaling

OCCAS ..... Ordered Comparison Case Series

P ..... P-value

P_C ..... corrected P-value

PCA ..... Principal Component Analysis

PCoA ..... Principal Coordinate Analysis

R^2 ..... Coefficient of determination

R^2_A .....Adjusted coefficient of determination

RDA .....Canonical Redundancy Analysis

DB-RDA ..... distance-based Canonical Redundancy Analysis

TB-RDA.....transformation-based Canonical Redundancy Analysis

SAD ..... Species abundance distribution

SU ..... Sampling Unit

Introduction

Understanding the relationships species have with each other and with their habitat is a main goal of ecology (Morin 2011). Because of the multivariate nature of community data, researchers often resort to dimension reduction tools such as ordinations to better understand the patterns defining these ecological communities. “Ordination” is the arrangement of objects (e.g., sites) in a particular order (Goodall, 1954, Legendre and Legendre 2012, Chapter 9). Central to these multivariate methods are resemblance measures that quantify the association between sites. They are at the core of modern statistical analysis in community ecology.

Many resemblance measures have been proposed in the ecological literature. To understand why so many of these measures have been introduced in ecology, I present a historical overview of resemblance measures in the context of ordinations.

Of the many resemblance measures used in ecology, the basic and most widely known one is Euclidean distance, which measures the geometric distance between two objects. Principal component analysis (PCA, Pearson 1901; Hotelling 1933) is based on the Euclidean distance. It was the first ordination method ever by ecologistsused to analyse community data, as introduced by Goodall (1954) in a plant community study. Even though PCA is a good dimension reduction method, its assumptions constrain its use to ecological problems focussing on total, not proportional, changes in abundances, biomass, or cover. Changes in total abundances, biomass, or cover are generally the result of

sudden changes in the environment such as upwelling, disturbance, or the introduction of a predator. On the other hand, changes in proportional abundances, biomass, or cover are more often associated to smoother variation from resource gradient, such as moisture or salinity gradients. For example, if a study focuses on the impact of an oil spill or ocean currents like the Gulf Stream (i.e. sudden changes) on a species community, it is appropriate to use PCA to extract ecological information because the absence (or presence) of a species can be directly associated to these changes. All resemblance measures that are symmetrical, like Euclidean distance, should only be used in situations where changes in total abundances are of interest (Anderson et al. 2011).

With the exception of a few probabilistic resemblance measures (e.g., Raup-Crick measure [Raup and Crick 1979]), most measures can be considered to be symmetrical or asymmetrical. When measuring the resemblance of two sites, a symmetrical resemblance measure will give importance to the absence of a species at both sites. In contrast, asymmetrical resemblance measures do not attribute any weight to a species absent at two sites. Symmetrical resemblance measures suffer from a problem known in ecology as the double-zero problem. This problem (and its name) stems from the difficulty in interpreting the absence of a species from two sites. There are many causes for a species to be absent from a site. For example, it is possible that the niche of a species happens to be occupied by other species due to some stochastic processes or that a species was missed because sampling effort was not important enough to catch all species. Therefore, the absence of a species at two sites is not necessarily an indication of resemblance between the two sites. In that context, PCA may be ill adapted to

extract community patterns because it relies on a symmetrical resemblance measure. As explained in the previous paragraph, it is justified to use it only when the reason for species to be absent is of interest.

With the double-zero problem, the need to perform ordinations based on asymmetrical resemblance measures became apparent. Ecologists have found a solution in correspondence analysis (CA, Greenacre 2007, Legendre and Legendre 2012, subsection 9.2), which relies on the χ^2 distance, an asymmetric resemblance measure (Legendre and Legendre 2012, Subsection 7.4.1). It was first applied to study plant communities by Roux and Roux (1967).

Parallel to the introduction of CA in community ecology, ecologists more versed in mathematics and statistics have also developed ordination methods appropriate for analyses of multivariate species data. An example of such development is the classic paper by Bray and Curtis (1957), which presents an ordination method based on the percentage difference distance. Note that Bray and Curtis (1957) proposed no distance measure, but referred to Motyka et al. (1950) for the resemblance measure that underlies the Bray-Curtis ordination. This resemblance measure was attributed to Steinhaus by Motyka (1947) and was rediscovered by Odum (1950) who named it percentage difference (Legendre and Legendre, 2012, Subsection 7.4.2).

With the introduction of PCA, CA, and the Bray-Curtis ordination to community ecology, researchers could perform ordinations based on the Euclidean, χ^2, and percentage difference distance, respectively. However, by the time these ordination methods were applied to the study of ecological

communities, many other resemblance measures had already been proposed both to and by ecologists. However, no ordination methods could be performed using these new measures. This problem was overcome in 1962 by Shepard who introduced non-metric multidimensional scaling (NMDS) and by Gower (1966) who proposed principal coordinate analysis (PCoA). Any resemblance measure can be used in ordinations performed with either of these two methods. These two methodological developments were of utmost importance because they offered the possibility for use of any resemblance measure through a single methodological framework.

As is true for all other ordinations, NMDS and PCoA are dimensional reduction methods. However, because of their flexibility, they also offer the possibility of comparing resemblance measures through an ordination framework to decide which one is better adapted to study species communities. Compared to PCA and CA, at pitfall of NMDS and PCoA is that the information about species is lost because NMDS and PCoA are performed on matrices calculating distance between sites and that the information about species is lost in these matrices.

Following the development of simple ordinations (PCA, CA, Bray-Curtis ordination, NMDS, and PCoA) came the widespread use of canonical ordinations. Canonical ordinations relate a matrix of response variables, such as a species community, with a matrix of explanatory variables. As for simple ordinations, the first canonical ordination methods introduced in the ecological literature were not flexible with regards to the resemblance measure that underpinned them. Canonical redundancy analysis (RDA, Rao 1964, Wollenberg 1977), for example, is based on Euclidean distance while canonical correspondence analysis (CCA,

ter Braak 1986) relies on the χ^2 distance. With the development of distance-based RDA (db-RDA, Legendre and Anderson 1999), the canonical equivalent of PCoA, it became possible to perform canonical ordinations based on any resemblance measure. However, similar to the implications for PCoA of this flexibility, the information on species is lost in db-RDA.

To adapt the use of PCA and RDA to a wider range of ecological problems, Legendre and Gallagher (2001) proposed to pre-transform community data using ecologically meaningful transformations. These pre-transformations enable researchers to perform PCAs and RDAs with distances other than the Euclidean distance because the distance preserved between sites after pre-transformation is different than the Euclidean distance and of course depends on the pre-transformation. Because PCA and RDA conserve the information about the response variables (species), as do CA and CCA, it became possible to extract ecological information from species using resemblance measures other than the Euclidean and χ^2 distances. Until recently the approach proposed by Legendre and Gallagher (2001) was the only one that allowed choice among a reduced group of resemblance measures to perform ordinations that conserved information about species. Borcard et al. (2011), Legendre and Legendre (2012, Subsection 9.3.3) and Oksanen et al. (2012), however, have now shown that it is possible to gather information on species even when using PCoA and db-RDA.

Methodological developments since the second half of the twentieth century allow researchers to perform ordinations (simple and canonical) using any resemblance measure. However, because of the large number of resemblance

measures used to model community data, ecologists face difficulties in deciding which resemblance measure to use. Although some evaluates the strength and weaknesses of different resemblance measures, most such studies do not apply to recent methodological developments, such as canonical ordinations.

The history of attemps to resolve this problem has been full of controversy. Hajdu (1981) proposed to construct a set of test cases, which he named "ordered comparison case series" (OCCAS), that may be used as reference to choose a resemblance measure among a set of candidate measures. These test cases present linear changes in the abundance of two species along a set of simulated gradients. A pitfall of the OCCAS approach is that simulated species generally have high abundance and thus poorly reflect the abundance patterns of real ecological communities. Moreover, linear relationships between a species and its environment are not necessarily a pattern found in nature.

Gower and Legendre (1986) used the OCCAS to evaluate the behaviour of 25 resemblance measures (15 for binary data and 10 for quantitative data). They concluded that only two of the measures compared should be avoided because they are strongly non-linear, a pattern that species simulated with the OCCAS does not have. These are the Yule (Sokal and Sneath 1963) and Kulczynski (1927) resemblance measures. These results suggest either that the OCCAS does not effectively discriminate among the resemblance measures compared by Gower and Legendre (1986) or that the information in the test data is presented equivalently by 23 of the 25 resemblance measures compared.

Also using simulations, Bloom (1981) compared four resemblance measures including the Morisita index (Morisita 1959), and concluded that the

percentage difference distance is the only one that should be used to study ecological communities. Although this result is in accordance with Faith et al. (1987), it contradicts Wolda (1981) who compared 22 resemblance measures with simulations, among them percentage difference, and concluded that only the Morisita index should be used. During the same time period, but using empirical data from the fungi genus *Chaetomium*, Hubàlek (1982) compared 43 binary resemblance measures using cluster analysis. Contrary to Gower and Legendre (1986), he suggested using the Kulczynski resemblance measure. He also highlighted the Jaccard (1901), Sørensen (Dice 1945, Sørensen 1948), and Ochiai (1957) resemblance measures as good measures to study communities sampled using presence-absence data.

More recently, Legendre and Gallagher (2001) have offered guidelines to select resemblance measures. Unlike the papers comparing resemblance measures discussed above (Bloom 1981, Wolda 1981, Hubàlek 1982, Gower and Legendre 1986, Faith et al. 1987), they put ordinations into constant perspective, focusing on simple ordinations. They compared seven resemblance measures using only abundance data and conclude that the Hellinger (Rao 1995) and chord (Orlóci 1967, Cavalli-Sforza and Edwards 1967) distances are good alternatives to the percentage difference distance, although the later distance was still deemed valid. Legendre and Gallagher (2001) also proposed generalizing their conclusions to canonical ordinations and to other data formats (e.g., presence-absence) but they did not test these ideas. Recently, Legendre and Legendre (2012, Subsection 7.6) proposed a decision key to choose resemblance measures based on their properties.

As can be understood from the various papers that compare resemblance measures as discussed above, it is difficult to decide which measure to use because studies have often contradicted each other, favouring or discrediting the same resemblance measures. All of these authors aimed at identifying one or a few resemblance measures that could be used for almost any situation where community data are analysed. One goal of this thesis is to compare resemblance measures with respect to addressing particular research questions. I started with the premise that it is unlikely that a single resemblance measure can be used to answer all statistical questions about species composition or community ecology.

The species abundance distribution (SAD) is another approach that has been developed to study ecological communities. Like resemblance measures it has a long history. McGill (2007) traces its origin back to 1909, however, it was popularized by Fisher et al. (1943) and Preston (1948). Since then, theoretical and empirical studies have been carried out to better understand the patterns defined by SAD. Although the general patterns of SAD are consistent regardless of the communities studied (rare species are common and common species are rare [McGill 2011]), there is still significant variation among SADs of different ecological communities. Understanding why species are distributed the way they are has been studied independently with SAD and with ordinations using resemblance measures. However, no research has been carried out using both approaches together. In this thesis, I used SAD to evaluate the strengths and weaknesses of a group of resemblance measures when used in canonical ordinations (Chapter 3), and to evaluate if it is possible to be more cost-efficient when sampling ecological communities (Chapter 4).

Legendre and Gallagher (2001) proposed to select a resemblance measure for a canonical ordination as the one that yielded the largest fraction of explained variance (R^2). However, they did not explicitly test if this was a good approach to selecting a resemblance measure. In the third chapter, I evaluated if SADs could serve as a reference to choose a resemblance measure to analyse community data through the RDA framework (that is with RDA, or db-RDA). Results using sixteen different resemblance measures were compared to evaluate how their R^2 varies. I assessed usefulness and effectiveness of the different resemblance measures in analysing community data using simulated communities based on different SADs. The results showed that all resemblance measures yielded similar R^2 regardless of the SAD from which the data were simulated. Thus, in chapter 3 I propose a new procedure to use when more than one resemblance measure is applicable to study species communities. This new method makes a consensus of a series of canonical ordinations performed on the same data using different resemblance measures.

Most community surveys involve sampling a large number of species. Counting the abundance of every species for each sampling unit can be tedious and laborious. In contrast, measuring species presence-absence is easier and more cost-effective, but the information lost by collecting only presence-absence data may reduce our ability to efficiently describe and reconstruct the studied communities. In the fourth chapter of this thesis, I propose a new approach that is intermediate to counting abundance and measuring presence-absence. This new data gathering approach is designed to be more cost-effective than counting species abundances and prevents the loss of information caused by recording only

presence-absence. When applied to a species community, we should first evaluate the importance of SADs, the species aggregation level, and the choice of resemblance measure to decide how cost-effective the sampling would be. The approach proposed in this chapter can be applied to virtually any multivariate count datasets (ecological or others).

In addition to presenting two methodological developments that can be applied directly to investigate patterns on ecological communities, this thesis also presents a study aimed at understanding the factors that influence the distribution of ground beetles (Carabidae) in a mature boreal forest. I chose to present this study in the second chapter of this thesis because the third and fourth chapters will use the same data for ecological illustrations. In the second chapter, I examine the effect of landscape disturbances, habitat heterogeneity, and spatial autocorrelation on a ground beetle assemblage in boreal forest. Ecologically, boreal ground beetles are diverse and sensitive to variation in environmental factors making them good bioindicators (Rainio and Niemelä 2003). Carabids have also been shown to react strongly to harvesting (Pearce and Venier 2006). In a spatial context, the way ground beetles interact with the environment, and the effect disturbances have on this relationship may vary considerably, depending on the scale of the study (Niemelä and Spence 1994). Many carabid studies have been carried out at a scale of a few hectares (Thiele 1977, Lövei and Sunderland 1996), but very few studies have been conducted at scales larger than a square kilometre (Vanbergen et al. [2005] and Woodcock et al. [2010] are two notable exceptions). The second chapter considers carabid in 70 km^2 of boreal forest where 194 sites were sampled in a near-regular grid. Forest floor cover, soil drainage, and tree

composition were used to define habitat heterogeneity. To evaluate the importance of spatial autocorrelation, Moran's eigenvector maps (Dray et al. 2006) were used. To study landscape disturbances I referred to the shortest distance to an anthropogenic disturbance (road, seismic line, or harvest block). In this chapter, all analyses were performed using the Hellinger distance. This resemblance measure was chosen based on knowledge of the carabid community and its property when applied to the statistical methods used to carry out the analyses.

In summary, the factors structuring species in a community are often of different nature, as illustrated in Chapter 2. However, the methodological approaches used to count and model ecological communities can influence the interpretations made of these data. This thesis presents methodological developments that will help ecologists more efficiently gather community data and better analyse them when using resemblance measures. To show how these new analytical developments can increase the understanding of real ecological systems, I used Carabidae sampled in northwestern Alberta as illustration.

LITERATURE CITED

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.

Bloom, S. A. 1981. Similarity indices in community studies: potential pitfalls. Marine Ecology-Progress Series **5**:125–128.

Borcard, D., F. Gillet, and P. Legendre. 2011. Numerical Ecology with R. Use R!, Springer, New York.

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

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

Dice, L. R. 1945. Measures of the amount of ecologic association between species. Ecology **26**:297–302.

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

Faith, D., P. Minchin, and L. Belbin. 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio **69**:57–68.

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

Goodall, D. W. 1954. Objective methods for the classification of vegetation. III. An essay in the use of factor analysis. Australian Journal of Botany **2**:304–324.

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

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

Greenacre, M. 2007. Correspondence Analysis in Practice. Chapman & Hall.

Hajdu, L. J. 1981. Graphical comparison of resemblance measures in phytosociology. Vegetatio **48**:47–59.

Hotelling, H. 1933. Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology **24**:417–441.

Hubálek, Z. 1982. Coefficients of association and similarity, based on binary (presence-absence) data: an evaluation. Biological Reviews **57**:669–689.

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

Kulczynski, S. 1927. Zespoly roslin w Pieninach. Bulletin International de l'Académie Polonaise des Sciences et des Lettres, Classe des Sciences Mathématiques et Naturelles, Série B (Sciences Naturelles) Supplément II, 57–203.

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

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

Legendre, P., and L. Legendre. 2012. Numerical Ecology. 3rd edition. Elsevier.

Lövei, G. L., and K. D. Sunderland. 1996. Ecology and behavior of ground beetles (Coleoptera: Carabidae). Annual Review of Entomology **41**:231–256.

McGill, B. J., 2011. Species abundance distributions. Pages 105–122 *in* A. E. Magurran and B. J. McGill, editors. Biological diversity: Frontiers in Measurement and Assessement. Oxford University Press.

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

Morin, P. J. 2011. Community Ecology. 2nd edition. Wiley-Blackwell.

Morisita, M. 1959. Measuring of interspecific association and similarity between communities. Memoirs of the Faculty of Science, Kyushu University, Series E (Biology) **3**:65–80.

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

Motyka, J., B. Dobrzanski, S. Zawadzki, et al. 1950. Preliminary studies on meadows in the south-east of Lublin province. Annales Universitatis Mariae Curie-Sklodowska **5**:367–447.

Niemelä, J. K., and J. R. Spence. 1994. Distribution of forest dwelling carabids (Coleoptera): spatial scale and the concept of communities. Ecography **17**:166–175.

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

Odum, E. P. 1950. Bird populations of the Highlands (North Carolina) plateau in relation to plant succession and avian invasion. Ecology **31**:587– 605.

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

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

Pearce, J. L., and L. A. Venier. 2006. The use of ground beetles (Coleoptera: Carabidae) and spiders (Araneae) as bioindicators of sustainable forest management: A review. Ecological Indicators **6**:780–793.

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

Preston, F. W. 1948. The commonness, and rarity, of species. Ecology **29**:254–283.

Rainio, J., and J. Niemelä. 2003. Ground beetles (Coleoptera: Carabidae) as bioindicators. Biodiversity and Conservation **12**:487–506.

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

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

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

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

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

Sokal, R. R., and P. H. A. Sneath. 1963. Principles of numerical taxonomy. W. H. Freeman and company.

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

Thiele, H.-U. 1977. Carabid Beetles in their Environments. Springer-Verlag, Berlin.

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

van den Wollenberg, A. L. 1977. Redundancy analysis an alternative for canonical correlation analysis. Psychometrika **42**:207–219.

Vanbergen, A. J., B. A. Woodcock, A. D. Watt, and J. Niemelä. 2005. Effect of land-use heterogeneity on carabid communities at the landscape scale. Ecography **28**:3–16.

Wolda, H. 1981. Similarity indices, sample size and diversity. Oecologia **50**:296–302.

Woodcock, B. A., J. Redhead, A. J. Vanbergen, L. Hulmes, S. Hulmes, J. Peyton, M. Nowakowski, R. F. Pywell, and M. S. Heard. 2010. Impact of habitat type and landscape structure on biomass, species richness and functional diversity of ground beetles. Agriculture, ecosystems and environment **139**:181–186.

Chapter 2 – Landscape effects of disturbance, habitat heterogeneity and spatial autocorrelation for a ground beetle (Carabidae) assemblage in mature boreal forest

A version of this chapter has been accepted for publication with minor corrections: Blanchet F. G., J. A. C. Bergeron, J. R. Spence, and F. He. 2012. Ecography

INTRODUCTION

Many factors influence the distribution of species. The traditional autecological paradigm suggests that environmental factors exert the main influence on location of species in landscapes (Hutchinson 1957). Over the past forty years, however, researchers have emphasized the impact of factors such as disturbances (Sousa 1984) on establishing and maintaining patterns of species distribution and have studied how these species are spatially autocorrelated (Legendre and Fortin 1989). Clearly, spatial aggregation of species in communities can result from species interacting either with their environment or among themselves (Legendre 1993). In this paper, we use the concept of 'space' in a general sense, to reflect unmeasured spatially structured habitat variables (e.g., geological formations, moisture, soil conditions) and/or dispersal limitations (caused by reproduction strategies, mortality, migration, predation, etc.) that may affect distribution of species in an area.

Inclusion of disturbance and space in models of community structure can significantly increase the understanding of both species interactions and the dynamics of the ecosystem in which they occur. The effects of disturbances, habitat heterogeneity, and space on species assemblages have been the focus of studies in various ecosystems, with this theme having been particularly relevant for forests (Parisien and Moritz 2009), coral reefs (Connell et al. 1997) and agricultural landscapes (Kleyer et al. 2007). Effects of disturbance, habitat heterogeneity and space have also been featured in numerous studies on boreal landscapes. For example, several Canadian studies of forest ecosystems [e.g., the

EMEND (Ecosystem Management Emulating Natural Disturbance) project in northwestern Alberta (Spence et al. 1999) and the SAFE (Silviculture et Aménagement Forestiers Écosystémique) project in Québec (Harvey et al. 1997)] consider disturbances, habitat heterogeneity and space on a broad spatial scale in an effort to understand forest dynamics and conserve biodiversity. Increasing knowledge about factors influencing species distributions through such long-term studies is central to ecosystem–based management of boreal forests because resource extraction patterns (e.g., forest harvesting, mining) involves broad spatial scales and can have long lasting effects.

In this study we focus on boreal ground beetles (Carabidae), as they are diverse and have been identified as good bioindicators due to their sensitivity to changes in environmental factors (Rainio and Niemelä 2003). Changes in ground beetle assemblages are tightly linked to the edaphic conditions of habitats (Lövei and Sunderland 1996) and, as such, carabid community dynamics may be a model for other organisms responding to these same habitat conditions (e.g., epigaeic fauna, understory flora). As a functional group, ground dwelling arthropods, including carabid beetles, react strongly to industrial harvesting (Buddle et al. 2006, Pearce and Venier 2006).

Although it is well understood that harvest disturbances and environmental factors interact to widely affect assemblages of epigaeic invertebrates in boreal forests (e.g., Niemelä et al. 1993, Niemelä 1997, Work et al. 2004), it remains a challenge to partition the effects of environment and disturbance in space. Thus, by examining the landscape scale response of a ground dwelling beetle

assemblage to disturbances, habitat heterogeneity, and space, we provide insights about how the species-rich epigaeic fauna may be affected by these factors. Such insights are highly relevant to basic understanding of forest function and for implementing biodiversity conservation in development of sustainable forest management (Burton et al. 2003; Lindenmayer and Franklin 2003).

Interactions between an epigaeic fauna and the environment, as well as the effect of disturbances on this relationship vary notably from fine to broad scales (Niemelä and Spence 1994); however, many aspects of biodiversity can only be studied meaningfully at a broad scale (Wiens 1989). In contrast, studies of ground dwelling invertebrates are rarely conducted in areas larger than a few square kilometres; most are undertaken in single locales. Among the largest scale studies reported in the literature, Vanbergen et al. (2005) sampled six 1 km^2 quadrats while Woodwork et al. (2010) sampled an area of 10 km^2. The carabid dataset analysed here is exceptional as it covers 70 km^2, an area sufficiently large to capture the spatial variation in a species assemblage and to understand the possible impact of disturbances that operate over broad scales. A study at this scale is important to understand more completely what structures the epigaeic community at scales corresponding to the phenomena shaping the structure and variation of the boreal forest.

The general objective of this study is to assess the effects of habitat heterogeneity and anthropogenic disturbances on boreal carabid diversity at a landscape scale. Specifically, we aimed to answer the following three questions: (1) Is the spatial pattern of ground beetle assemblages in mature forest affected by

surrounding anthropogenic disturbance? From the numerous studies carried out at finer local scales, we expected anthropogenic disturbance to influence the distribution of carabids in our study. (2) What portion of the environmental factors structuring beetle assemblages is spatially autocorrelated? (3) What commonly recognized components of the habitat (e.g., forest floor cover, soil drainage, vegetation structure or topography) are most important for structuring boreal carabid assemblages? Carabids disperse mainly by running on the forest floor. We thus expected that descriptors of ground conditions would be the main factors structuring boreal ground beetles. However, because descriptors of ground conditions and vegetation structure interact strongly, we hypothesized that both would notably influence the structure of the carabid community.

MATERIAL AND METHODS

Study area and Forest Sampling

The data analysed here were collected at the EMEND study site located in northwestern Alberta, Canada (56º46'13''N, 118º22'28''W) on the southeast slopes of the Clear Hills formation. The site is dominated by boreal mixedwood forest, as defined by Rowe (1972). In the summer of 2002, 194 sites distributed on a near-regular grid were established in an area of 70 km^2 (Figure 2.1). Each site was positioned in never-harvested mature forests, a minimum of 40 meters away from any anthropogenic disturbance or major water bodies. On average, sites were 700 meters apart, with a minimum distance between any two sites of 260 meters. Forest harvesting (tree felling), two roads and a number of seismic lines produce some forest fragmentation in the sampling area. In addition to the never-harvested

mature forests stands in which the sites were located, the EMEND experiment also includes burned, slash burned (harvested forest in which the harvest residuals have been spread and subsequently burned) and slash harvested (harvested forest in which the harvest residuals have been spread) stands as well as sites with different levels of harvesting (0%, 10%, 20%, 50%, and 75% of unharvested forest).

Prior to c. 1960 the forest on our study site was strictly natural, including limited aboriginal use, and the fragmentation mentioned above resulted from the first industrial anthropogenic disturbances in this area. In short, the area studied was probably about as 'natural' as can be found on the western boreal plain of Canada or in similar latitudes of boreal forest anywhere on Earth. Although the sampled sites were all located in never-harvested mature forests, the epigaeic fauna may still be influenced by surrounding anthropogenic disturbances due for example to forest fragmentation (Davies and Margules 1998).

Carabid samples

Carabids were sampled with three pitfall traps (Spence and Niemelä 1994) at each site during the summer of 2003 (Bergeron et al. 2011). Traps were distributed around the centre point of each site on the circumference of a circle 15 meters in radius, starting at due north and with traps separated from each other by 120 degrees. Such spacing is sufficient to render depletion effects insignificant for ground beetles in the boreal forest (Digweed et al. 1995). Pitfall traps consisted of a 500 mL circular plastic cup of 11 cm in diameter filled with silicate-free ethylene gycol (GM Dex-Cool®) used as preservative. They were placed into the

ground with the lip at soil surface and covered by a wooden roof to reduce accumulation of debris and water.

Traps were emptied every three weeks from early May to the end of August, and so the samples comprised a total of 99 possible trapping days with traps emptied 4 times. Prior to any analysis, total beetle catches at each site were divided by the number of days over which traps were operating. When a trap was non-functional at collection time, the trap content was discarded and the sampling effort for the site was reduced by a third for that specific collection period. This was done to correct for non-demonic intrusions (Hurlbert 1984) such as flooding or destruction of traps by large mammals. To ensure that this correction does not over-emphasize common species we examined the species richness, abundance, and composition at each site in relation to the number of days traps were active. No distinct patterns were found, suggesting that traps were active long enough to effectively capture the carabid composition at each site. All of the 9729 beetles collected in the samples were identified to species using Lindroth (1961–1969) and various updates available in the literature. Overall, the relative activity-density of 43 carabid species was recorded in this study.

Habitat heterogeneity

Variables representing habitat heterogeneity were grouped in two broad categories, comprising descriptors of ground condition and vegetation structure. These variables were recorded as the grid was laid out in 2002 (Bergeron et al. 2011).

Descriptors of ground condition included forest floor cover, soil drainage and topography. The 1st, 2nd and 3rd most dominant floor cover were recorded at each site; each was considered as an independent factor. Throughout the sampling area, seven floor covers were found: mosses, peat, leaves, lichen, needles, grass, and litter. Lichen and litter were never the most dominant cover, and litter was never 3rd most dominant. Each of the three factors describing a dominant level of floor cover was transformed into a set of dummy variables, providing a total of 18 binary descriptors, one per floor cover for each dominance level. For example, if a factor measures the levels A and B, binary descriptors can be constructed for each level. Whenever A was measured a binary descriptor is coded as 1, otherwise 0 is used. The same procedure was used for all levels generating as many binary descriptors as there are levels in a factor.

Soil drainage was characterized in two 0.5 m^2 soil pits dug to mineral soil substrate at opposite ends of each site. Drainage was recorded following a modified version of the Beckingham et al. (1996) classification. We attributed intermediate levels to sites exhibiting characteristics of two adjacent categories of the original classification. A detailed presentation of the 13 level classification we used in this study is presented in Appendix A where it is contrasted to the classification of Beckingham et al. (1996). Nine of the 13 levels were found in our study (from rapidly drained to poorly drained soil). Soil drainage was also transformed into dummy variables, as above for floor cover.

Lastly, three standard topographic measurements were recorded for each site: slope, aspect (cardinal direction toward which the slope faces), and elevation.

Because aspect was recorded in degrees, it was transformed into aspect easting (sine of aspect values) and aspect northing (cosine of aspect values). In all, we used 31 descriptors of ground conditions.

Characterization of vegetation structure was based on the 25 individual trees closest to the centre of each site. Eight tree species were found: aspen (*Populus tremuloides* Michx.), balsam fir (*Abies balsamea* (L.) Mill.), balsam poplar (*Populus balsamifera* L.), black spruce (*Picea mariana* (Mill.) BSP), larch (*Larix laricina* (Du Roi) K. Koch), lodgepole pine (*Pinus contorta* Doug. ex Loud. var. *latifolia* Engelm.), paper birch (*Betula papyrifera* Marsh.) and white spruce (*Picea glauca* (Moench) Voss). For each individual tree, the diameter at breast height (DBH) and the basal area was measured. The DBH average per site was used as surrogate for site productivity. The coefficient of variation (CV) of the DBH was used as a measure of forest structure. We also calculated the relative basal area per species for each site as the area covered by one tree species at a site divided by the area covered by all trees at the same site. We also derived the Shannon diversity (Shannon 1948) and the Pielou's evenness (Pielou 1966) indices from the tree basal area for each sites. Together, 11 variables were used to define vegetation structure.

Statistical analyses

We first performed a Hellinger transformation (Rao 1995) on the ground beetle catch per trap-day per site. The transformation increases the weight given to rare species and yields little 'horseshoe effect', as described by Legendre and Legendre (2012, Subsection 9.2.5). Essentially, when using ordinations, it may

happen that the first axis presents a gradient where sites at the end of the gradient are folded inward and this results in a horseshoe shaped configuration of points on the ordination. The horseshoe effect is often explained as a mathematical construct resulting from progressive changes in species composition of sites along an environmental gradient (Legendre and Legendre 2012, Subsection 9.2.5). Nonetheless, the Hellinger transformation is well adapted to extract ecological patterns from species community data (Legendre and Gallagher 2001). All analyses were performed using the Hellinger transformed data. We also standardized (centred and divided by their standard deviation) all continuous explanatory variables to remove any unit effects.

We analysed the carabid data with a principal component analysis (PCA) calculated on the correlation matrix to identify outlier sites with clearly different species composition. As a result, two sites were removed from subsequent analyses and examined separately.

To test if anthropogenic disturbances (forest harvesting, roads, and seismic lines) influence the spatial distribution of ground beetle assemblages we grouped the sites based on species composition using spatial constrained clustering (Legendre and Legendre 2012, Subsection 13.3.2) following Ward's (1963) criterion and tested if the shortest distance to an anthropogenic disturbance differed between groups using a permutation-based one-way analysis of variance (ANOVA) with 5000 permutations. This ANOVA does not require the data to be normal because it is permutation-based, however it does assume that variance between groups is homogeneous and this was confirmed by Levene's test. To find

the best cluster solution for the carabid data, we tested clustering results with 2 to 15 groups using cross-validation with 1000 iterations for each set of groups. The clustering result with the lowest cross-validation residual error was retained. The spatial constraint was defined by a connexion diagram (Figure 2.1), where fictitious (open circles) and outlier sites (open squares) were not considered. Given the experimental design of EMEND, we considered all stands with 75% or less of forest retained as disturbed. All burned, slash burned, and slash harvested stands were also considered disturbed. Spatial constrained clustering was performed with the 'const.clust' package (Legendre 2011) within the R statistical language (R Development Core Team 2011).

We used Moran's eigenvector maps (MEM, Dray et al. 2006) to evaluate spatial patterns in the carabid assemblage. The spatial variables (eigenfunctions) were constructed based on a connexion diagram (Figure 2.1). To ensure spatial continuity, outlier sites and six fictitious sites as described above were included in the construction of the eigenfunctions. These sites were removed from all spatial variables so that MEMs could be used in subsequent analyses. Five sets of spatial eigenfunctions constructed with different weights were compared: (1) All links have equal weights (presence of link = 1, absence = 0), (2-3) following a concave-up function of the distance ($f_1(d_{ij}) = 1/d_{ij}^{\alpha}$), and (4-5) a concave-down function of the distance ($f_2(d_{ij}) = 1 - (d_{ij} / \max(d_{ij}))^{\alpha}$). In f_1 and f_2, d_{ij} is the distance between sites i and j, and α is either 1 or 2. Spatial variables were constructed with the package 'spacemakeR' (Dray 2011) within the R statistical language.

The MEMs and the spatially constrained clustering introduced in the two previous paragraphs are based on the same connexion diagram (Figure 2.1). They are designed to highlight the same spatial pattern. However, they have different purposes. The spatially constrained clustering defines spatial groups by partitioning the carabid assemblages, while the selected MEMs highlight spatial patterns at varying scales across the sampling area.

Moran's *I* coefficients of spatial correlation (Legendre and Legendre 2012, Subsection 13.1.1) were used to test the significance of each spatial eigenfunction (using 999 random permutations). A spatial eigenfunction associated with a Moran's *I* larger than the expected value of the Moran's *I* models positive spatial correlation. Similarly, a spatial eigenfunction associated with a Moran's *I* smaller than the expected value of the Moran's *I* describes negative spatial autocorrelation. We retained the spatial eigenfunctions that had a significant (P-value < 0.05) Moran's *I* and that modelled either positive or negative autocorrelation.

Both groups of spatial variables were tested independently for use in modelling the ground beetle assemblage. A Šidák correction (Šidák 1967) was applied to the P-values of each test because carabid species were tested twice. If the corrected P-value was significant ($P_c < 0.05$), the adjusted coefficient of multiple determination (R^2_a, Zar 1999) calculated for the carabid assemblage was retained. This procedure was proposed by Blanchet et al. (2008) as a way to coarsely choose which group of spatial variables (positively or negatively autocorrelated) should be considered in more detail. This procedure is carried out

on the detrended species data if the linear trend is significant (Borcard and Legendre 2002). The linear trend models a spatial gradient broader than the sampling area using the X and Y coordinates of the sites. It increases (or decreases) in a north-south and/or east-west direction. We tested it using a permutation test (using 999 random permutations). Because MEMs are akin to sine waves, half of all MEMs are required to model a linear gradient.

The values of R^2_a calculated with different weighting functions were compared and the highest one was retained for further analyses. Forward selection (Blanchet et al. 2008) was then used to find the spatial variables that best modelled distribution of the ground beetle assemblage. We used the 'packfor' package (Dray et al. 2011) in R to select the variables. If significant, the linear trend described in the previous paragraph was re-introduced as an extra spatial variable after the forward selection for all following analyses to be performed on the non-detrended data. Note that even after including the linear trend, the spatial variables present low collinearity among each other. The largest correlation between any pair of spatial variables was 0.29.

Effects on the carabid community of the descriptors of ground condition, those reflecting vegetation structure and the selected spatial variables were quantified using variation partitioning (Borcard et al. 1992). R^2_a was used to measure the importance of each fraction in the variation partitioning results because it corrects for the number of variables which may vary between each group in the variation partitioning analyses (Peres-Neto et al. 2006). To perform this analysis, we used the 'varpart' function in the 'vegan' package (Oksanen et

al. 2012) in R. Variation partitioning was also performed independently on the descriptors of ground condition, using five sub-groups: soil drainage, forest floor cover (from the 1st most to 3rd most dominant), and topography. We also partitioned the variance of the forest vegetation structure into three sub-groups: tree basal area, forest productivity (DBH average) and structure (coefficient of variance of the DBH), and the Shannon and Pielou diversity indices. Appendix B describes the two types of Venn diagram used to present the results of variation partitioning.

To better understand the impact of ground condition and vegetation structure on the carabid community, we performed a partial canonical redundancy analysis (partial RDA, Davies and Tso 1982; Legendre and Legendre 2012, Subsection 11.1.6), using the selected spatial eigenfunctions as covariates. Controlling the spatial component allowed us to focus specifically on the aspects of the beetle community that are not spatially autocorrelated, making this analysis complementary to the spatially constrained clustering described earlier. All canonical axes were tested using 1000 permutations; only the significance (P-value ≤ 0.05) axes were retained for interpretation.

RESULTS

Influence of anthropogenic disturbances on ground beetles

The PCA performed on the ground beetle data isolated two of the 194 sites as having notably different species structure (Figure 2.1, open squares); these were the only two sites located in forest patches retained within harvest blocks. Six carabid species [*Agonum cupreum* Dejean, *Amara erratica* (Duftschmid), *Amara*

laevipennis Kirby, *Harpalus laevipes* Zettersted, *Notiophilus semistriatus* Say and *Poecilus lucublandus* (Say)] were present only at these two sites, making the total number of species at these sites double than that of the other sites. These six species are all open-area specialists (Lindroth 1963, 1968, 1969a), suggesting that their presence expresses local sensitivity of the carabid assemblage to harvest. We removed these two sites for subsequent analyses, bringing the ground beetle species count to 37 in 192 sites.

Six groups of the remaining sites were suggested as the best clustering solution by the spatial constrained clustering (Figure 2.2), as determined by a cross-validation residual error of 0.614. Levene's test showed that the variance was homogeneous among the six groups of sites (P-value = 0.288). We found no significant differences among these groups with respect to the minimum distance to any of the modest anthropogenic disturbance in the area (i.e., road, seismic lines, edges of harvested blocks) (ANOVA, P = 0.1074). Thus, even harvest effects appeared to have impacts that were quite locally expressed.

Influence of space on ground beetles

The carabid data were detrended both in the east-west and north-south directions. The spatial eigenfunctions constructed using the second order ($\alpha = 2$) concave-up function (f_1) explained the largest amount of variance ($R^2_a = 5.8\%$, $P_c < 0.002$). Of the 68 positively autocorrelated spatial variables with a significant Moran's *I*, eight eigenfunctions were chosen by the forward selection procedure: eigenfunctions numbered 3 and 9 defined broad scale patterns; 13, 21 and 23 encompassed patterns at an intermediate scale; 31, 40, and 44 described a fine

scale of pattern (Appendix C). Together, these spatial eigenfunctions explained 4.1% of the detrended species data. With the addition of the linear trend defined by the site coordinates to the selected eigenfunctions, purely spatial variables had an R^2_a = 16.0% for explaining the non-detrended ground beetle data.

Influence of environment on ground beetles

The three groups of explanatory variables describing the ground beetle community at EMEND were employed for variation partitioning (Figure 2.3a). The descriptors of ground condition were the most important group of explanatory variables (R^2_a = 43.8%). These also explained the largest independent fraction of the variance ([a], R^2_a = 12.0%). Vegetation structure was the second most important group of variables, both independently (fraction [b], R^2_a = 4.6%) and as a group (R^2_a = 34.6%). The selected spatial variables accounted for the least amount of variance (R^2_a = 16.0%). For space alone, the independent fraction [d] was 1.4%.

Interaction among the five sub-groups of descriptors of ground condition is shown in Figure 2.3b. The most important sub-group of variables structuring the carabid community was drainage (R^2_a = 31.1%). The most dominant forest floor cover alone explained 22.9% of the variance in carabid assemblages, while the second and third most important covers explained 16.0% and 3.9%, respectively. Combining the three sub-groups of floor cover variables yielded an overall R^2_a of 39.1%. Topography alone explained 10.3% of the variance in the structure of ground beetle assemblages.

Relationships among the three sub-groups of vegetation structure variables are illustrated in Figure 2.3c. The relative basal area of tree species explained the largest portion of the variance with an R^2_a = 31.5%, followed by forest productivity and forest structure (R^2_a = 18.1%). Finally, indices of tree species diversity were the least important group of variables (R^2_a = 7.3%).

In a partial RDA triplot calculated with space as covariate (Figure 2.4), axis 1 seems to mainly represent a gradient from higher well-drained sites with relief on the right side to lower, flatter and more poorly drained sites on the left side. Axis 2 suggests a gradient from deciduous sites with high leaf litter at the top of the figure to coniferous sites with high amounts of lichen toward the bottom.

Most of the relationships found between carabid species and their habitats (Figure 2.4) are typical of what has been found in boreal forests. *Pterostichus punctatissimus* (Randall) (Pterpunct), for example, was generally found at sites dominated by black spruce that have low relief and are poorly to imperfectly drained. *Agonum gratiosum* (Mannerheim) (Agongrati) and *Platynus mannerheimii* (Dejean) (Platmanne) have very similar habitat requirements; they were found at sites dominated by larch having low relief and poor to somewhat poor drainage. *Agonum retractum* LeConte (Agonretra), *Patrobus foveocollis* (Eschscholtz) (Patrfoveo), *Platynus decentis* (Say) (Platdecen) and *Trechus chalybeus* Dejean (Trecchaly) have similar habitat requirements, and these species were commonly trapped at rapidly drained sites with balsam poplar and paper birch. Leaves were the dominant forest floor cover at these sites. It seems that

topography has little influence on the distributions of these four species. In contrast, *Calathus ingratus* Dejean (Calaingra) and *Pterostichus adstrictus* Eschscholtz (Pteradstr) were found at somewhat rapid to somewhat well-drained sites dominated by aspen, with steeper and higher slopes. *Calathus advena* (LeConte) (Calaadven) was collected mainly in steeper, higher and well-drained sites dominated by white spruce. Finally, *Stereocerus haematopus* (Dejean) (Sterhaema) was found in moderately well drained sites where balsam fir and lodgepole pine are present. Topography had little influence on the distribution of this species.

In the partial RDA triplot, binary variables were re-projected by averaging the corresponding data points. They are projected as points on the plot, not arrows, because they do not represent gradients. Eight axes of the partial RDA were significant, but only the first two are shown in Figure 2.4, together accounting for 25.5% of the variance in R^2_a (Borcard et al. 2011). Although statistically significant, axes 3 to 8 explain so little of the variance in carabid assemblages (2.8%, 1.7%, 1.1%, 0.9%, 0.5%, and 0.4%, respectively) that they cannot be interpreted ecologically. Also, in order to simplify interpretation of Figure 2.4, all species close to the centre of the triplot were removed. The interpretable data were about species sampled at a minimum of 18 sites and these contributed to at least 29% of the variance of all canonical axes. All other species were located consistently close to the centre of the triplot, making any ecological interpretation impossible.

DISCUSSION

Ground beetle assemblages at our study site are structured by both environmental and spatial factors. However, while a combination of environmental (i.e. descriptors of ground condition and vegetation structure) and spatial variables explains half of the variability in species' composition ([abcdefg] in Figure 2.3a), the spatial component is almost entirely shared with the environmental fractions [efg]. This demonstrates that the heterogeneity of resource gradients is the main driver of spatial autocorrelation for this local epigaeic assemblage.

It is well understood that carabid species move away from unsuitable habitat (Niehues et al. 1996, Riecken and Raths 1996). However, the small but significant portion of the overall variance in the data uniquely explained by the spatial variables (1.4%; Figure 2.3a) suggests that neutral dispersal processes also affect to a lesser extent the structure of these communities (Cottenie 2005). This may happen when species are mostly excluded from a local community because of immigration from nearby populations constantly tests the suitability of surrounding habitats.

Alternatively, the portion of the ground beetle data uniquely explained by the spatial variables may reflect response to an environmental gradients not sampled in this study (Peres-Neto and Legendre 2010). Nonetheless, because the fraction attributed solely to space (Figure 2.3a, fraction [c]) is so small compared to all other fractions we are confident that landscape variation in dispersal-limited epigaeic invertebrate communities is mostly controlled by environmental gradients and that neutral dispersal of species contributes little to structuring local

assemblages. Because epigaeic organisms appear to disperse mostly in relation to environmental gradients, even coarse filter biodiversity conservation strategies employed in forest land management should focus on observable environmental factors known to evoke biotic responses.

Overall, the variance explained uniquely by the spatial variables (fractions [cefg]) is quite small compared to that explained by the other groups of variables. In itself, this is an interesting finding that may be partly attributed to the scale at which this study was carried out. As noted in the *Introduction*, most studies of ground beetle community patterns have been conducted at much finer scale. Given the broad scale of our study, we can confidently conclude the spatial distribution of carabids is not strongly autocorrelated on the landscape scale, but rather it is the heterogeneity of the habitat that was partly autocorrelated, and this in turn seems to be a main influence on ground beetles distribution.

We discovered that carabid species composition was quite distinctive in the only two forest remnants isolated by harvesting. Note that edge effect was not enough to characterize carabid species composition at these two sites because some sites located in larger intact forest patches were as close to a disturbance as the two isolated sites but did not have a carabid species composition that differed markedly from sites located far from any disturbances. This result supports the practice of connecting forest patches retained in harvested areas with unharvested areas in efforts to conserve the local structure of epigaeic assemblages characteristic of mature forest.

Although a low density of roads and seismic lines, as well as a variable retention harvest experiment are included on the studied landscape, we did not identify significant broad effects of such disturbances on the spatial organisation of carabid assemblages. Thus, the much discussed conservation issue of maintaining forest connectivity on working boreal landscapes, may simply involve determining tolerable disturbance thresholds. For carabids and likely for other epigaeic arthropods with similar lifestyles using forest-floor habitats that remain after harvest, such thresholds appear to be relatively high, at least in the immediate aftermath of variable retention harvests. Studies of longer duration will be required to delimit harvesting thresholds that support full recovery of biodiversity on local sites over the longer term (Work et al. 2010). Our results, however, suggest that the overall level of anthropogenic disturbance on the EMEND landscape (27.4%) is not associated with detectable landscape effects on the epigaeic fauna in forest that remains undisturbed.

Use of Moran's Eigenvector Maps for investigating organization of insect assemblages is a salient and innovative feature of this study. MEMs using a second order concave-up weighting function (f_1) effectively represented ground beetle species of mature boreal forest, probably because these species have restricted dispersal abilities (Larochelle and Larivière 2003). This pattern suggests that spatial processes driving the distribution of such epigaeic invertebrates in forested landscapes can act on a very local scale and that their dispersal abilities are constrained to short distances, at least when considered at the landscape scale.

Thus, it is not surprising to find highly heterogeneous epigaeic faunas within relatively small areas (Niemelä et al. 1992, Niemelä and Spence 1994).

The selected MEMs highlighted significant spatial patterns in carabid assemblages on the boreal landscape, allowing us to identify the spatial scale at which the carabid community was structured. The finest scales of autocorrelation detected in our study represent spatial patterns of roughly 2 km (MEMs 31, 40 and 44; Appendix C). All other selected eigenfunctions describe patterns at broader scales. This is an indication that even when dealing with an epigaeic fauna strongly affected by local conditions, patterns at a scale of at least 2 km can clearly affect community structure (Niemelä and Spence 1994). Furthermore, the fact that we found fine (MEMs 31, 40 and 44; Appendix C), medium (MEMs 13, 21 and 23; Appendix C) and broad (MEMs 3 and 9, Appendix C) scale spatial patterns emphasises the importance of considering a variety of scales in both ecological studies and forest management.

Interestingly, anthropogenic disturbances other than forest harvest that completely segregates small forest patches in large harvested blocks seem to have little effect on the spatial distribution of carabids in mature forests. At the scale of this study, modest fragmentation through roads and seismic lines registered no detectable influence on the spatial structure of these epigaeic invertebrates. However, we note that the results of a PCA performed on the ground beetle assemblage, which isolated the two sites in forest retention plots, showed clear differences in species composition in patches with radius ≤ 40 m.

It is noteworthy that the factors found to be influential in our analyses do not operate independently but interact to determine composition and structure of the beetle assemblages. This is clearly shown by the large overlaps in explanatory power among the spatial variables, the descriptors of ground condition, and vegetation structure in Figure 2.3a (fraction [dfg], $R^2_a = 31.9\%$). Thus, understanding of singular cause and effect remains elusive and we favour the hypothesis that complex interactions are the norm for most ecological systems (Pickett et al. 2007). This hypothesis can, and should, be used as a null hypothesis to increase understanding of factors structuring ecological communities such as these boreal ground beetles. A next logical step is to decipher these complex interactions through careful experiments coupled to theoretical development.

The importance of variables reflecting substrate characteristics and vegetation structure for understanding the structure of carabid assemblages is consistent with results of previous studies (Thiele 1977, Lövei and Sunderland 1996). However, the close association found here between these two groups of explanatory variables (Figure 2.3a, fractions [dfg], $R_a^2 = 31.9.\%$) illustrates well that their interactive influence on the ground beetle community cannot be effectively separated. This is not surprising. Trees contribute significantly to soil development and soil reciprocally influences growth of tree species (Perry 2008). It is thus expected to find concomitant variation in spatial patterns of soil and trees. The ecological link between ground beetles and vegetation or soil is not understood well enough to identify the mechanisms responsible for the patterns observed. However, the fact that these two groups of variables explained the

largest proportion of the variance in beetle assemblages should help focus a new generation of hypotheses about how such factors influence the structure of carabid and other epigaeic assemblages.

Among the descriptors of ground condition, soil drainage and forest floor cover were the most important. Interestingly, topography itself showed little impact on carabid assemblages (Figure 2.3b fraction [a], R_a^2 = 1.6%). Although drainage explained 31.1% of the variance (Figure 2.3b fraction [bgnpuwAD]), only a small portion (3.4%, fraction [b]) was explained solely by this factor. Taken together, information about the three dominant floor covers was the best of the descriptors of ground condition for defining carabid assemblages. This is consistent with the suggestion that epigaeic beetles depend greatly on the edaphic factors prevailing in their environment (Lövei and Sunderland 1996). The first dominant floor cover explained not only the most variance (Figure 2.3b fraction [cghpA], R_a^2 = 22.9%), but its independent fraction explained more variation in the carabid data than any other variable ([c], R_a^2 = 7.3%). The 2nd dominant floor cover was less important than the 1st, and the 3rd dominant forest floor cover showed negligible ecological (and statistical) significance. Many levels of the same forest floor cover types cluster together in the partial RDA (Figure 2.4), suggesting that subdominant floor covers frequently occupy a large enough area to support beetle species associated with them. A test of this hypothesis awaits investigation at a finer scale than considered here.

The variables defining vegetation structure were chosen to understand the potential influence of trees species composition and diversity, stand structure and

tree productivity on ground beetle assemblages. The composite variable that we have called 'tree relative basal area' was quite important in structuring the ground beetle assemblages (Figure 2.3c, fractions [bdg], R^2_a = 31.5%). Although tree relative basal area explained more of the carabid data than other descriptors of vegetation structure, forest structure and productivity also contributed notably in defining the boreal forest beetle assemblage (Figure 2.3c, fractions [adg], R^2_a = 18.1%). In fact, the larger number of beetle species associated with high forest productivity and tree diversity (Shannon index) in Figure 2.4 suggests that highly productive forests rich in tree species support more diverse ground beetle assemblages. Because harvesting operations generally target highly productive sites, unharvested residuals relegated by default to poorer sites may be insufficient to meet conservation goals. Furthermore, our results together with those of Bergeron et al. (2012) suggest that reforestation operations that maximize the match between pre- and post-harvest species diversity of boreal trees will more effectively maintain ground beetle diversity.

By considering descriptors of ground condition and vegetation structure together, patterns in carabid assemblages are more easily understood. The dynamic characteristics of boreal forests generate distinct habitats that have influenced evolution of particular invertebrate species. The sustainable forest management perspective being adopted across the Canadian boreal forest and elsewhere demands that this diversity of habitats be maintained in the wake of industrial activity on these landscapes if biodiversity is to be maintained. Our study underscores that habitat diversity is critical for conserving the full range of

ground beetle diversity that occupies the forest. Research can provide further guidance. Although the ground characteristics were somewhat stronger predictors of assemblage structure here, vegetation structure is likely easier to manage and other studies at EMEND have identified a sound rationale for doing so (Bergeron et al. 2011, 2012).

Applying variation partitioning to study the relationship between environmental variables and space, as we did in Figure 2.3a, has recently been criticized by Gilbert and Bennett (2010) and Smith and Lundholm (2010). However, both papers agree that it can yield useful results if used carefully, and we hold that the present analysis supports optimism about this approach.

Gilbert and Bennett (2010) explain that both sampling scale and analytical techniques should be carefully chosen to answer the ecological question at hand. They suggest, in particular, that the literature about spatial sampling should be consulted to explore the trade-offs amongst sampling designs for particular study organisms. Our study provides data highly germane for design of sampling regimes for carabids. Moreover, the landscape scale at which our study was carried out is unique with respect to carabid studies in general, and allowed us to evaluate the relationship between space and environment using variation partitioning.

Smith and Lundholm (2010) show that the common fraction of variance explained by environment and space is of utmost importance and should not be neglected because it represents patterns generated by both environmental factors and dispersal limitation. They recognize that at some scale all environmental

patterns are spatially correlated. In our paper, we carefully chose the scale at which we sampled ground beetles to focus on landscape patterns. Literature in boreal forest ecology (e.g., Beckingham 1996, and Kimmins 2004) strongly suggests that the environmental variables we measured are spatially autocorrelated at the scale of our study. We controlled for space specifically to obtain the ordination result in Figure 2.4, so as to more clearly depict the independent effects of various environmental factors. Many of the relationships we found between the ground beetle distributions and the different environmental constraints are consistent with other studies (Lindroth 1961–1969, and Larochelle and Larivière 2003), validating to some extent the results of the variation partitioning used here.

Patterns described for the carabid community studied here suggest that composition of local assemblages is strongly influenced at several scales on boreal landscapes by interactions among vegetation structure, soil drainage and forest floor cover. Conversely, assemblages were not markedly influenced by anthropogenic disturbances (harvesting, roads and seismic lines) on the same landscape. Because of its unusually broad scale, this study contributes a new much needed perspective that is highly relevant to improved understanding of arthropod communities on forest landscapes.

The statistical methodology used to obtain our conclusions provides insights to improve understanding about why particular carabid species were found at particular sites, and were absent from others on the landscape. These results connect statistical theory, now the backbone of spatial ecology, to natural

history and provide confidence that our process-oriented investigations are not missing much that is important. Although uncommon in epigaeic invertebrate community research, the methods used here have a solid foundation in the statistical literature and have overdue potential to inject more rigorous spatial reasoning into work about distributions of particular organisms in nature. From an ecological perspective, our study suggests that underlying gradients in environmental factors on landscapes regenerating after harvest will determine the structure and spatial distribution of organisms like these epigaeic invertebrates. And thus, further attention to understanding these gradients, their spatial configuration and their impacts on the biota will contribute significantly to both the science of spatial ecology and to making better predictions useful for managing boreal landscapes in a manner that conserves biodiversity.

LITERATURE CITED

Beckingham, J. D., I. G. W. Corns, and J. H. Archibald, 1996. Field guide to ecosites of West-central Alberta. Special report 9, Natural resources of Canada, Canadian forest service, Northwest region, Northern forestry center, Edmonton, Alberta.

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

Bergeron, J. A. C., J. R. Spence, and W. J. A. Volney. 2011. Landscape patterns of species-level association between ground-beetles and overstory trees in

boreal forests of western Canada (Coleoptera, Carabidae). ZooKeys **147**:577–600.

Blanchet, F. G., P. Legendre, and D. Borcard. 2008. Forward selection of explanatory spatial variables. Ecology **89**:2623–2632.

Borcard, D., F. Gillet, and P. Legendre. 2011. Numerical Ecology with R. Use R!, Springer, New York.

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

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

Burton, P. J., C. Messier, D. W. Smith, and W. L. Adamowicz. 2003. Towards Sustainable Management of Boreal Forest, NRC Research Press.

Buddle, C. M., D. W. Langor, G. R. Pohl, and J. R. Spence. 2006. Arthropod responses to harvesting and wildfire: Implications for emulation of natural disturbance in forest management. Biological Conservation **128**: 346–357.

Connell, J. H., T. P. Hughes, and C. C. Wallace. 1997. A 30-year study of coral abundance, recruitment, and disturbance at several scales in space and time. Ecological Monographs **67**:461–488.

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

Davies, K. F., and C. Margules. 1998. Effects of habitat fragmentation on carabid beetles: experimental evidence. Journal of Animal Ecology **67**: 460–471.

Davies, P. T., and M. K. S. Tso. 1982. Procedures for reduced-rank regression. Applied Statistics **31**:244–255.

Digweed, S. C., C. R. Currie, H. A. Carcamo, and J. Spence. 1995. Digging out the ”digging-in effect” of pitfall traps: Influences depletion and disturbance on catches of ground beetles (Coleoptera: Carabidae). Pedobiologia **39**:561–576.

Dray, S., 2011. spacemakeR: Spatial modelling, version 0.0-5. URL `http://R-Forge.R-project.org/projects/sedar/`.

Dray, S., P. Legendre, and F. G. Blanchet, 2011. packfor: Forward Selection with permutation (Canoco p.46), version 0.0-7/r58. URL `http://R-Forge.R-project.org/projects/sedar/`.

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

Gilbert, B., and J. R. Bennett, J. R. 2010. Partitioning variation in ecological communities: do the numbers add up? Journal of Applied Ecology **47**:1071–1082.

Harvey, B. D., Y. Bergeron, A. Leduc, and S. Gauthier. 1997. Sylviculture et aménagement forestier écosystémique, peut-on concilier les deux ? L'exemple de la forêt boréale mixte de l'Abitibi. Pages 22-24, 31 (Vol. 121-122) dans L’Aubelle. Ressources naturelles Canada, Service canadien des forêts, Centre de foresterie des Laurentides, Sainte-Foy (Québec).

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

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

Kimmins 2004. Forest Ecology - A Foundation for Sustainable Forest Management and Environmental Ethics in Forestry, fourth edn. Prentice Hall, New Jersey.

Kleyer, M., R. Biedermann, K. Henle, E. Obermaier, H.-J. Poethke, P. Poschlod, B. Schroeder, J. Settele, and D. Vetterlein. 2007. Mosaic cycles in agricultural landscapes of Northwest Europe. Basic and Applied Ecology **8**:295–309.

Larochelle, A., and M.-C. Larivière. 2003. A natural history of the ground-beetles (Coleoptera: Carabidae) of America north of Mexico. Pensoft Publishers.

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

Legendre, P. 2011. const.clust: Space- and time-constrained clustering package.

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. 2012. Numerical Ecology. third edition. Elsevier, Amsterdam.

Lindenmayer, D. B. and J. F. Franklin. 2003. Towards Forest Sustainability, Island Press.

Lindroth, C. H. 1961–1969. The ground beetles (Carabidae, excl. Cincindelinae) of Canada and Alaska. – Opusc. Entomol. Suppl. 20 (1961): 1–200; 24 (1931): 201–408; 29 (1966): 409–648; 33 (1968): 649–944; 34 (1969): 945–1192; 35 (1969): I–XLVII.

Lövei, G. L., and K. D. Sunderland. 1996. Ecology and behavior of ground beetles (Coleoptera: Carabidae). Annual Review of Entomology **41**:231–256.

Nieheus, F.-J., P. Hockmann, and F. Weber et al. 1996. Genetics and dynamics of a *Carabus auronitens* metapopulation in the Westphalian Lowlands (Coleoptera: Carabidae). Annales Zoologici Fennici **33**: 85–96.

Niemelä J. 1997. Invertebrates and boreal forest management. Conservation Biology **11**: 601–610.

Niemelä, J., Y. Haila, E. Halme, T. Pajunen, and P. Punttila. 1992. Small-scale heterogeneity in the spatial distribution of carabid beetles in the southern Finnish taiga. Journal of Biogeography **19**: 173–181.

Niemelä, J., D. W. Langor, and J. R. Spence. 1993. Effects of clear-cut harvesting on boreal ground-beetle assemblages (Coleoptera: Carabidae) in western Canada. Conservation Biology **7**:551–561.

Niemelä, J. and Spence, J. R. 1994. Distribution of forest dwelling carabids (Coleoptera): spatial scale and the concept of communities. Ecography **17**:166–175.

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

Parisien, M.-A., and M. A. Moritz. 2009. Environmental controls on the distribution of wildfire at multiple spatial scales. Ecological Monographs **79**:127–154.

Pearce, J. L., and L. A. Venier. 2006. The use of ground beetles (Coleoptera : Carabidae) and spiders (Araneae) as bioindicators of sustainable forest management: A review. Ecological Indicators **6**:780–793.

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

Peres-Neto, P. R., P. Legendre, S. Dray, and D. Borcard. 2006. Variation partitioning of species data matrices: Estimation and comparison of fractions. Ecology **87**:2614–2625.

Perry, D.A., R. Oren, S. C. Hart. 2008. Forest ecosystems, second edn. Johns Hopkins University Press, Baltimore, MD.

Pickett, S. T., J. Kolasa, and C. G. Jones. 2007. Ecological understanding, the nature of theory and the theory of nature, second edition. Academic Press, Burlington, MA.

Pielou, E. C. 1966. Measurement of diversity in different types of biological collections. Journal of Theoretical Biology **13**:131–144.

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

Rainio, J., and J. Niemelä. 2003. Ground beetles (Coleoptera: Carabidae) as bioindicators. Biodiversity and Conservation **12**:487–506.

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

Riecken, U. and U. Raths. 1996. Use of radio telemetry for studying the dispersal and habitat use of *Carabus coriaceus* L. Annales Zoologici Fennici **33**: 109–116.

Rowe, J. S. 1972. Forest regions of Canada. Department of Fisheries and the Environment, Canadian Forestry Service.

Shannon, C. E. 1948. A mathematical theory of communication. Bell System Technical Journal **27**:379–423.

Šidák, Z. 1967. Rectangular Confidence Regions for the Means of Multivariate Normal Distributions. Journal of the American Statistical Association **62**:626–633.

Smith T. W. and J. T. Lundholm. 2010. Variation partitioning as a tool to distinguish between niche and neutral processes. Ecography **33**:648–655.

Sousa, W. P. 1984. The role of disturbance in natural communities. Annual Review of Ecology and Systematics **15**:353–391.

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

Spence, J. R., W. J. A. Volney, V. J. Lieffers, M. G. Weber, S. A. Luchkow, and T. W. Vinge, 1999. The Alberta EMEND project: recipe and cooks' argument. Pages 583–590 in The sustainable forest management network conference - Science and practice: Sustaining the boreal forest.

Thiele, H.-U. 1977. Carabid Beetles in their Environments. Springer-Verlag, Berlin.

Vanbergen A. J., B. A. Woodcock, A. D. Watt, and J. Niemelä. 2005. Effect of land-use heterogeneity on carabid communities at the landscape scale. Ecography **28**:3–16.

Ward, J. H. 1963. Hierarchical grouping to optimize an objective function. Journal of the American Statistical Association **58**:236–244.

Wiens, J. A. 1989. Spatial scaling in ecology. Functional Ecology **3**:385–397.

Woodcock, B. A., J. Redhead, A. J. Vanbergen, L. Hulmes, S. Hulmes, J. Peyton, M. Nowakowski, R. F. Pywell, and M. S. Heard. 2010. Impact of habitat type and landscape structure on biomass, species richness and functional diversity of ground beetles – Agriculture, Ecosystems and Environment **139**:181–186.

Work, T. T., D. P. Shorthouse, J. R. Spence, W. J. A. Volney, and D. Langor. 2004. Stand composition and structure of the boreal mixedwood and epigaeic arthropods of the Ecosystem Management Emulating Natural Disturbance (EMEND) landbase in northwestern Alberta. Canadian Journal of Forest Research **34**:417–430.

Work, T. T, J. M. Joshua, J. R. Spence, and W. J. Volney. 2010. Higher levels of variable retention required to maintain ground beetle biodiversity in boreal mixedwood forests. Ecological Application **20**: 741-751.

Zar, J. H. 1999. Biostatistical Analysis. Fourth edition. Prentice Hall.

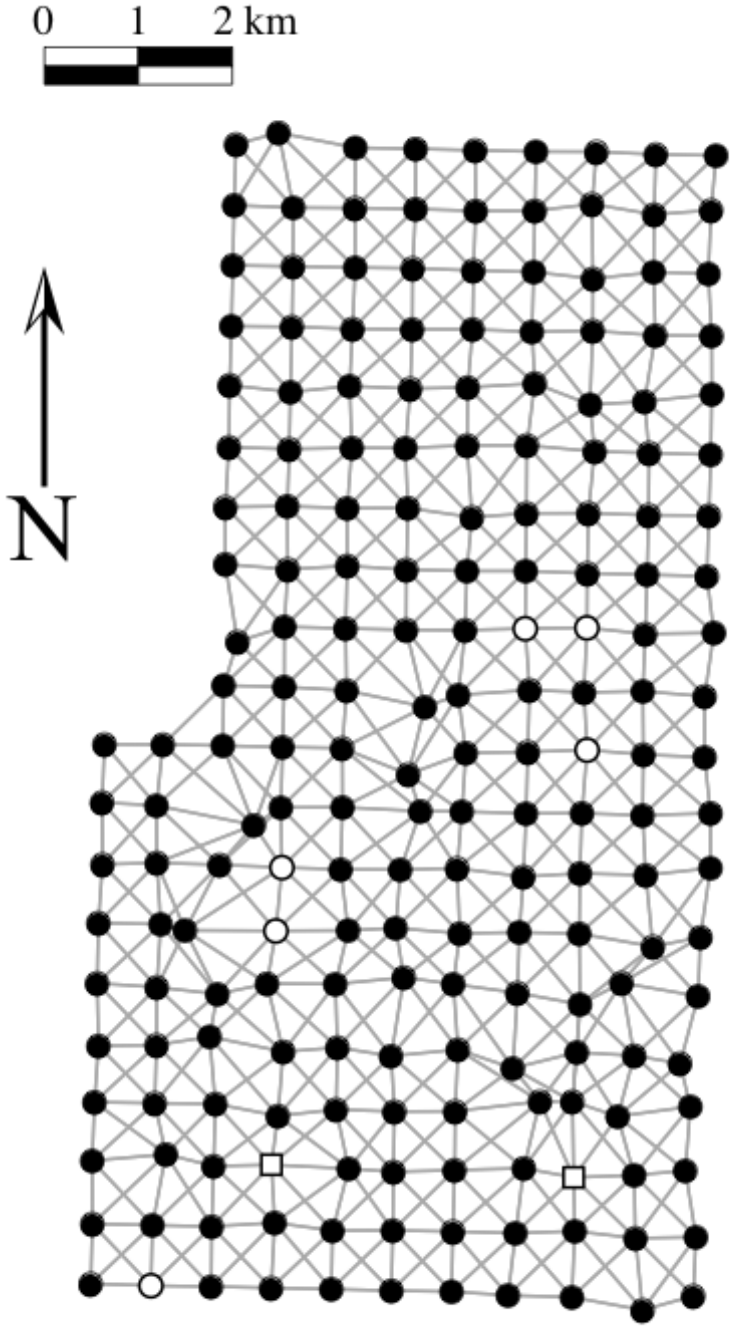

FIGURE 2.1 Map of the studied area. Full circle represent sampled sites, empty squares are outlier sites, and empty circles illustrate fictitious sites added to ensure continuity in the construction of spatial variables. The lines linking the sites represent the connexion diagram used to perform the spatially constrained clustering and to construct the spatial eigenfunctions.

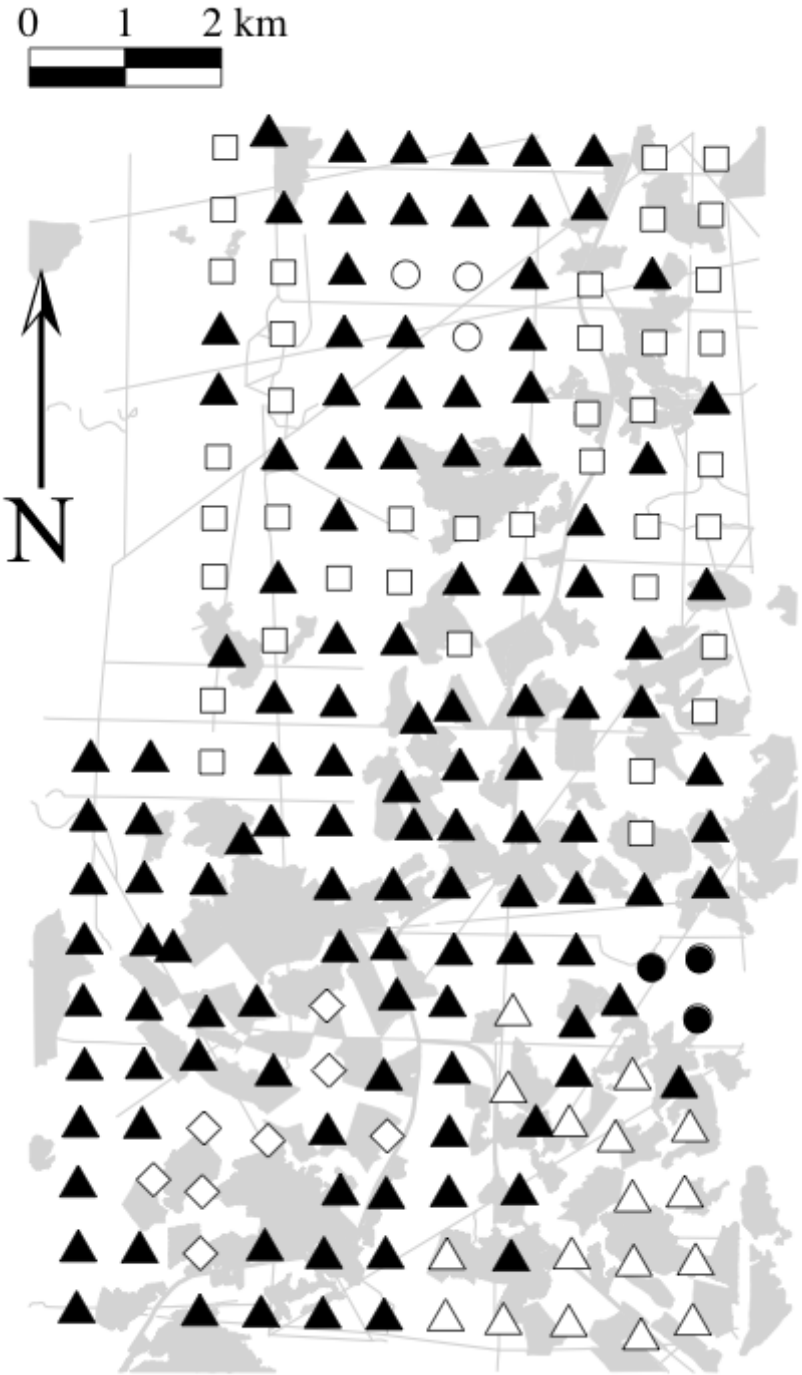

FIGURE 2.2 Spatially constrained clustering constructed using Ward's criterion constrained by the connexion diagram presented in Figure 2.1. The six groups solution yielded the lowest cross-validation residual error (0.614). Each group is defined by a specific symbol. Anthropogenic disturbances occurring in the landscape are illustrated in light grey in the background of the figure.

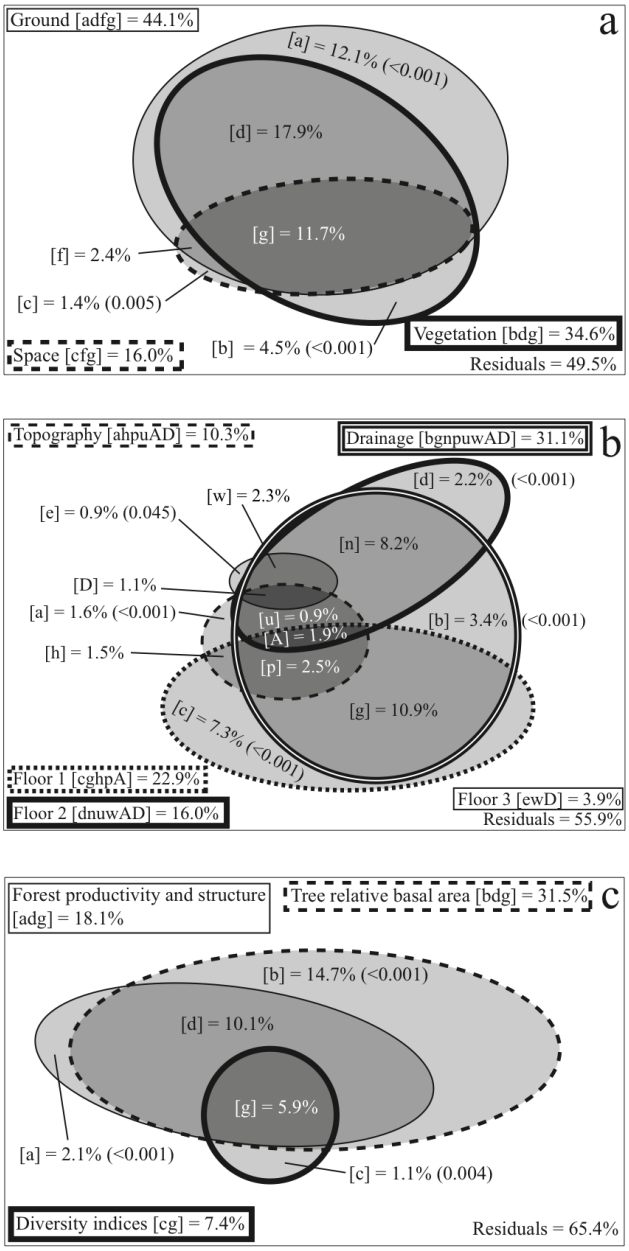

FIGURE 2.3 (a) Venn diagrams presenting the results of the variation partitioning between descriptors of ground condition, vegetation structure and the spatial structure (Moran's eigenvector maps and the sites spatial coordinates). (b) Venn diagram presenting the variation partitioning results between soil drainage, the three sets of forest floor cover variables (Floor 1, 2 and 3 are the 1st, 2nd and 3rd most dominant floor cover), and topography. These variables are only representing descriptors of ground condition. (c) Venn diagram presenting the variation partitioning results between tree relative basal area by species and forest productivity (mean tree diameter at breast height [DBH]) and structure (coefficient of variation of tree DBH). The Pielou and Shannon indices were also calculated on the tree species. These variables are only representing vegetation structure. Fraction sizes for all Venn diagrams are approximations. Appendix B presents the conceptual representation the two types of Venn diagram used in the variation partitioning presented in this figure. All fractions that had an R_a^2 smaller than 1% were not plotted.

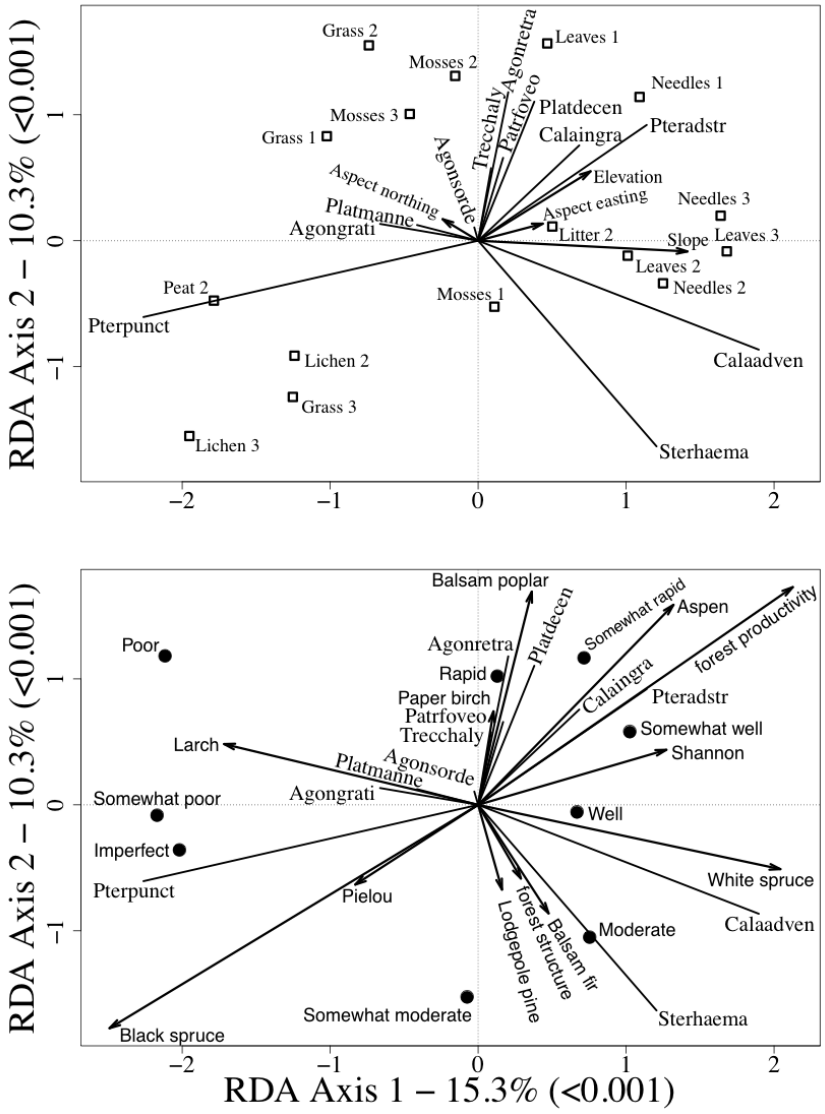

FIGURE 2.4 Partial canonical redundancy analysis (partial RDA) of the carabid species with spatial structure (Moran's eigenvector maps and the sites spatial coordinates) as covariate. Results are presented in two triplots to make them easier to read. Top plot – Forest floor cover is represented by empty squares and topography by the black arrows. Floor covers followed by 1, 2 and 3 are the 1st, 2nd and 3rd most dominant, respectively. Bottom plot – Drainage is represented by full circles and vegetation structure variables are the black arrows. Drainage is defined following Appendix A. Forest productivity is the mean tree diameter at breast height (DBH) and forest structure is the coefficient of variation of the trees DBH. Carabid species are represented in both plots by segments. Ground beetles species close to the center of the triplot were removed. The remaining species were sampled at a minimum of 18 sites and contributed to at least 29% of the variance of all canonical axes. Agongrati = *Agonum gratiosum* (Mannerheim), Agonretra = *Agonum retractum* LeConte, Calaadven = *Calathus advena* (LeConte), Calaingra = *Calathus ingratus* Dejean, Patrfoveo = *Patrobus foveocollis* (Eschscholtz), Platdecen = *Platynus decentis* (Say), Platmann = *Platynus mannerheimi* Dejean, Pteradstr = *Pterostichus adstrictus* Eschscholtz, Pterpunct = *Pterostichus punctatissimus* (Randall), Sterhaema = *Stereocerus haematopus (*Dejean), Trecchaly = *Trechus chalybeus* Dejean. Each axis represents the variance in R^2_a. The dimensions of the triplot are proportional to the explained variance of each axis. The partial RDA was drawn using a correlation triplot.

APPENDIX 2A

TABLE 2A1. Soil drainage classification used in the study in comparison with Beckingham et al. (1996) classification.

Beckingham et al. (1996) classification	Drainage classification used in this study
Very Rapidly Drained	Extremely Rapidly Drained
	Very Rapidly Drained
Rapidly Drained	Rapidly Drained
	Somewhat Rapidly Drained
Well Drained	Well Drained
	Somewhat Well Drained
Moderately Well Drained	Moderately Well Drained
	Somewhat Moderately Well Drained
Imperfectly Drained	Imperfectly Drained
	Somewhat Poorly Drained
Poorly Drained	Poorly Drained
	Very Poorly Drained
Very Poorly Drained	Extremely Poorly Drained

APPENDIX 2B

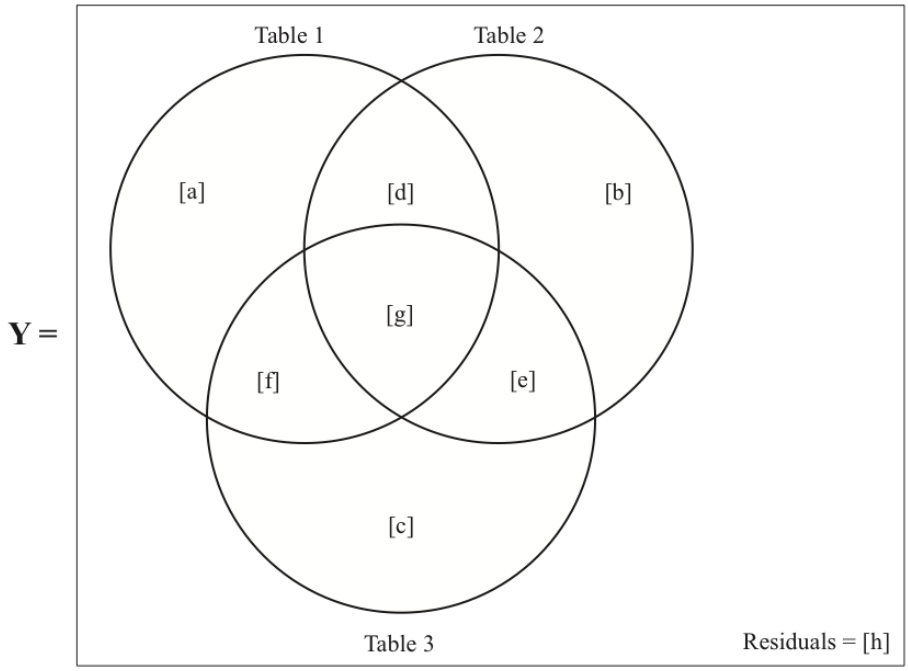

FIGURE 2B1 Conceptual Venn diagram presenting the variation partitioning results between three groups of explanatory variables. Each letter presents an independent fraction of explained variance.

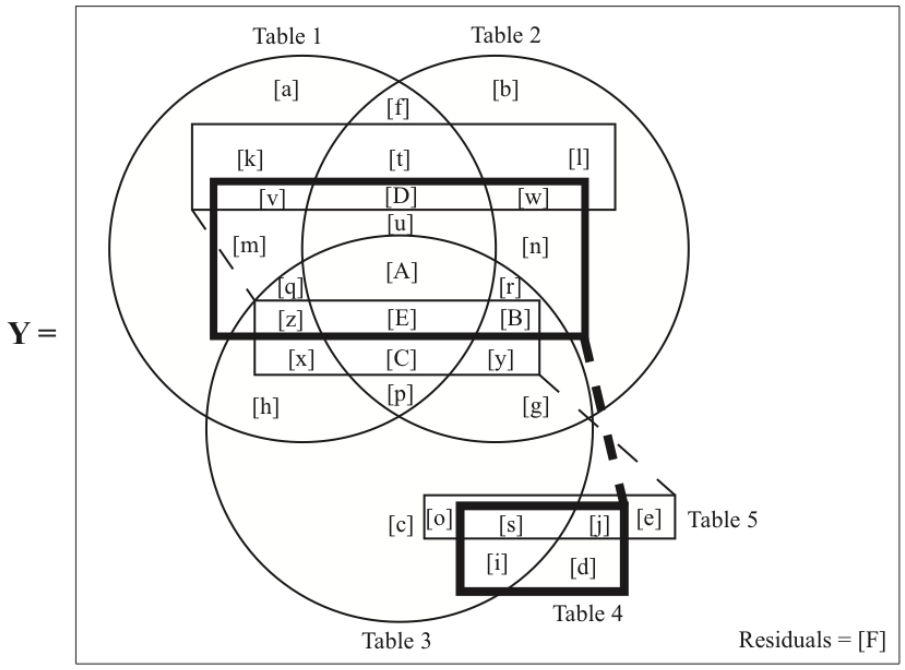

FIGURE 2B2 Conceptual Venn diagram presenting the variation partitioning results between five groups of explanatory variables. The contribution of the fourth set of explanatory variables is illustrated by the two rectangles in bold. The contribution of fifth set of explanatory variables is illustrated by the three rectangles non-bold rectangles. Each letter presents an independent fraction of explained variance.

APPENDIX 2C

TABLE 2C1 R^2_a and significance of spatial models (MEM) constructed with the different values of α, f_1 and f_2 proposed in the main text. The ground beetle data were Hellinger transformed. The MEM models were computed on detrended data. The results presented for MEMs refer to a subset of eigenfunctions measuring positive autocorrelation only (all variables in the subset were associated to a positive Moran's *I*).

Distance function	α	R^2_a	P	P_{corr}
f_1	1	4.8%	< 0.001	0.002
	2	**5.8%**	**< 0.001**	**0.002**
f_2	1	4.3%	0.002	0.004
	2	4.3%	< 0.001	0.002
Binary		4.3%	< 0.001	0.002

* A Šidák correction was applied to the *P*-values.

** Bold highlights the weight used to compute the results reported in the main text.

Note: The subset MEMs measuring negative autocorrelation are not presented because they yield a corrected *P*-value always equal to 1.

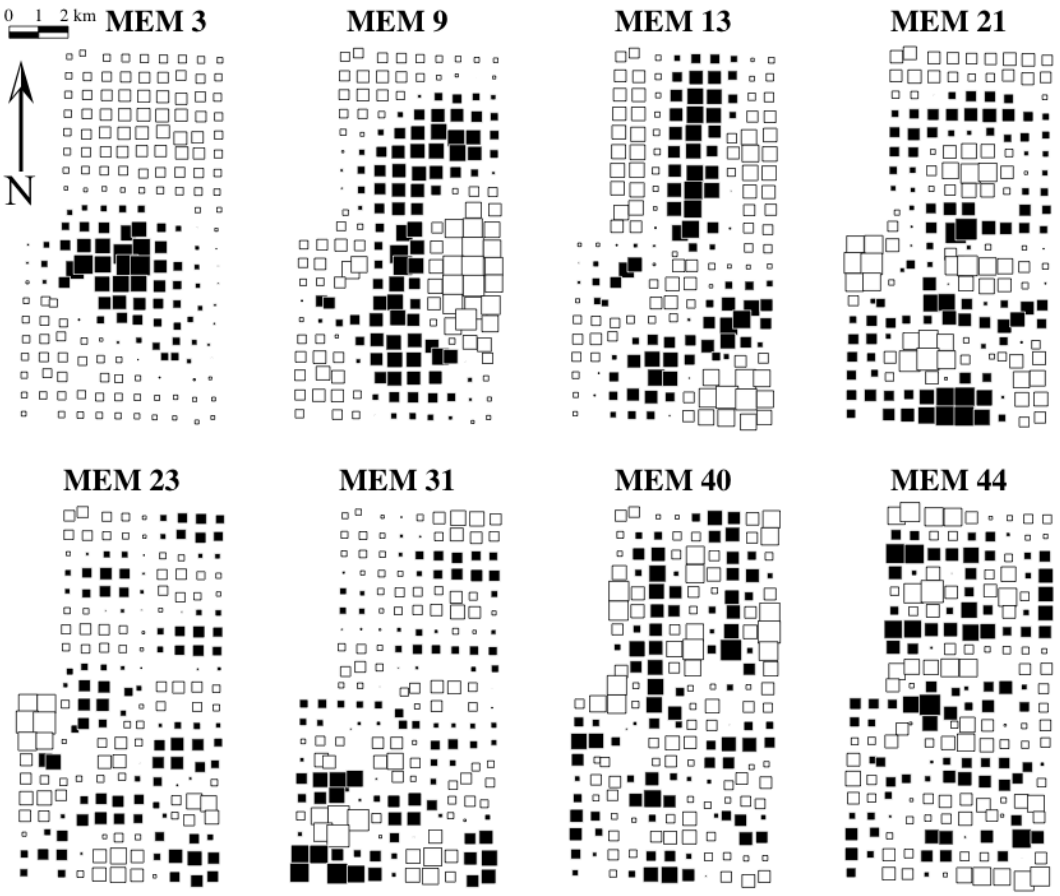

FIGURE C1 MEM eigenfunctions selected to model ground beetle distribution. The square bubble size is proportional to the value associated to it, whereas the color reflects the sign of the value associated to the bubble (black = positive, white = negative).

Chapter 3 – Consensus RDA across association coefficients for canonical ordination of community composition data

INTRODUCTION

The species composition of an ecological community is heavily influenced by local variation in habitats. In theory, this intimate species-habitat relationship results because many characteristics of the environment influence organisms and these influences structure the niches of the species living together in the same community (Hutchinson 1957). Whittaker (1972) expanded this idea using the concept of environmental gradient. These gradients are essentially spatial and 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 using a complex configuration of local habitats are composed of a few common species, and a large proportion being less abundant, enve quite rare. 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 observed for each species in samples from a community. SADs were first mathematically described in the pioneering papers of Fisher et al. (1943) and Preston (1948). McGill et al. (2007)

review the various types of SADs and explain 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) and correspondence analysis (Hirschfeld 1935) 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) have been used to more directly evaluate how specific habitat components affect species assemblages. Two variants of RDA have also been proposed to ecologists in the last decade: distance-base RDA (db-RDA, Legendre and Anderson 1999) and transformation-based RDA (tb-RDA, Legendre and Gallagher 2001). These two variants, in contrast with earlier approaches, make it possible to use a range of association coefficients to perform canonical ordinations, offering much more flexibility in analysing community data. An association coefficient assesses the resemblance in species composition among sampled sites by condensing the community data into a symmetric square matrix of association among sites (using either similarity **S** or distance **D**). For example, the Euclidean distance (Table 3.1) uses Pythagore's formula between all pairs of sites, which result in a symmetric square matrix where the information from the species between two sites is condensed into one value.

Choosing an association coefficient well suited to study of a particular community and particular 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 association coefficients 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 association coefficient to use. Under this procedure, the abundance structure of a community can influence the selection of association coefficients used to describe it.

Although variation in SADs complicates coefficient selection, little is known about how variations in SADs affect the performance of association coefficients. In this study, we compare the performance of association coefficients most commonly used in canonical ordinations 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 performance of association coefficients within data type, we show that the choice of coefficients 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 association coefficients. Secondly, by comparing association 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 the boreal forest of northwestern Alberta, Canada.

DEFINING A COMMUNITY WITH A SAD

There are many ways of displaying a SAD. In this paper, we use a variation on 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 presented in Figure 3.1 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 total number of species was fixed at 20 but the total abundance varied from 261 (the sum of the abundance classes' lower bound for each species of the community depicted by Figure 3.1a) to 20460 (the sum of the abundance classes' upper bound for each species of the community depicted in Figure 3.1j). Therefore, the SADs of Figure

3.1 represent a huge variation of species-abundance distribution 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 3.1a-b present communities with the largest number of rare species. Note that communities with a larger number of rare species are often found in nature, however because the SADs in Figure 3.1 will later be used to define the abundance of species in simulated communities, the SADs in Figure 3.1a-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 during 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. Depending on the group of organisms studied, removing rare species can yield SADs similar to what is found in Figure 3.1c-g, m-o, u-v.

In a recent paper, Gaston (2011) emphasized the importance of also studying common species. In light of this work, we included a few SADs (Figure 3.1h-j, w-y) that define 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 the ones in Figure 3.1k-l. Finally, the SADs presented in Figure 3p-t are mainly theoretical and unlikely to be found in nature. We chose them because analysis of such extreme cases may lead to a better understanding of association coefficients.

RDA AND ASSOCIATION COEFFICIENTS

In this study, we used the RDA framework to compare commonly used association coefficients (Table 3.1), 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 association coefficients are mathematically equivalent in tb-RDA and db-RDA (Legendre and Legendre 2012).

Euclidean distance is linearly related to the square root of the complement of the simple-matching coefficient (first entry of Table 3.1). This relationship was established by Gower (1966) when he described PCA based on binary descriptors. PCA based on binary data produces the same ordination as the principal coordinate analysis of a matrix of $\sqrt{1 - \text{simple matching coefficient}}$. 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 association 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

$$\begin{aligned} \mathbf{B} &= (\mathbf{X}^t\mathbf{X})^{-1}\mathbf{X}^t\mathbf{Y} \\ \widehat{\mathbf{Y}} &= \mathbf{X}\mathbf{B} \end{aligned} \tag{3.1}$$

where *t* is the transpose of a matrix 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 3.1, **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 3.2:

$$\mathbf{Y}_{\mathbf{res}} = \mathbf{Y} - \widehat{\mathbf{Y}} \tag{3.2}$$

By performing a PCA on $\widehat{\mathbf{Y}}$, 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** (3.3.1) or **Y** (3.3.2).

$$\mathbf{Z} = \mathbf{X}\mathbf{B}\mathbf{U} = \widehat{\mathbf{Y}}\mathbf{U} \tag{3.3.1}$$

$$\mathbf{F} = \mathbf{Y}\mathbf{U} \tag{3.3.2}$$

If required, the canonical coefficients can be calculated following Equation 3.4:

$$\mathbf{C} = \mathbf{BU} \tag{3.4}$$

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, an association coefficient is applied to a community matrix, yielding a dissimilarity matrix. A principal coordinate analysis (PCoA, Gower 1966) 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 3.3) and canonical coefficients (Equation 3.4) 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:

$$\mathbf{U} = \frac{\mathbf{Y}^t\mathbf{Z}\mathbf{\Lambda}^{-1/2}}{\sqrt{n-1}} \tag{3.5}$$

All binary 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

because two sites with exactly the same species will not necessarily result in a dissimilarity of 0; 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 present a solution to the double-zero problem.

The double-zero problem stems from the difficulty in relating two sites where a species was not found (Legendre and Legendre 2012, Subsection 7.2.2). Asymmetrical association coefficients are designed to ignore double zeros altogether; for binary association coefficients, this amounts to ignoring the value *d* (Table 3.2) in the calculation of association coefficients. Conversely, symmetrical coefficients treat double zeros as any other relation. For example, for presence-absence community data, double zeros (or double absences) are considered as an indication of similarity in the same way as double presence (value *a* in Table 3.2). Symmetrical coefficients should be used only when the goal of a study is to evaluate total changes in a community such as the effect 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).

In the present study, we performed simulations that reflected species variation in undisturbed communities and where predation was not considered. In that instance, the Euclidean and simple matching coefficients are ill adapted to these types of ecological problems because they are 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, often 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. 2011) with the exception of the Raup-Crick coefficient, which was programmed independently using McCoy et al. (1986) permutation procedure. We used McCoy et al. (1986) permutation approach following Legendre and Legendre (2012, Subsection 7.3.5) who 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 of 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 3A1 of Appendix 3A 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 3.6, 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.

$$\mathbf{y}_{ref} = \omega\left(\mathbf{x}_i + \mathbf{x}_j\right) + \varepsilon \tag{3.6}$$

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 lower weights representing weak relationships (if $\omega = 1$ or -1) or an absence of relationship (if $|\omega| < 1$), between a species and the descriptors defining it. There are no relationships for species with $|\omega| < 1$ because the standard deviation of the error $\boldsymbol{\varepsilon}$ is equal to 1 in the simulation.

As will be explained at the end of this section, additional sets of communities were simulated where the error $\boldsymbol{\varepsilon}$ was smaller, giving importance to species with lower absolute weights. Note that Equation 3.6 without the error $\boldsymbol{\varepsilon}$ 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

3A1). 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 to define 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 3.7.1 was used if ω was positive and Equation 3.7.2 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.

$$\mathbf{y}_{prob} = \frac{|\mathbf{y}_{ref}|}{\sum |\mathbf{y}_{ref}|} \tag{3.7.1}$$

$$\mathbf{y}_{prob} = \frac{1}{|\mathbf{y}_{ref}|} \times \frac{1}{\sum |\mathbf{y}_{ref}|} \tag{3.7.2}$$

Equation 3.7.1 defines the probability of sampling a species directly related to the patterns in $\mathbf{y}_{ref}$ whereas Equation 3.7.2 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 followed one of the predefined SADs presented in Figure 3.1. 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. To allocate these individuals to specific sites, we sampled, with replacement, the sites using the species probability distribution $\mathbf{y}_{prob}$.

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 3.1. Four other sets of 25 000 communities were also constructed where the error terms $\boldsymbol{\varepsilon}$ in Equation 3.6 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 ASSOCIATION COEFFICIENTS WITH EXPLAINED VARIANCE

The amount of explained variance in canonical ordinations was estimated with the coefficient of determination (R^2) according to the procedure proposed by Legendre and Gallagher (2001). Coefficients of determination are 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).

In the present study, only the canonical eigenvalues associated with significant canonical axes ($P \leq 0.05$ after 999 random permutations) were considered in calculation of R^2. Figure 3.2 compares the performance of RDAs for different association coefficients for each of the 25 SADs presented in Figure 3.1. In Figure 3.2, 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 3B1-3B4. All simulations yielded the same conclusions (see next paragraph), regardless of the error size. The only difference between the sets of simulations is that the larger the error when constructing species is associated to lower R^2. The inverse relation between error term and variance explained, which is consistant for all association 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 (above the 95%) 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 3.2, the most striking feature is that the confidence intervals for all asymmetrical association 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 association coefficient generally also has high R^2 for other coefficients.

The R^2 values for the Euclidean distance differ most from the other association coefficients, although its confidence intervals still overlap with the other coefficients (Figure 3.2, top panel). This is because the Euclidean distance is a symmetrical association coefficient. It is for the same reason that the confidence intervals are much wider for the Euclidean distance than for any other coefficients. 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 symmetrical association 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 3.1h-j, w-y). This is probably because variations in the most abundant species contribute predominantly in reducing the coefficient value when it is used (Legendre and Legendre 2012, Subsection 7.4.1). For this reason, the distance between species

profiles suffers from the same problem as Euclidean distance in the presence of common species, but to a lesser extent.

When comparing association coefficients with simulated presence-absence data, the R^2 are very similar between association coefficients across the different SADs. 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 (Figures 3B5-3B9). 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 3.1 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. 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 symmetrical coefficient that considers double-zeros. 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 will not be 0 even though it is clear that these sites are quite similar. For this reason, the Euclidean distance down-weights the

importance of abundant species, 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 3.1). 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 3B5-3B9). Similar conclusions were found with communities simulated with larger error (Figures 3B1-3B9).

A NEW WAY TO PERFORM CANONICAL ORDINATIONS

The previous simulations have shown that within data types, association coefficients yield similar value of R^2 independently for each SAD compared (Figure 3.2 and Figures 3B1-3B9). This is shown by the substantial overlap between confidence intervals of all association coefficients calculated for any particular SAD. Each association coefficient has particularities making it more appropriate for specific ecological situations or research questions, and less so for others. With the wealth of association coefficients available in the ecological literature, it is common for more than one coefficient to be appropriate for a particular ecological study. In that aspect, the question "Which association coefficient should be used?" remains unanswered.

Here, we propose a three-step procedure to handle this problem. Even though most of the information highlighted by the different association

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 instance, the first step is to compare association coefficients and evaluate how different the information they explain diverge. This is carried out 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 of association coefficients may be carried out if necessary. The RDA models constructed using association coefficients that differ markedly from the others should be considered separately or their usage 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 the species. In a nutshell, the differences between RDA models can be found in all aspect of the models. Comparison and selection of association coefficients is recommended because if an association coefficient is markedly different from the others, its inclusion in the following step may blur ecological relationships that could be apparent if this coefficient was removed. Thirdly, only the information common to RDA models that differ strickly by their association coefficients should be considered. 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 calculates a consensus among canonical ordinations that differ only by the association coefficients used to construct them. The consensus focuses on the patterns found by all RDA models, leaving out the

information emphasized by only one or a few association coefficients. We call this new approach "consensus RDA". A detailed explanation of how these three steps are carried out is presented bellow.

Comparison of RDA models.—To compare RDA models where only the association coefficients differ, the first step is to isolate the significant axes (P-value ≤ 0.05) found in each **Z** matrix (Equation 3.3.1). Model comparisons rely on the **Z** matrices, which contain the 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, canonical eigenvalues are measures of variance explained by canonical axes. The significant canonical axes of the **Z** matrix obtained with each association coefficient are correlated to those obtained with the other association coefficients 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 association coefficients are recorded in a matrix of pairwise RV coefficients. Using this matrix of pairwise RV coefficients, we propose to draw a minimum spanning tree (MST, Legendre and Legendre 2012, Section 8.2) to compare association coefficients. This requires the matrix of RV coefficients to be transformed into a dissimilarity matrix. We use (1 – RV) to perform the transformation because it ensures that the correlation information brought by the RV coefficient is conserved. This dissimilarity ranges from 0 to 1.

Although the procedure proposed compares **Z** matrices (sites scores), matrices **U** (species scores) and **C** matrices (canonical coefficients) should also be compared to ensure that all aspects of the models are considered.

Selection of RDA models.—After an examination of the MST a selection of association coefficient can then be made. We leave it at the discretion of the user to decide how association coefficients should be selected. For example, the association coefficients linked by the longest branches in the MST can be removed. If the longest branch in the MST links two groups of association coefficients, it may be interesting to calculate two consensus RDAs, one for each group of coefficients.

Consensus RDA.—To calculate a consensus RDA, the significant axes of the **Z** matrices selected to compare RDA models are used again (Figure 3.3a). Of course, only the **Z** matrices from coefficients that have been selected in the previous step should be considered. In consensus RDA, all significant axes are grouped in a large matrix (Figure 3.3b). A PCA is then performed on this large matrix to obtain $\mathbf{Z}^*$, the site-by-axes consensus RDA site scores (Figure 3.3c). This PCA also yields eigenvalues, which express the amount of variance represented by each $\mathbf{Z}^*$ axis, and more generally by each axis of the consensus RDA (Figure 3.3b). These eigenvalues can be used to measure the strength of the consensus.

Using $\mathbf{Z}^*$ as a reference, it becomes possible to compute a rotation matrix $\mathbf{H}_k$ for the $\mathbf{Z}_k$ matrix of significant axes associated to the k^{th} association coefficient (Figure 3.3c). This rotation matrix can be obtained through an orthogonal

Procrustes analysis (Peres-Neto and Jackson 2001; Gower and Dijksterhuis 2004, Chapter 4; Legendre and Legendre 2012, Section 10.5). In essence, the rotation matrix $\mathbf{H}_k$ pivots an RDA result calculated with one association coefficient to best fit the consensus RDA by a sum-of-squares minimization.

The rotation matrix $\mathbf{H}_k$ defines the orthogonal rotation needed to maximize correlation between $\mathbf{Z}^*$ and the $\mathbf{Z}_k$ matrix of significant axes. The rotation matrix $\mathbf{H}_k$ is obtained following Equation 3.8, where $\mathbf{V}_k$ is a matrix of eigenvectors and $\mathbf{D}_k$ a diagonal matrix of eigenvalues; both are extracted from $(\mathbf{Z}^{*t}\mathbf{Z}_k)^t(\mathbf{Z}^{*t}\mathbf{Z}_k)$. In this calculation, $\mathbf{Z}^*$ and $\mathbf{Z}_k$ are scaled by dividing each matrix by the square root of its sum-of-squares. When applied to any matrix whose values have been centred by columns, this scaling gives it a sum-of-squares of 1 (Gower and Dijksterhuis 2004), allowing $\mathbf{H}_k$ to be optimal.

$$\mathbf{H}_k = (\mathbf{Z}^{*t}\mathbf{Z}_k)\mathbf{V}_k\mathbf{D}_k^{-1/2}\mathbf{V}_k^t \tag{3.8}$$

Equation 3.8 applies only when $\mathbf{Z}^*$ has the same number of axes as $\mathbf{Z}_k$, a situation that does not always happen.

When the dimensions of $\mathbf{Z}^*$ and $\mathbf{Z}_k$ differ, a different procedure needs to be used to construct $\mathbf{H}_k$. Firstly, columns of zeros need to be added to $\mathbf{Z}_k$ to ensure that it has the same dimensions as $\mathbf{Z}^*$. With the zero-inflated $\mathbf{Z}_k$, $\mathbf{V}_k$ and $\mathbf{D}_k$ can be extracted from $(\mathbf{Z}^{*t}\mathbf{Z}_k)^t(\mathbf{Z}^{*t}\mathbf{Z}_k)$ as explained in the previous paragraph. The columns of zeros in $\mathbf{Z}_k$ are necessary for the matrix multiplication to be carried out. Note that only the eigenvectors and eigenvalues that correspond to the rank of $\mathbf{Z}_k$ are meaningful. The zero-inflated $\mathbf{Z}_k$, and the computed eigenvectors and

eigenvalues are used to calculate $(\mathbf{Z}^{*t}\mathbf{Z}_k)\mathbf{V}_k\mathbf{D}_k^{-1/2}$. Columns of random values orthogonal to each other and to all eigenvectors in $\mathbf{V}_k$ are then constructed to inflate the rank of the non-zero inflated $\mathbf{Z}_k$ for it to be equal to the rank of $\mathbf{Z}^*$. The resulting matrix is then multiplied with $\mathbf{V}_k^t$ (using all the eigenvectors it includes) to obtain $\mathbf{H}_k$.

With the rotation matrices, it becomes possible to compute the consensus of species scores and canonical coefficients (Equations 3.9 and 3.10, Figure 3.3d). Because the position of the site scores in an RDA is directly related to the species scores and the canonical coefficients, if site scores are rotated, the species scores and the canonical coefficients also need to be rotated for the information presented by an RDA to be consistent before and after rotation. The key to making this rotation is matrix $\mathbf{H}_k$. To construct the consensus RDA species scores $\mathbf{U}^*$, the matrices $\mathbf{U}_k$ are rotated with their respective $\mathbf{H}_k$ and averaged (Equation 3.9). The same procedure is carried out to obtain the consensus RDA canonical coefficients (Equation 3.10) and, if required, for the $\mathbf{F}_k$ matrices (Equation 3.11) (Figure 3.3d).

$$\mathbf{U}^* = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{U}_k\mathbf{H}_k) \tag{3.9}$$

$$\mathbf{C}^* = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{C}_k\mathbf{H}_k) \tag{3.10}$$

$$\mathbf{F}^* = \frac{1}{K}\sum_{k=1}^{K}(\mathbf{F}_k\mathbf{H}_k) \tag{3.11}$$

In equation 3.9 to 3.11, K is the number of association coefficients whose consensus is sought. Note that the $\mathbf{U}_k$, $\mathbf{C}_k$ and $\mathbf{F}_k$ matrices are all scaled by dividing each matrix by the square-root of their respective sum of squares to ensure that the consensus obtained by averaging is not influenced by the scale imposed by individual association coefficients.

When performing an RDA, the results can be presented either in 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 $\mathbf{Z}^*$ or $\mathbf{F}^*$) need to be rescaled following $\mathbf{Z}^*\mathbf{\Lambda}^{*-1/2}$ (or $\mathbf{F}^*\mathbf{\Lambda}^{*-1/2}$). A similar procedure is used to apply a scaling 2 on species scores consensus ($\mathbf{U}^*\mathbf{\Lambda}^{*-1/2}$).

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

The explanations to perform a consensus RDA were given so 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 if 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. The RV coefficient was used for the correlation. 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 association coefficients discussed in this paper were used in the construction of the consensus site scores.

The results in Figure 3.4 were obtained using abundance data where the error was the largest ($\boldsymbol{\varepsilon}$ in Equation 3.6 followed a Normal distribution with a mean = 0 and a standard deviation = 2), which yielded the largest variations in the comparison made. In Figure 3.4 (note the fine ordinate scale), 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 -0.05. 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 larger error terms are presented in Appendix 3C. In these simulations, presence-absence and abundance data were considered. For abundance data the results yield the same conclusions. For presence-absence data, it seems 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 association coefficients and a consensus RDA is performed 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 abundance is missing. The results of our simulations confirm this statement; the R^2 are consistently higher for abundance data (Figure 3.2, 3B1-3B4) than for presence-absence data (Figures 3B5-3B9). 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 level of error in the data (Figures 3.2, 3B1-3B9). 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 ordination results of abundance and presence-absence data, we first need to measure how much information from the true species (Equation 3.6 without the error term) structure is extracted by the canonical analyses. As explained at the end of section *RDA and association coefficients*, the Euclidean and simple matching coefficients are 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, symmetrical association 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 3.6 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 3.5). According to the results obtained by comparing association coefficients within data type (Figure 3.2 and Figures 3B1-3B9), it is valid to group association coefficients used on the same data type because no association coefficient dominates over the others for any SAD. Figure 3.5 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 3.1, 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 3.5, 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 by measuring occurrences should not be interpreted the same way in canonical ordinations as results obtained from canonical ordinations on abundance data. Similar results were obtained for data simulated with larger errors (Figures 3D1-3D4). 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 boreal mixedwood forest of northwestern Alberta, Canada (see Bergeron et al. 2011, Chapter 2 of this thesis). 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 only difference here is that we used all asymmetrical resemblance measures discussed in this paper and performed our analyses on abundance and presence-absence data. Bergeron et al. (2011) performed all their analyses using the percentage difference distance calculated on abundance data.

The sites, which covered an area of 70 km^2, were located in the Ecosystem Management Emulating Natural Disturbances (EMEND) experimental area. The community data are composed of 37 ground beetle species sampled with pitfall traps (Spence and Niemelä 1994) 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. Because the relative basal area of all trees sums to 1 for each site, only seven tree species may be drawn in the ordination triplot. Further analysis of this data set may be found in Chapter 2 of this thesis and in Bergeron et al. (2011, 2012). The Hellinger distance was used in Chapter 2 of this thesis and by Bergeron et al. (2012), and the percentage difference distance was employed by Bergeron et al. (2011). Note that Bergeron et al. (2012) used non-metric multidimentional scaling (Legendre and Legendre 2011, Section 9.4) to study carabids, unlike the Chapter 2 of this thesis and Bergeron et al. (2012) who used RDAs.

In this ecological illustration, we compare canonical ordinations calculated on abundance and presence-absence data, considering results from all association coefficients used in our simulations, with the exception of the symmetrical coefficients. We did not use 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. In Chapter 2 of this thesis it is also 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 symmetrical association coefficients inappropriate for studying this particular carabid community.

A comparison of the RDA models constructed with different association coefficients is presented using MSTs in Figure 3.6b for abundance data and in Figure 3.6e 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 association coefficients following the procedure presented in the section *A new way to perform canonical ordinations*. As a reference, we included in Table 3.3 the amount of variance explained (R^2) by the full RDA models with the different association 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 construct using the χ^2 distance is most different from the others (Figure 3.6b, e). This is probably because unlike the other association coefficients used, the χ^2 distance gives higher weights to species represented by only a few individuals (Legendre and Legendre 2012, Subsection 7.4.1). Because we did not want to give undue importance to rare species, we did not further consider the χ^2 distance in analysis of this carabid community.

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

the tree basal areas. The first two axes of the consensus RDA represent 88.4% of the variance for abundance data and 85.2% for presence-absence data, and thus represent well the information present in the different RDA models. All other consensus axes for abundance as well as presence-absence data described less than 7% of the variance, making the information they present too small to justify use of additional axes.

Although the amount of information explained by the first two axes of the consensus RDAs based on abundance (Figure 3.6c) and presence-absence data (Figure 3.6f) is similar, the underlying information is different. For example, the data about *Agonum gratiosum* (Agongrat), *Agonum sordens* (Agonsord), *Carabus chamissonis* (Caracham), *Nebria gyllenhali* (Nebrgyll), *Platynus mannerheimii* (Platmann), and *Pterostichus brevicornis* (Pterbrev), and *Trechus apicalis* (Trecapic) are impossible to interpret in the consensus RDA performed using species abundance because they were too close to the ordination centre. However, in the presence-absence ordination, information about these species is interpretable. Also, relationships between beetle and tree species were not always consistent between the two ordinations. For example, *Calathus ingratus* (Calaingr) and *Pterostichus adstrictus* (Pteradst) are more closely related to *Populus tremuloides* (Pt) in the abundance ordination (Figure 6c) than they are in the presence-absence ordination (Figure 6f).

The SAD (Figure 3.6a) 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 3.6c). Our simulations suggest that a community composed of many rare and many abundance species (Figure 3.1, l and t) does not preserve well community patterns after having been transformed into (Figure 3.5, l and t). 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. To know if the differences between the two consensus RDA is directed by ecological processes, a detailed study of this carabid community needs to be carried out contrasting presence-absence and abundance data at multiple scale using other variables characterizing the habitat of Carabidae in addition to tree basal area.

It is not the goal of this paper to present a detailed ecological study of northwestern Alberta boreal carabids. However, by comparing the consensus RDA calculated on the carabid abundance data (Figure 3.6c) with the ordination results from Bergeron et al. (2012, Figure 4) who also studied the relationship between carabids and tree relative basal area with the same data using RDA, differences can be found that are solely attributed to the association coefficient used. For example, in our result, *Stereocerus haematopus* (Sterhaem) is more closely related to *Pinus contorta* (Pc) than it is in the analyses of Bergeron et al.

(2012). To prevent a bias interpretation that stems from the use of a specific association coefficient, as was the case for *S. haematopus*, consensus RDA is a better option.

DISCUSSION

A surprising result of this study is that the SAD of a community is unimportant for choosing an association coefficient (Figure 2 and 5, Figure 3B1-3B9 and 3D1-3D4) when used with canonical ordinations. This is what prompted us to developed consensus RDA. These results may also bring insight in the comparison of SADs, an important line of research (McGill et al. 2007). Using the result in Figure 3.5 (and Figure 3D1-3D4) obtained from abundance data, we can compare SADs because the communities simulated with different SADs were correlated with the same true underlying structure of the data (Equation 3.6 without the error term). The true underlying structure of the data serves as a reference to know how well a SAD defines the raw community data because it is the basic information from which all species are constructed. From the discussion in McGill et al. (2007) on SAD comparison, it can be expected that SADs defining abundance patterns as different as the ones in Figure 3.1b, l, m, q, and u would correlate differently with the true underlying structure of the data. However, in Figure 3.5 they all correlate equally well with the true underlying structure of the simulated communities. Moreover, the fairly broad range of the 95% confidence intervals for any one of the 25 SADs indicates that the variations in the raw multivariate community data can be surprisingly important even if species have the same abundance structure. Such results may suggest that the

SAD of a community may present only a small fraction of the information that characterizes a community matrix. However, further research still need to be carried out to confirm the findings we made that the information lost when constructing SADs may make it difficult to develop a valuable approach to compare communities using SADs.

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

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

model(s) differs notably from the others. This comparison can be used to decide if any association coefficients should be discarded. Comparing RDA models constructed using different association coefficients through an MST is a first step to better understand an ecological community by studying the ordination results produced by more than one association coefficient.

When more than one association coefficient presents similar information, a consensus RDA allows to extract the most information out of the data because it focuses on the common information brought out by different association coefficients. Using only one coefficient may put too much emphasis on a particular aspect of the data because each association 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 association coefficients. In that instance, consensus RDA indirectly solves the technical problem of choosing an association coefficient by using all the ones that can be suitable to analyse community data. Also, because it diminishes the importance of the information highlighted by one or a few association coefficients it gives a result less influenced by the mathematical properties of an association 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 weights than the poor ones. This can be related to the selection procedure we propose where association coefficients are considered independently, discarded, or used to construct a consensus model. However, the association coefficients used in consensus RDA are all weighted equally. If a model is given more weight than another, it would mean that a particular association coefficient should be favoured. In that instance, why use a "weighted" consensus of RDA models if one association coefficient ought to be favoured? It would be simpler and better to build an RDA model using the association coefficient that is best adapted to answer the ecological problem at hand.

A problem that we have not approached but warrants further investigation is selection of explanatory variables in RDA. Methods such as forward selection (e.g., Blanchet et al. 2008) assume that an RDA is performed using only one association coefficient. Consensus RDA requires all explanatory variables to be the same and that only the association 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 that a consensus analysis should employ the union of all explanatory variables selected for the various association coefficients. That is, if for an association coefficient, explanatory variables A and B are selected and with another association 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.

Species abundance data are more informative than presence-absence data in understanding community variations through RDA. However, for certain organisms sampling abundances is not reliable. For example, in palynology presence-absence data are sometimes favoured because abundance data are subject to large bias (Davis 2000). 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 data. This is likely to be true for any communities where variations in size between species require that different trapping methods be used to catch enough species to have a representative fraction of the studied species community.

When working with presence-absence data, we suggest that one should first draw a species presence distribution, as we did in Figure 3.6c. The ratio between common and rare species should serve as a general guideline when devising the ecological conclusions. Although it is possible that canonical ordinations performed on presence-absence data present biased results, it is more likely that such ordinations can be complementary to those performed on abundance data. Certain environmental factors may be necessary for a species to occur in an area (e.g., certain plant species are found only in the presence of

certain geological formations) while other factors may make species abundance vary (e.g., precipitation). Variation in abundance is efficient in describing how a species is related to a gradient (environmental, physical, or others). However species abundances may conceal the strict relationship a species has with its habitat defining if that species occurs or not at a site. The nature of presence-absence data may be more efficient in capturing the strict relationship a species may have with its habitat. In that instance, considering both abundance and presence-absence data may be ecologically valuable to better understand the factors structuring a community. As explained in the previous paragraphs, abundance data may be unreliable when sampling certain group of organisms. However, for all communities where species abundances can be sampled without diminishing the value of the data, presence-absence data can be easily obtained by transforming all abundances larger than 0 to the value 1, allowing ecologists to get a more complete understanding of the data they gathered.

In conclusion, besides presenting a new approach to perform canonical ordination using a group of association coefficients, in this paper we also propose a new framework to analyze species communities using abundance and presence-absence data together.

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.

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

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

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

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

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

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

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

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

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

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

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

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

Gaston, K. J. 2011. Common Ecology. BioScience **61**:354-362.

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

Gower, J. C., and G. B. Dijksterhuis. 2004. Procrustes problems. Oxford University Press.

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

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

Hirschfeld, H. O., 1935. A connection between correlation and contingency. Pages 520–524 *in* Proceedings of the Cambridge Philosophical Society, volume 31. Cambridge University Press.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Sneath, P. H. A., and R. R. Sokal. 1973. Numerical Taxonomy. W. H. Freeman and company.

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

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

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

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. F. M. Verdonschot. 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology. Aquatic Sciences-Research Across Boundaries **57**:255–289.

Whittaker, R. H. 1972. Evolution and Measurement of Species Diversity. Taxon **21**:213–251.

TABLE 3.1. List of association coefficients compared. All coefficients are presented in a dissimilarity (distance) format. The association coefficients in bold can be applied to presence-absence as well as abundance data directly.

Association coefficient	Equation	Reference	Comment
Binary symmetrical			
Simple-matching	$\sqrt{1 - \frac{a+d}{a+b+c+d}}$ [a]	Sokal and Michener (1958)	Binary equivalent of Euclidean (Sneath and Sokal 1973)
Binary probabilistic			
Raup-Crick	$1 - p(a_{hi})$ [ab]	Raup and Crick (1979) McCoy et al. (1986)	
Binary asymmetrical			
Jaccard	$\sqrt{1 - \frac{a}{a+b+c}}$ [a]	Jaccard (1901)	Binary equivalent of any variation of the modified Gower dissimilarity
Sørensen	$\sqrt{1 - \frac{2a}{2a+b+c}}$ [a]	Sørensen (1948)	Binary equivalent of percentage difference
Ochiai	$\sqrt{1 - \frac{a}{(a+b)+(a+c)}}$ [a]	Ochiai (1957)	Binary equivalent of chord and Hellinger
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)	
Hellinger	$\sqrt{\sum_{j=1}^{p}\left(\sqrt{\frac{y_{1j}}{y_{1+}}} - \sqrt{\frac{y_{2j}}{y_{2+}}}\right)^2}$ [c]	Rao (1995)	
χ^2	$\sqrt{y_{++}}\sqrt{\sum_{j=1}^{p}\frac{1}{y_{+j}}\left(\frac{y_{1j}}{y_{1+}} - \frac{y_{2j}}{y_{2+}}\right)^2}$ [c]	Lebart and Fénelon (1971)	Dissimilarity preserved in CCA

TABLE 3.1. Continue

Abundance asymmetrical			
Distance between species profiles	$\sqrt{\sum_{j=1}^{p}\left(\frac{y_{1j}}{y_{1+}}-\frac{y_{2j}}{y_{2+}}\right)^2}$ [c]	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[e]
$\sqrt{\text{Percentage difference}}$	$\frac{\sum_{j=1}^{p}\lvert\sqrt{y_{1j}}-\sqrt{y_{2j}}\rvert}{\sum_{j=1}^{p}(\sqrt{y_{1j}}+\sqrt{y_{2j}})}$	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 $\log_2$	$\frac{\sum_{j=1}^{p} w_j\lvert\log_2(y_{1j})-\log_2(y_{1j})\rvert}{\sum_{j=1}^{p} w_j}$ [d]	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 $\log_5$	$\frac{\sum_{j=1}^{p} w_j\lvert\log_5(y_{1j})-\log_5(y_{1j})\rvert}{\sum_{j=1}^{p} w_j}$ [d]	Anderson et al. (2006)	
Modified Gower $\log_{10}$	$\frac{\sum_{j=1}^{p} w_j\lvert\log_{10}(y_{1j})-\log_{10}(y_{1j})\rvert}{\sum_{j=1}^{p} w_j}$ [d]	Anderson et al. (2006)	

[a] The letters a, b, c, and d are defined in Table 3.2.

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

[c] 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*.

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

[e] 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 3.2. Contingency table describing the similarity between two sites where species presence or absence were sampled. *a* is the number of species present at site 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*, or *d* from a community matrix **Y** composed of *p* species, where y_{1j} and y_{2j} present the species occurrence of site 1 and 2 for species *j*.

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

TABLE 3.3. Variance explained (R^2) by RDA models constructed independently with each association 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 abundance of carabids divided by the number of days 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 symmetrical association coefficients. All association coefficients are defined in Table 3.1.

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

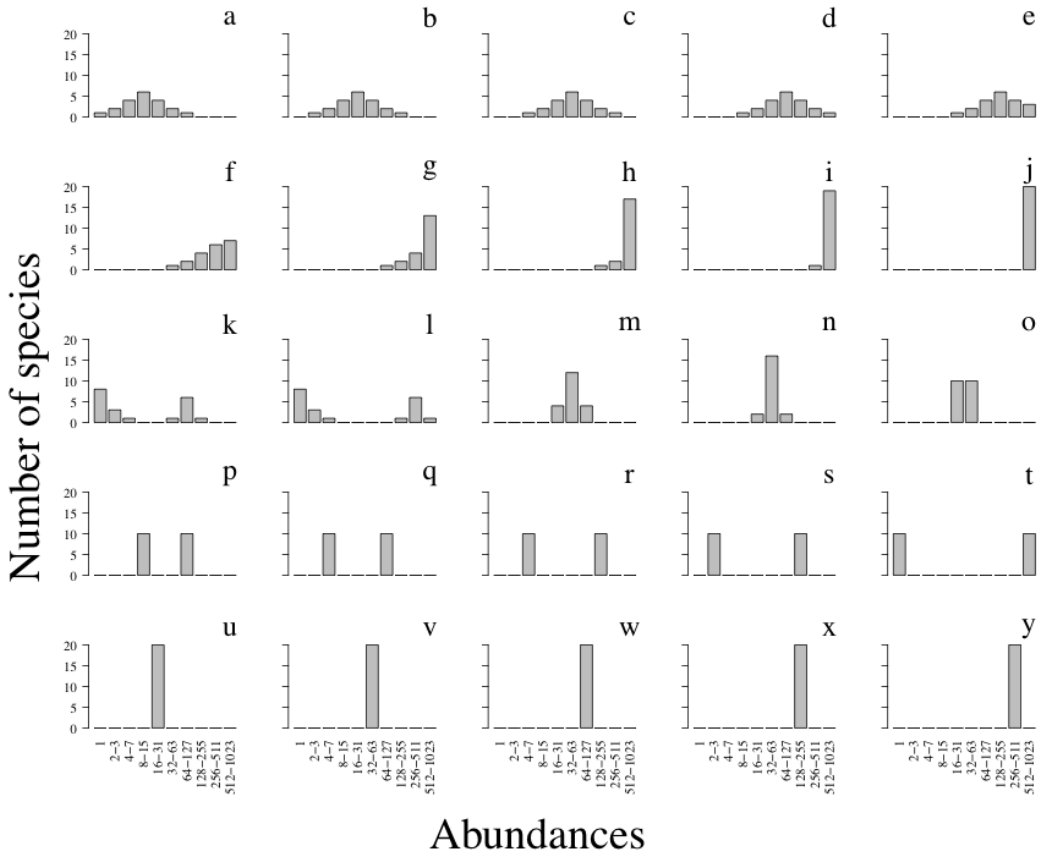

FIGURE 3.1. Species-abundance distributions (SAD) used in the simulations. These SADs are presented using Preston (1948) graphs where the abundance classes in the abscissa increase according to a geometric progression whose lower bound is made of the values 2^k with k being the successive integers from 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 presents a community of 20 species. They were constructed to encompass a wide range of variations in abundance patterns.

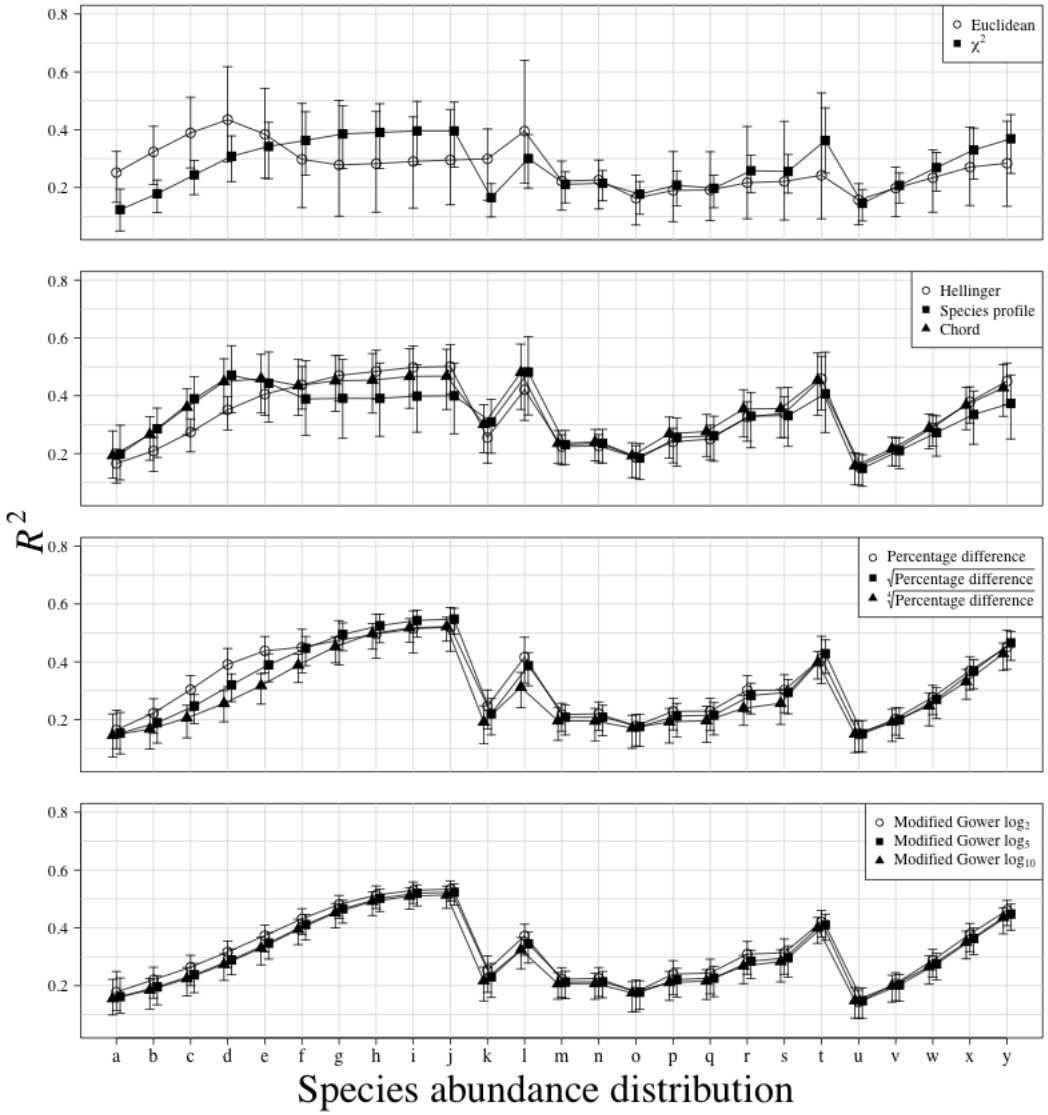

FIGURE 3.2. Comparison of explained variance (R^2) between 11 association coefficients calculated from simulated communities following different species abundance distributions (SAD) using abundance data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved to calculate R^2. Points are R^2 means of all simulations and error bars represent 95% confidence intervals. Association coefficients are presented in different panels for visual clarity. Letters along the abscissa refer to the SADs as presented in Figure 3.1. 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 = 0.001). A thousand simulations were run for each SAD.

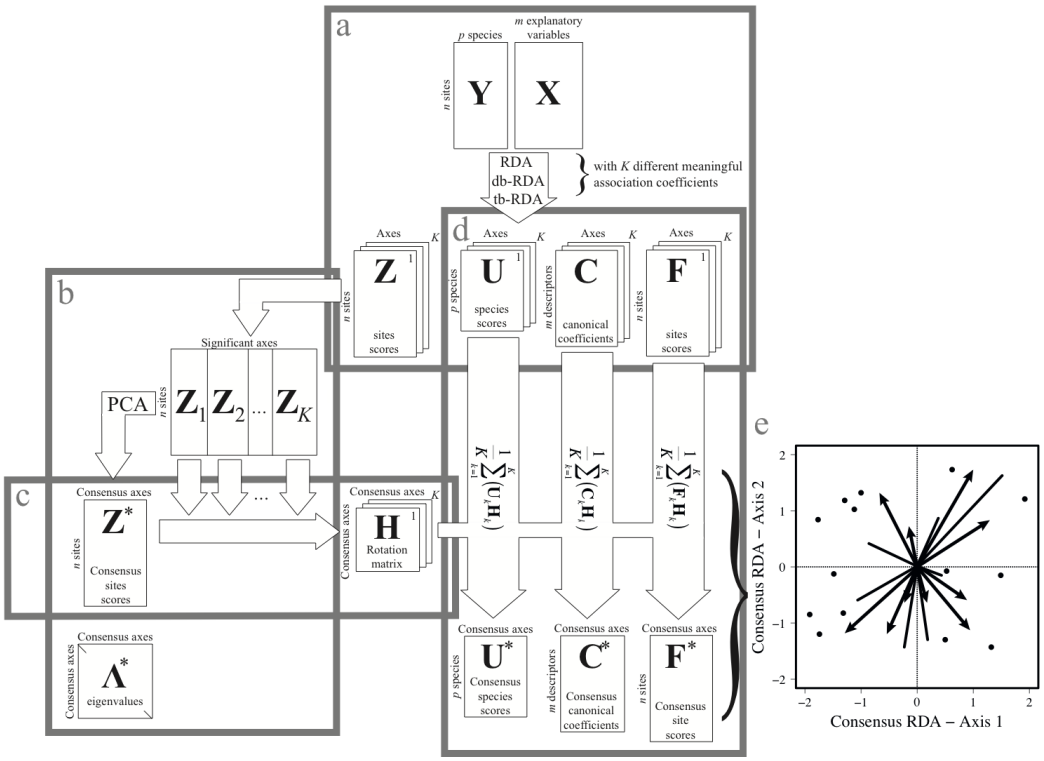

FIGURE 3.3. Schematic representation of consensus RDA. (a) The first step of the procedure is to perform a series of RDAs (tb-RDA or db-RDA) to model the community data **Y** using explanatory variables **X**. Each RDA is performed using a distance scaling (scaling 1) but with a different association coefficient. In the figure, *K* different association coefficients are used. An RDA result is composed of four principal matrices: the species scores **U**, the sites scores **Z** calculated in the space of **X**, the site scores **F** calculated in the space of **Y** and the canonical coefficients **C**. Because *K* different meaningful association coefficients are used and thus *K* RDAs are performed, *K* sets of matrices **U**, **Z**, **F**, and **C** are calculated. (b) For each of the *K* association coefficients, the significant axes within each **Z** matrix are grouped in a large matrix. A PCA is then performed on this large matrix yielding the site scores consensus matrix $\mathbf{Z}^*$ and a diagonal matrix of eigenvalues $\mathbf{\Lambda}^*$. (c) Using $\mathbf{Z}^*$ as a reference, a matrix of orthogonal rotation **H** is calculated for each **Z** matrix. The construction of **H** matrices are carried out using the scaled $\mathbf{Z}^*$ and $\mathbf{Z}_k$. By that we mean that $\mathbf{Z}^*$ and $\mathbf{Z}_k$ were divided by their sum of squares before being square root transformed. (d) The consensus species scores $\mathbf{U}^*$ can then be computed by multiplying each $\mathbf{U}_k$ by its respective orthogonal rotation matrix and averaging all the rotated $\mathbf{U}_k$ matrices. The same calculation is performed to obtain the consensus canonical coefficients $\mathbf{C}^*$ and if necessary $\mathbf{F}^*$. For $\mathbf{U}^*$, $\mathbf{C}^*$ and $\mathbf{F}^*$ to be optimal, they need to be calculated from the independent matrices with a sum of squares equal to 1. (e) $\mathbf{Z}^*$, $\mathbf{U}^*$, $\mathbf{C}^*$ and $\mathbf{F}^*$ can than be used to draw a consensus RDA triplot. The eigenvalues in $\mathbf{\Lambda}^*$ can also be used in the consensus RDA triplot to show the importance of each axis.

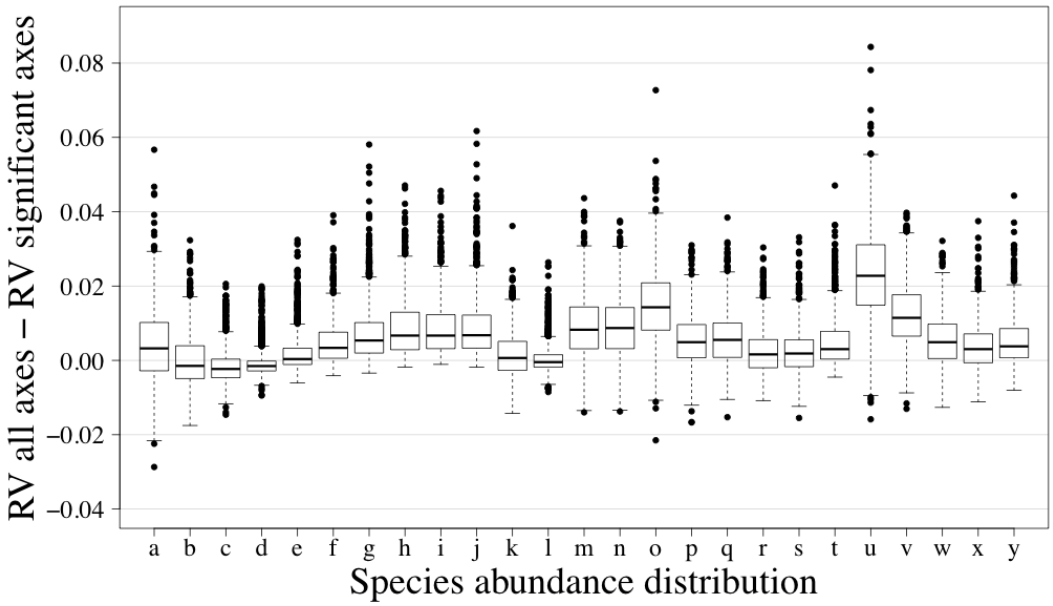

FIGURE 3.4. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from abundance data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (SAD) 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 = 2). A thousand simulations were run for each SAD.

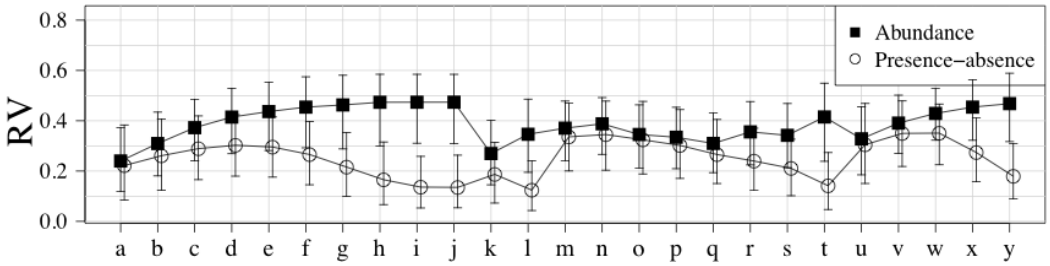

FIGURE 3.5. Comparison between abundance and presence-absence data to know how much of the true species structure (Equation 3.6 without the error term) is modelled by the canonical ordination models. For each data type (abundance and presence-absence), the significant canonical axes for all association coefficients (with the exception of the 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) as presented in Figure 3.1. A line was drawn between each SAD of each association coefficient to ease comparisons 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.

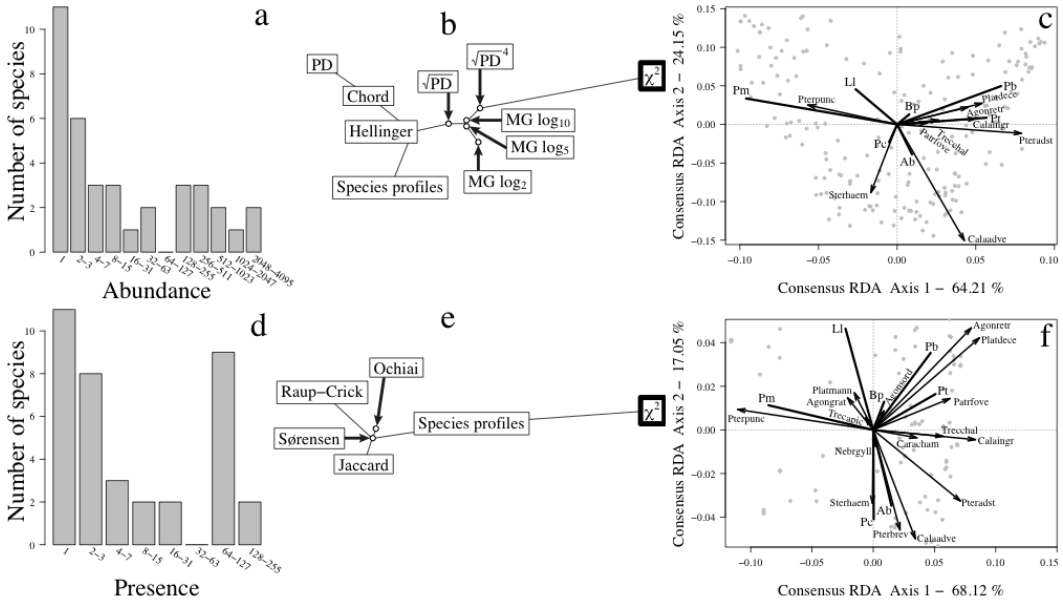

FIGURE 3.6. Comparison of (a) species-abundance distributions (SAD) and (d) species-presence distributions (SPD), and consensus RDA results for abundance (c) and presence-absence data (f) using Carabidae data sampled at the *Ecosystem Management Emulating Natural Disturbances* (EMEND) experimental area in Alberta, Canada. (b) The minimum spanning trees (MST) comparing association coefficients for abundance data and (e) the MST comparing association coefficient for presence-absence data show that the χ^2 distance presents an RDA very different from the other association coefficients. For both data types the χ^2 distance was the only association coefficient not used to compute the consensus RDA in (c) and (f). The SAD and SPD are constructed the same way, with the exception that for SPD it is the occurrence of species that is considered, not their abundance. The SAD (a) and SPD (d) are used as reference to relate the results presented in this figure to the simulation results presented in Figures 3.2 and 3.5. The consensus RDA triplots describe the relationship between ground beetle species (arrows), the relative basal area of trees by species (lines), and the sampling sites (grey points) using all but the symmetrical association coefficients and the χ^2 distance. The species codes for the Carabidae and trees are provided in Tables 3E1-3E2. In (b) and (e) MG stands for modified Gower and PD for percentage difference, the name of all other association coefficients are written fully.

APPENDIX 3A

Explanatory variables used in the construction of species during the simulations

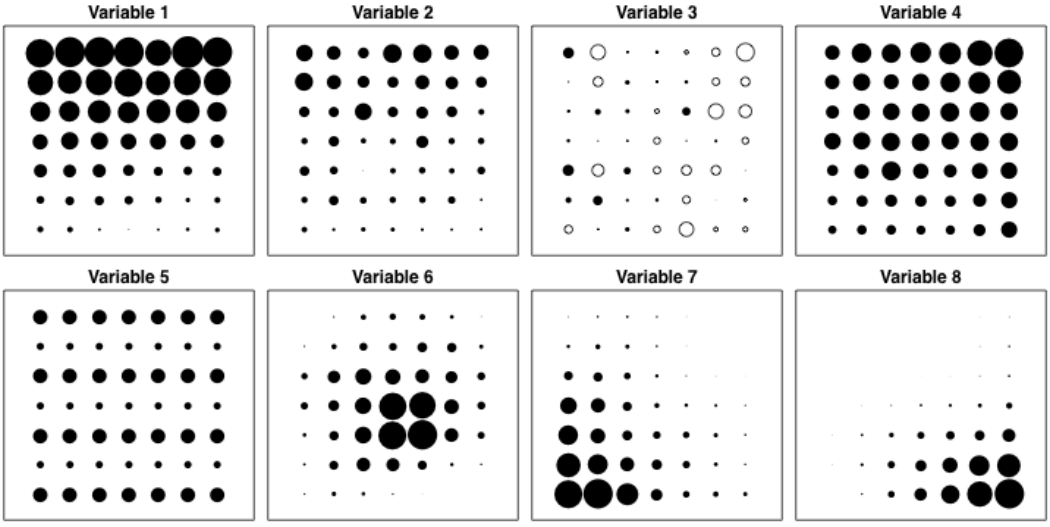

FIGURE 3A1. Bubble plots presenting the eight variables used in the construction of the simulated species through Equation 3.6. Each variable was constructed on a 7 × 7 regular grid. Each individual bubble is associated to a simulated site. The size and colour of the bubbles characterize the value associated to the bubble (black are positive values, white are negative, and bubble size is related to the associated value). The absence of bubble represents a value of 0. All variables have a range of 10 except for variable 5, which has a range of 2.5.

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 3A1: Explanatory variables and weight (regression coefficient) used to construct each species (following Equation 3.6) in the simulated communities. 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 3B

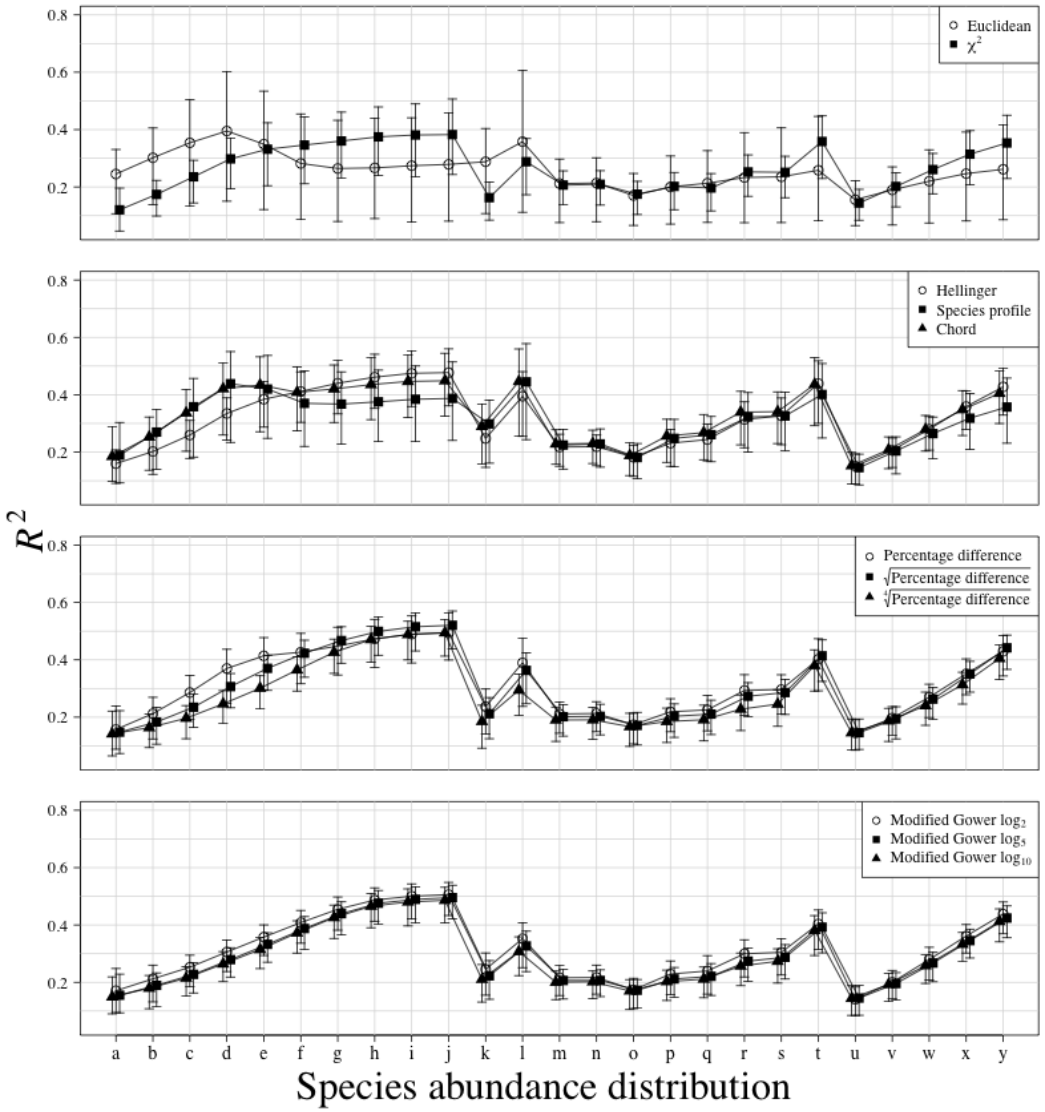

FIGURE 3B1. Comparison of explained variance (R^2) between 11 association coefficients calculated from simulated communities following different species abundance distributions (SAD) using abundance data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved to calculate R^2. Points are R^2 means of all simulations and error bars represent 95% confidence intervals. Association coefficients are presented in different panels for visual clarity. Letters along the abscissa refer to the SADs as presented in Figure 3.1. 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.

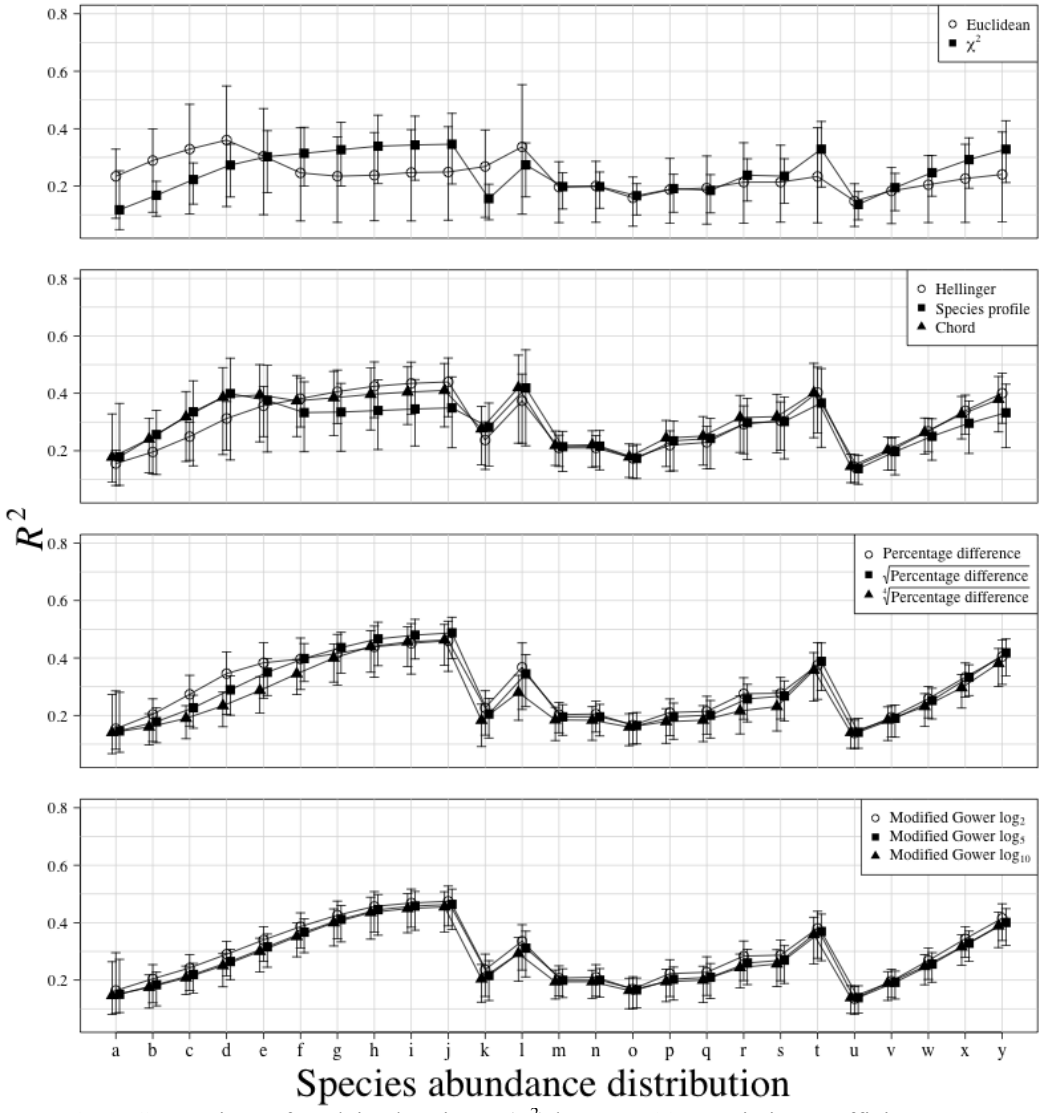

FIGURE 3B2. Comparison of explained variance (R^2) between 11 association coefficients calculated from simulated communities following different species abundance distributions (SAD) using abundance data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved to calculate R^2. Points are R^2 means of all simulations and error bars represent 95% confidence intervals. Association coefficients are presented in different panels for visual clarity. Letters along the abscissa refer to the SADs as presented in Figure 3.1. 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.

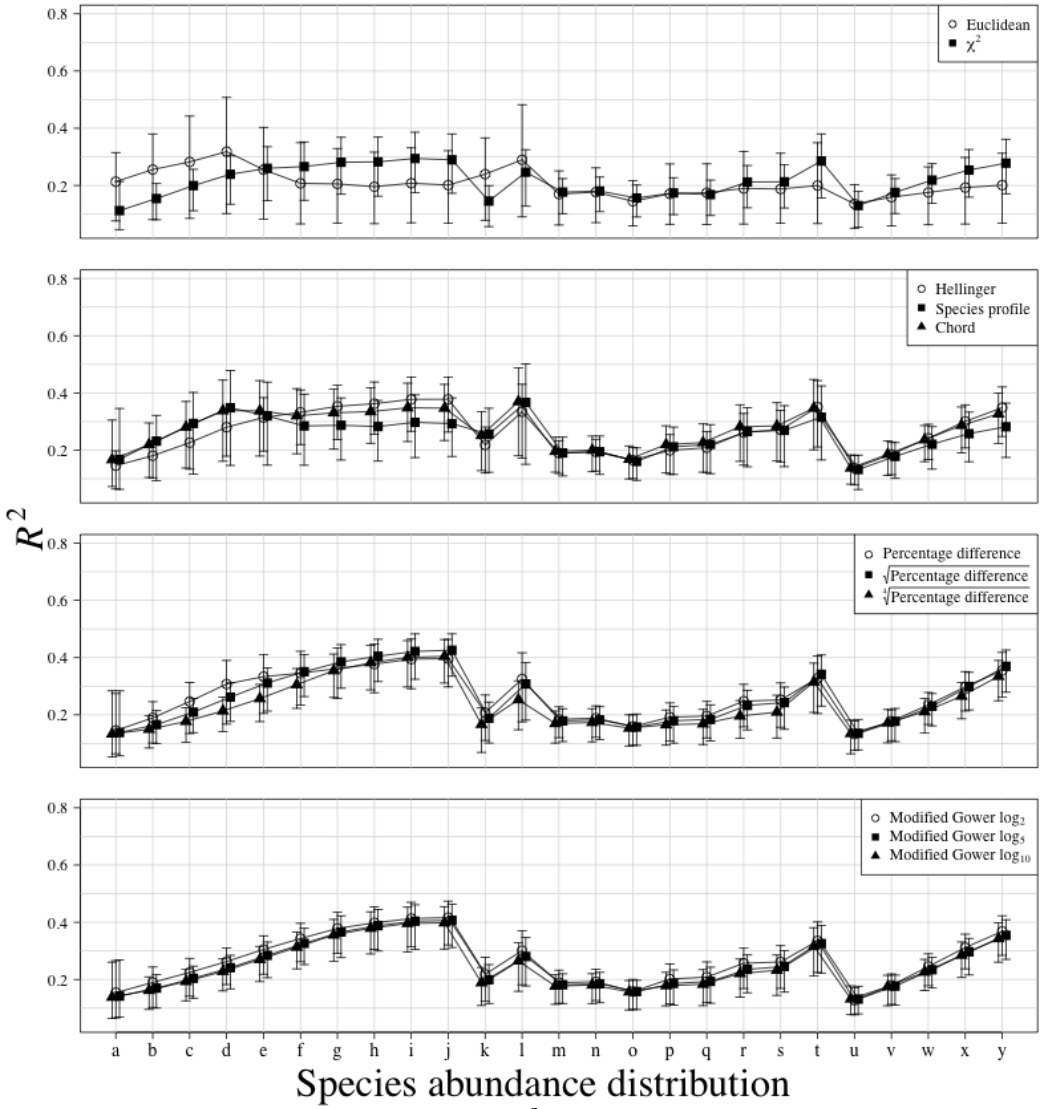

FIGURE 3B3. Comparison of explained variance (R^2) between 11 association coefficients calculated from simulated communities following different species abundance distributions (SAD) using abundance data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved to calculate R^2. Points are R^2 means of all simulations and error bars represent 95% confidence intervals. Association coefficients are presented in different panels for visual clarity. Letters along the abscissa refer to the SADs as presented in Figure 3.1. 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.

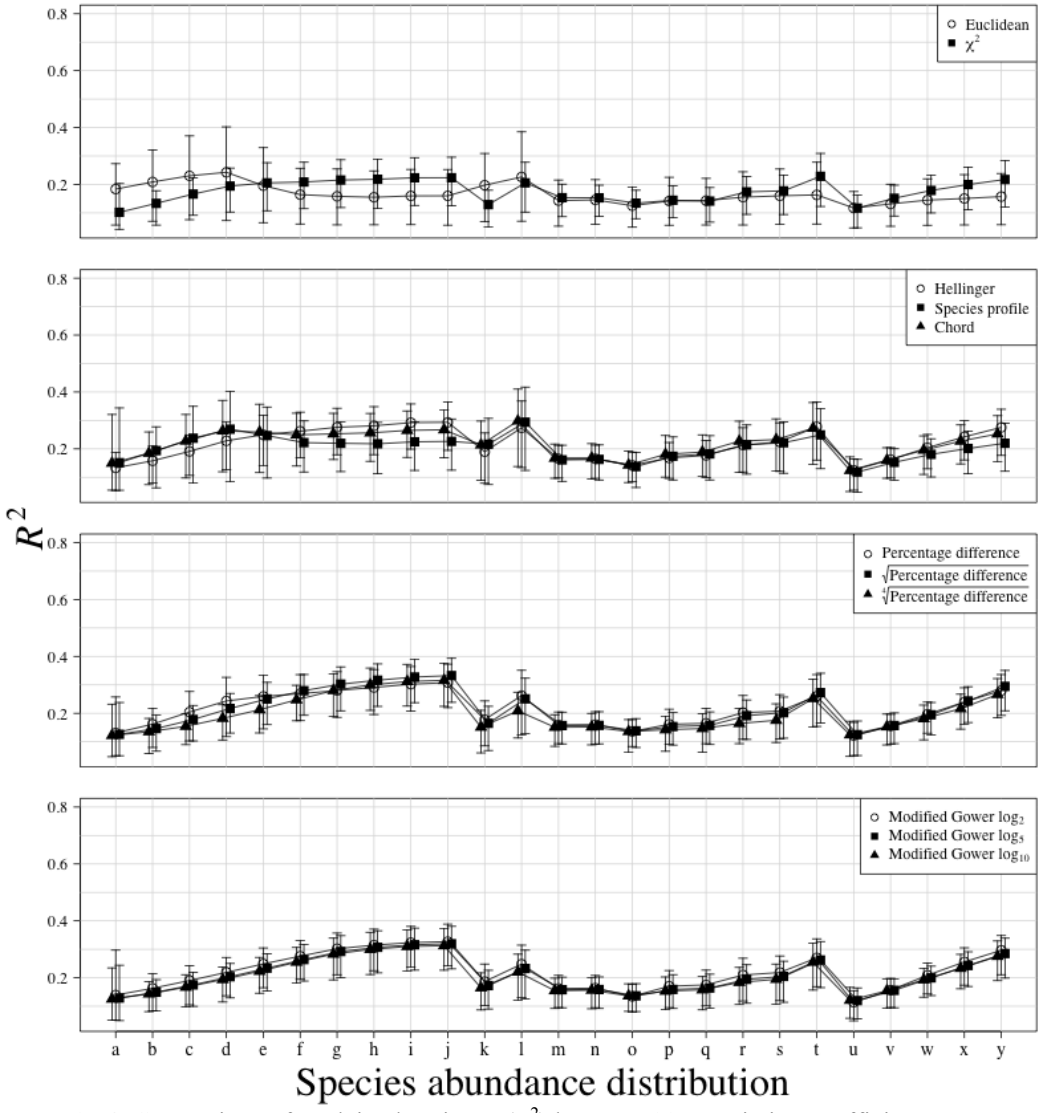

FIGURE 3B4. Comparison of explained variance (R^2) between 11 association coefficients calculated from simulated communities following different species abundance distributions (SAD) using abundance data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved to calculate R^2. Points are R^2 means of all simulations and error bars represent 95% confidence intervals. Association coefficients are presented in different panels for visual clarity. Letters along the abscissa refer to the SADs as presented in Figure 3.1. 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.

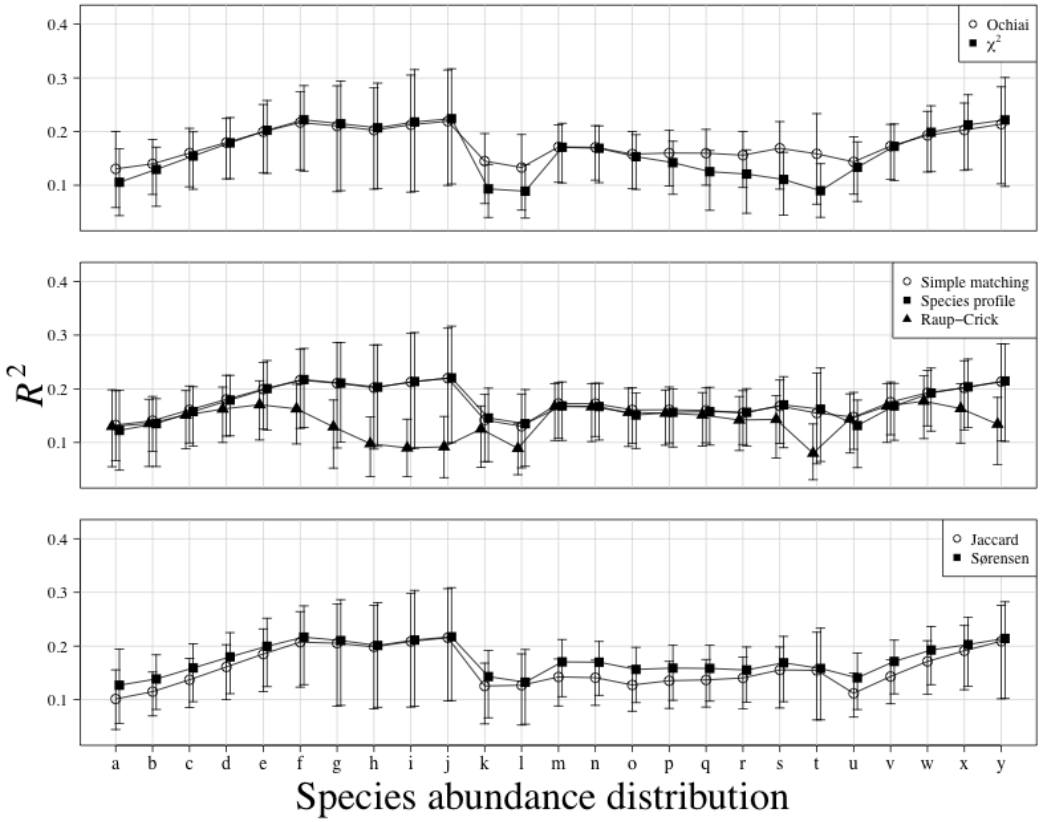

FIGURE 3B5. Comparison of explained variance (R^2) between 7 association coefficients calculated from simulated communities following different species abundance distributions (SAD) using presence-absence data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved 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 along the abscissa refer to the SADs as presented in Figure 3.1. 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 = 0.001). A thousand simulations were run for each SAD.

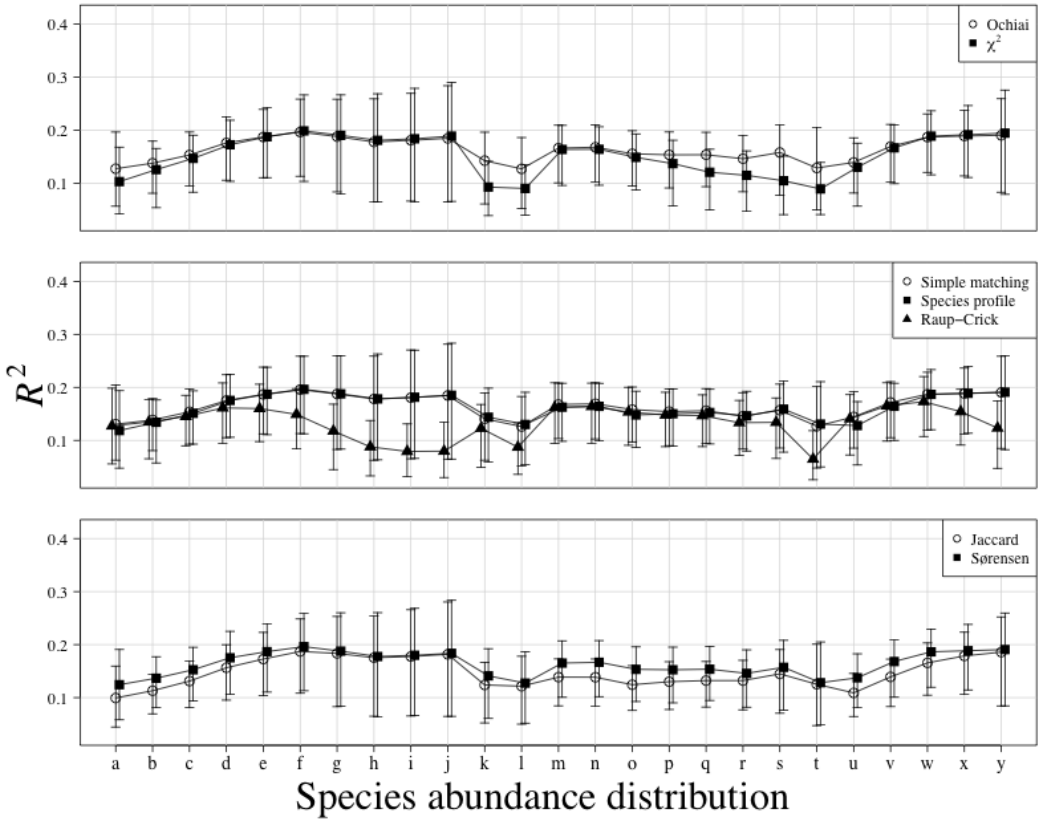

FIGURE 3B6. Comparison of explained variance (R^2) between 7 association coefficients calculated from simulated communities following different species abundance distributions (SAD) using presence-absence data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved 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 along the abscissa refer to the SADs as presented in Figure 3.1. 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.

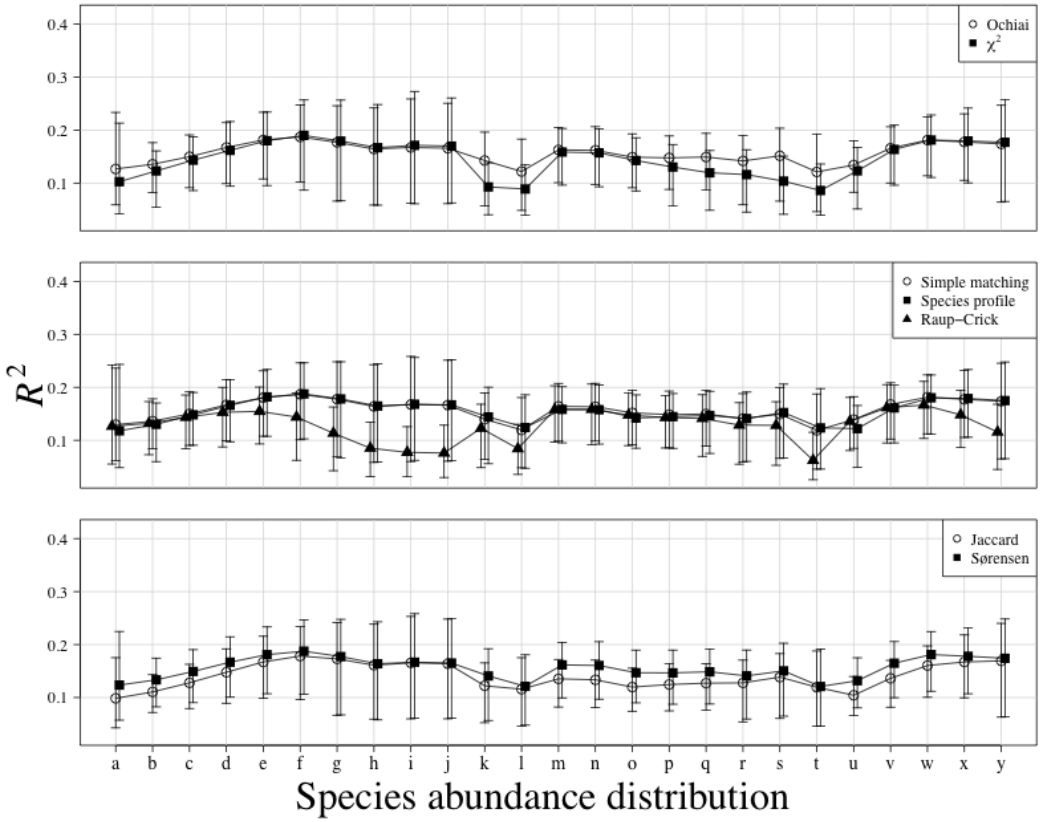

FIGURE 3B7. Comparison of explained variance (R^2) between 7 association coefficients calculated from simulated communities following different species abundance distributions (SAD) using presence-absence data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved 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 along the abscissa refer to the SADs as presented in Figure 3.1. 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.

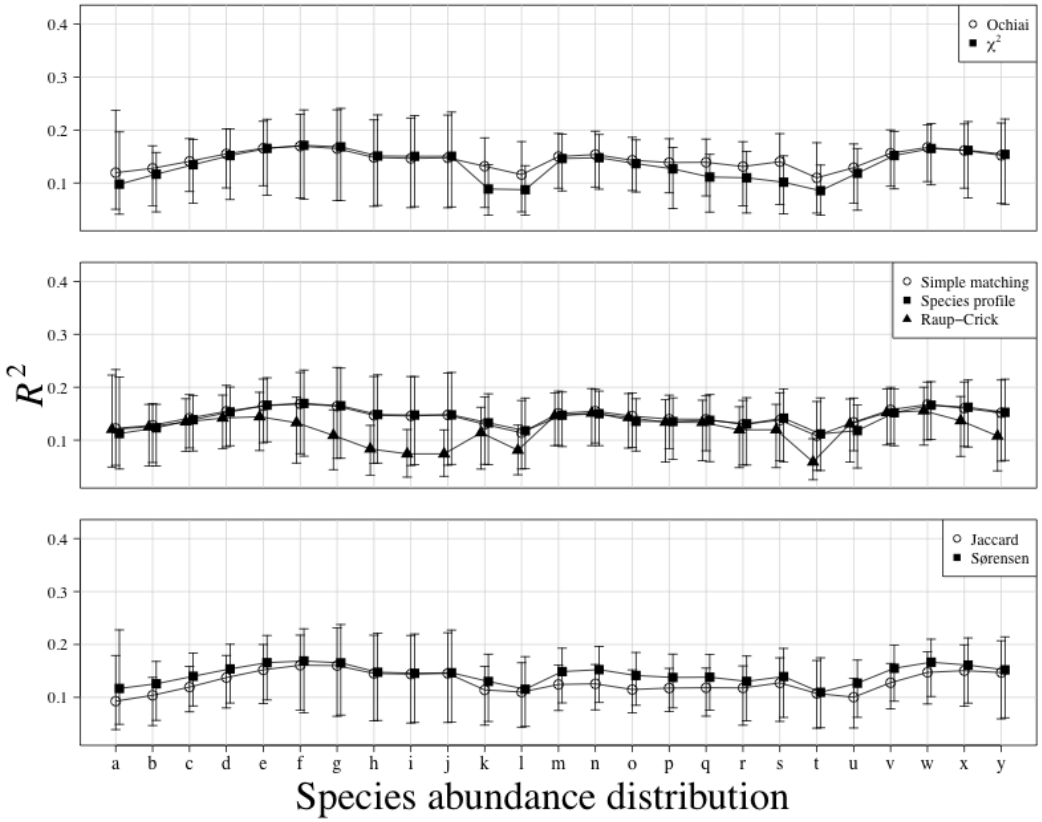

FIGURE 3B8. Comparison of explained variance (R^2) between 7 association coefficients calculated from simulated communities following different species abundance distributions (SAD) using presence-absence data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved 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 along the abscissa refer to the SADs as presented in Figure 3.1. 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.

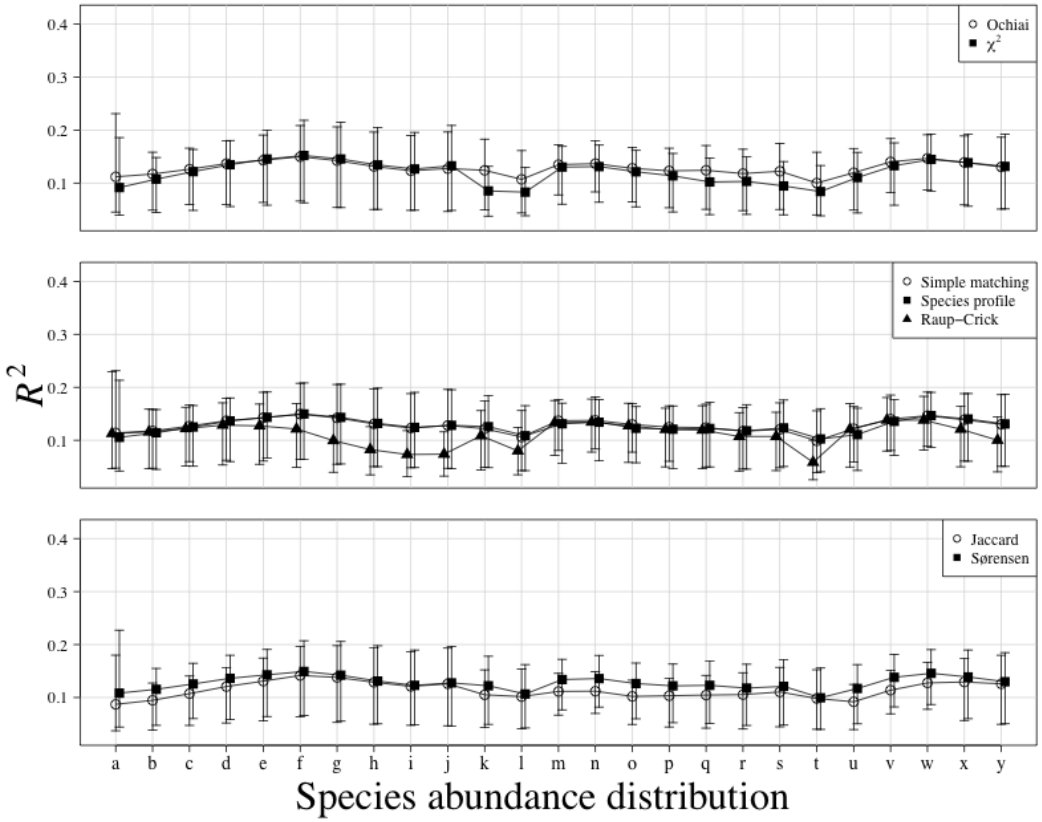

FIGURE 3B9. Comparison of explained variance (R^2) between 7 association coefficients calculated from simulated communities following different species abundance distributions (SAD) using presence-absence data. Only the significant ($P \leq 0.05$ after 999 permutations) canonical axes were conserved 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 along the abscissa refer to the SADs as presented in Figure 3.1. 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 3C

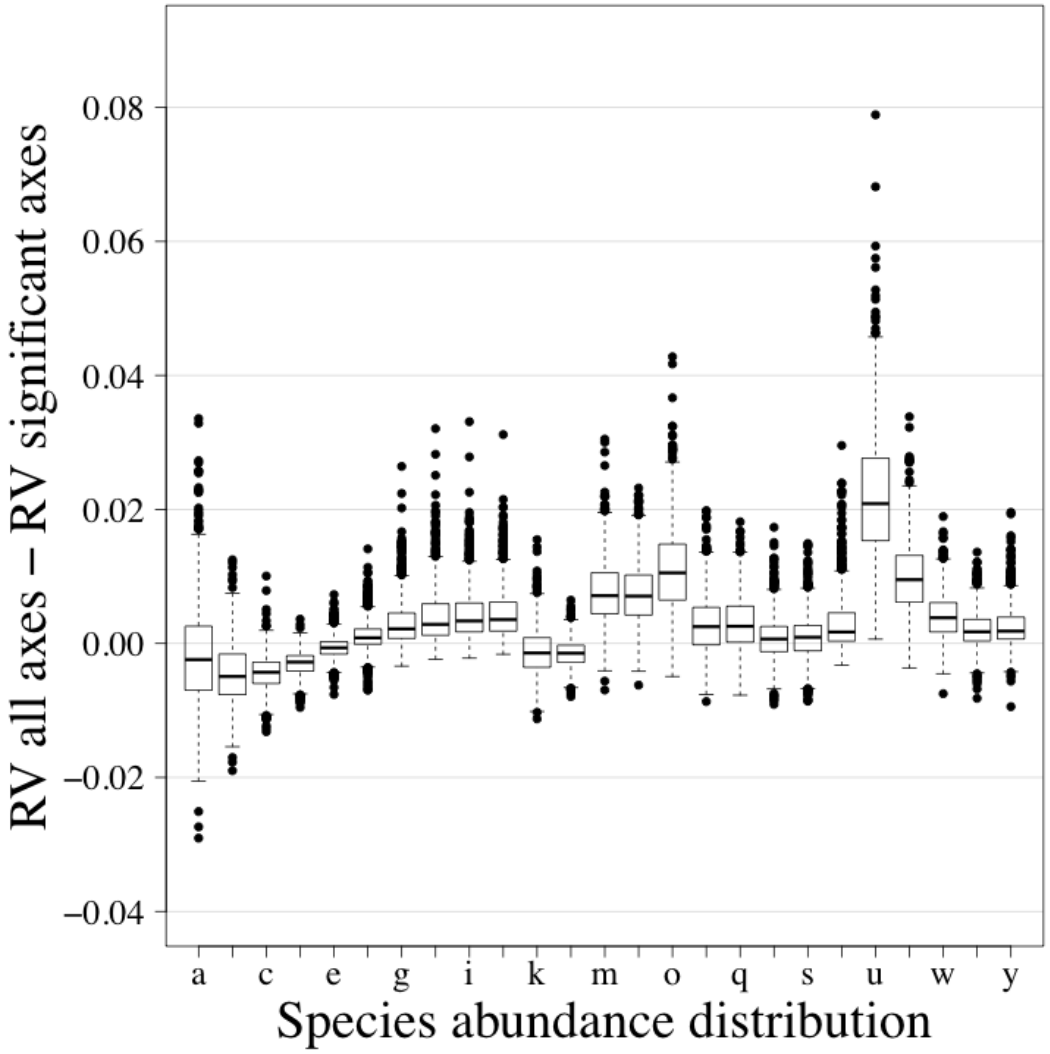

FIGURE 3C1. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from abundance data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (SAD) 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.

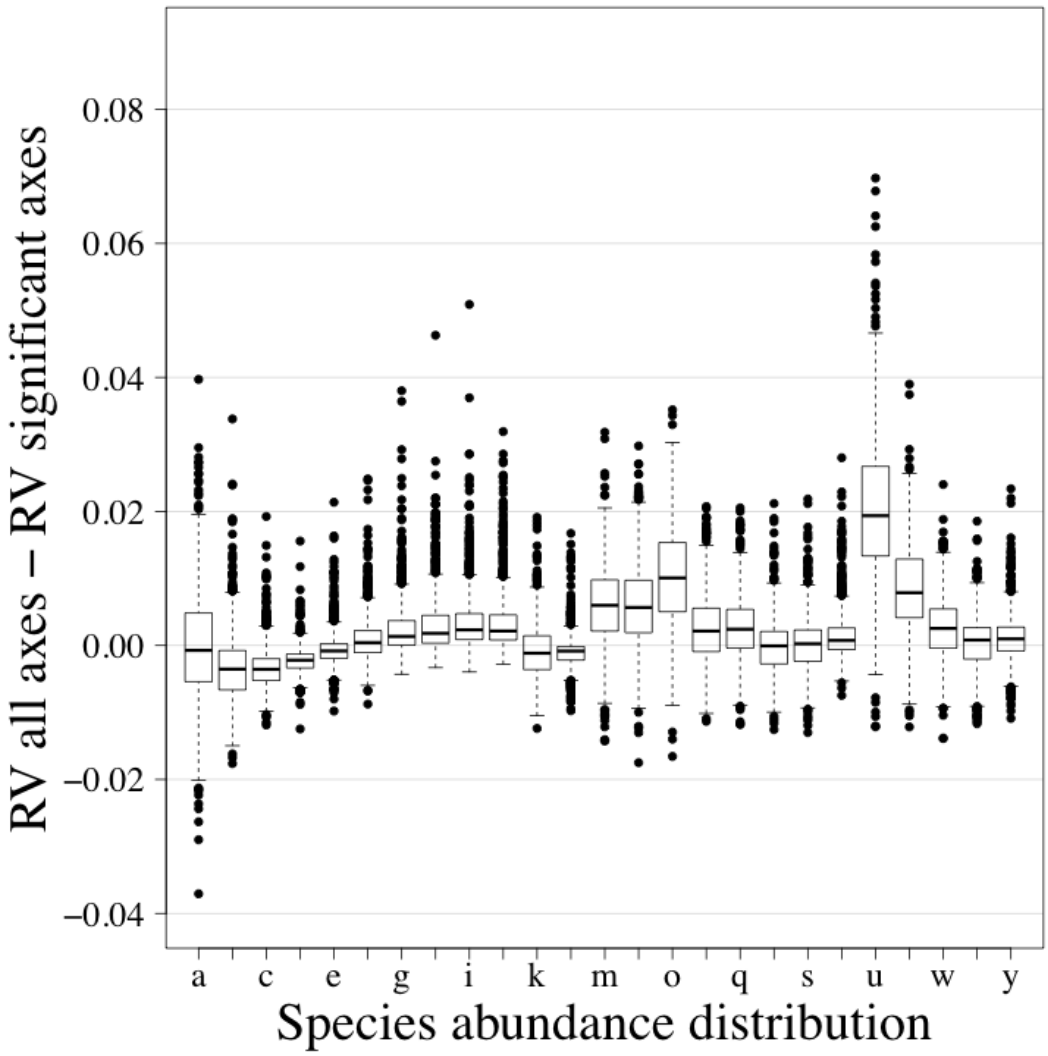

FIGURE 3C2. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from abundance data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (SAD) 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.

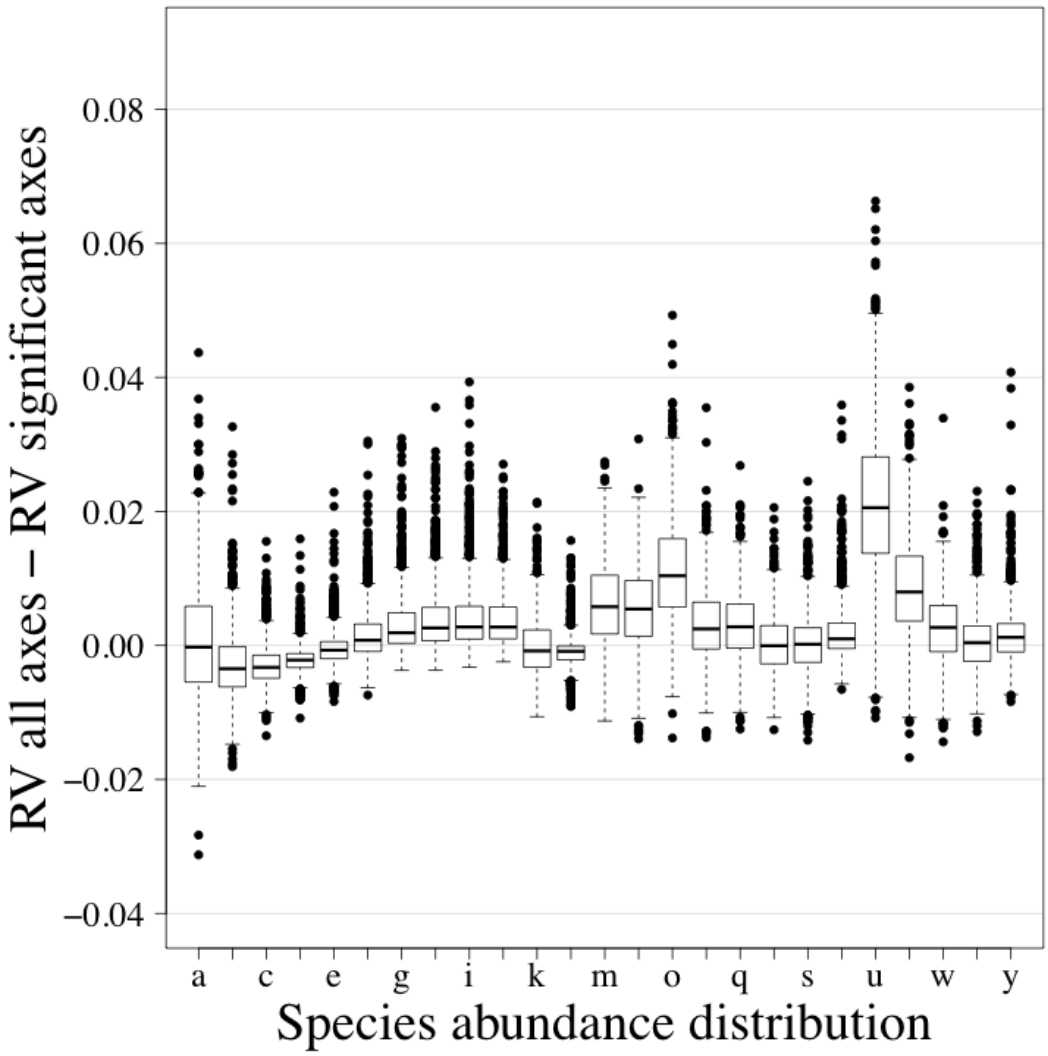

FIGURE 3C3. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from abundance data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (SAD) 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.

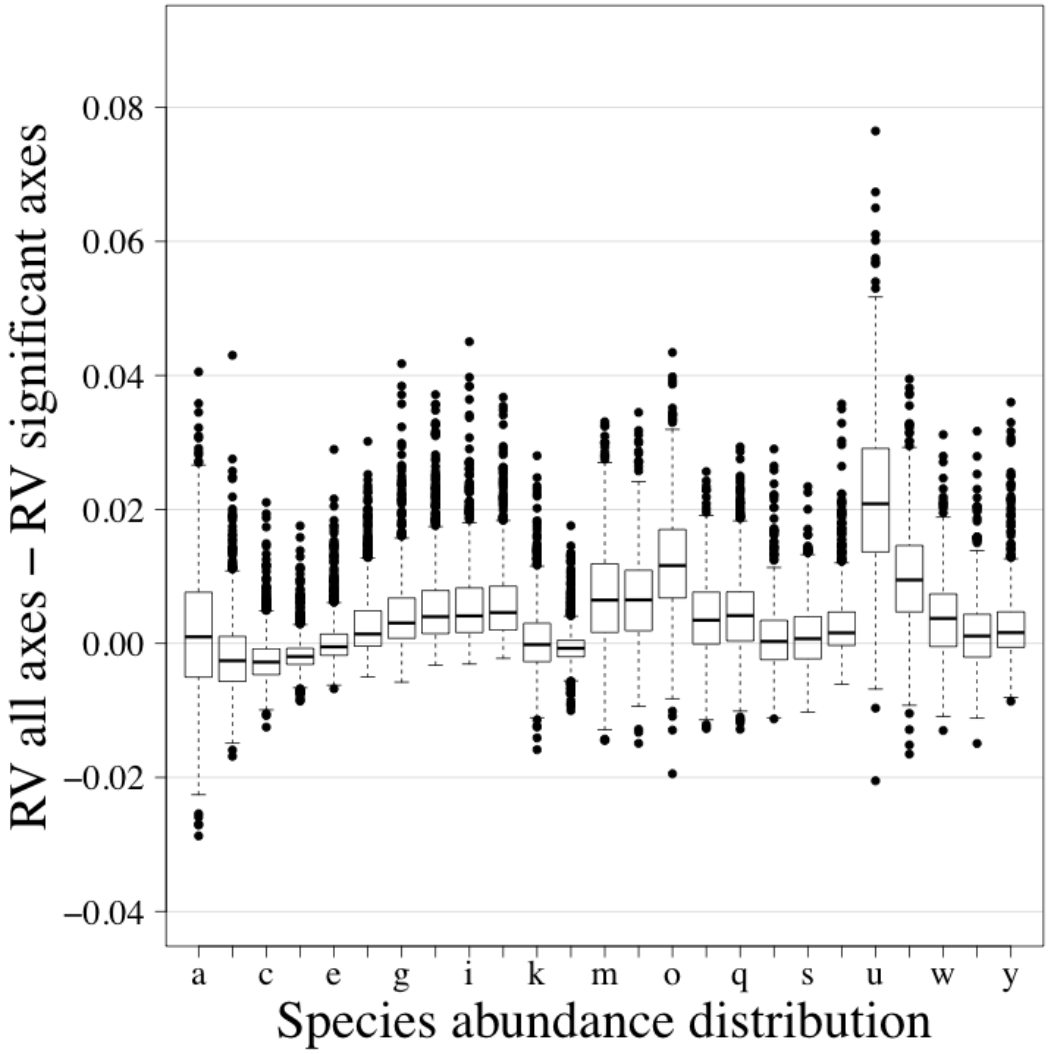

FIGURE 3C4. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from abundance data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (SAD) 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.

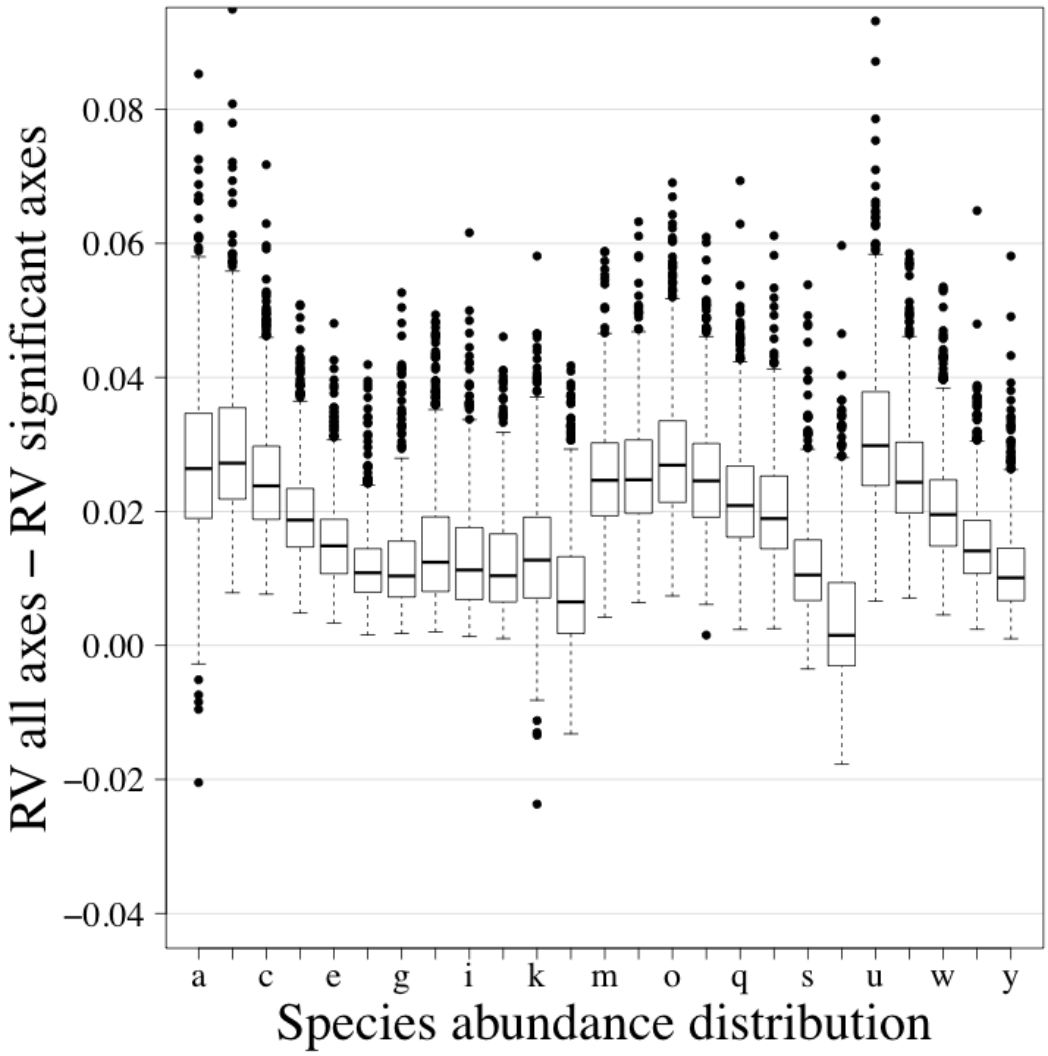

FIGURE 3C5. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from presence-absence data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (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.

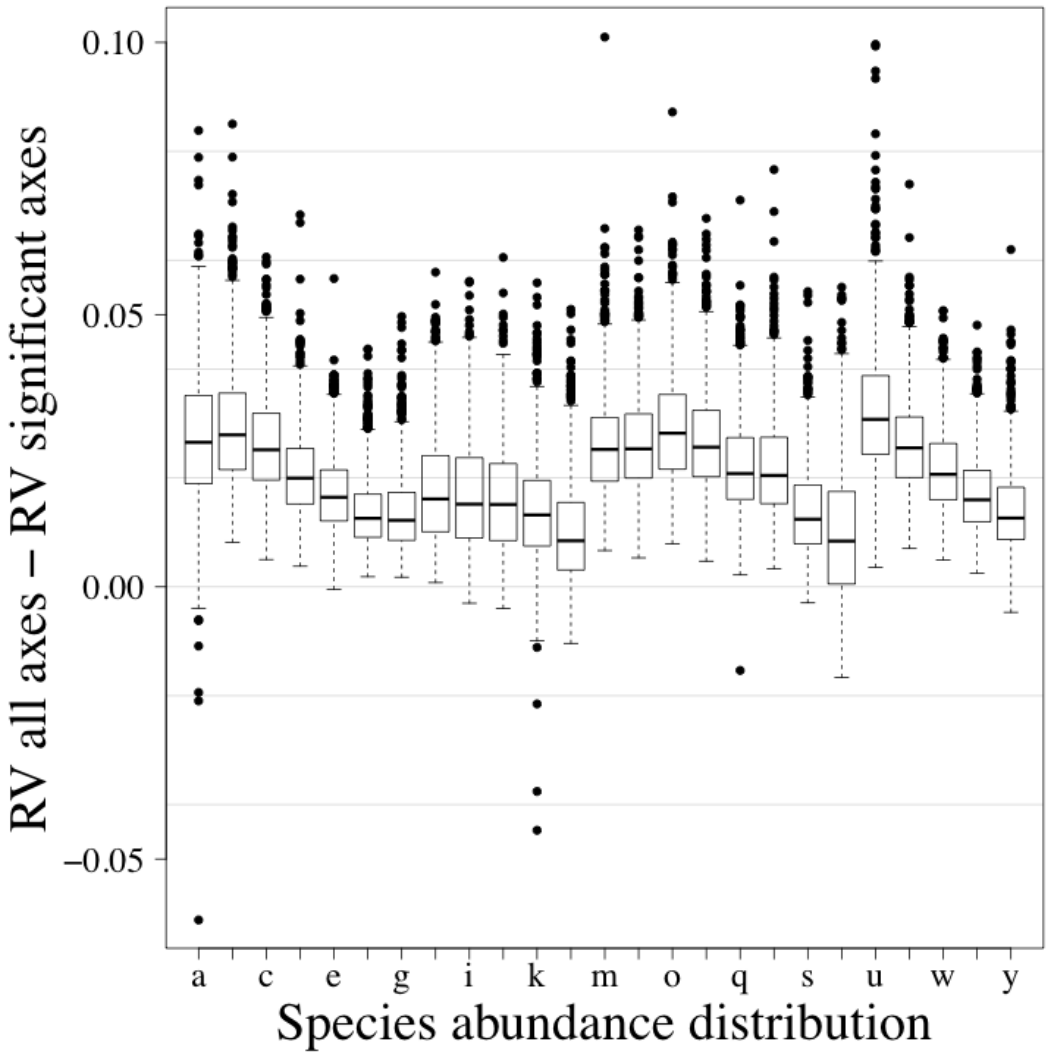

FIGURE 3C6. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from presence-absence data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (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.

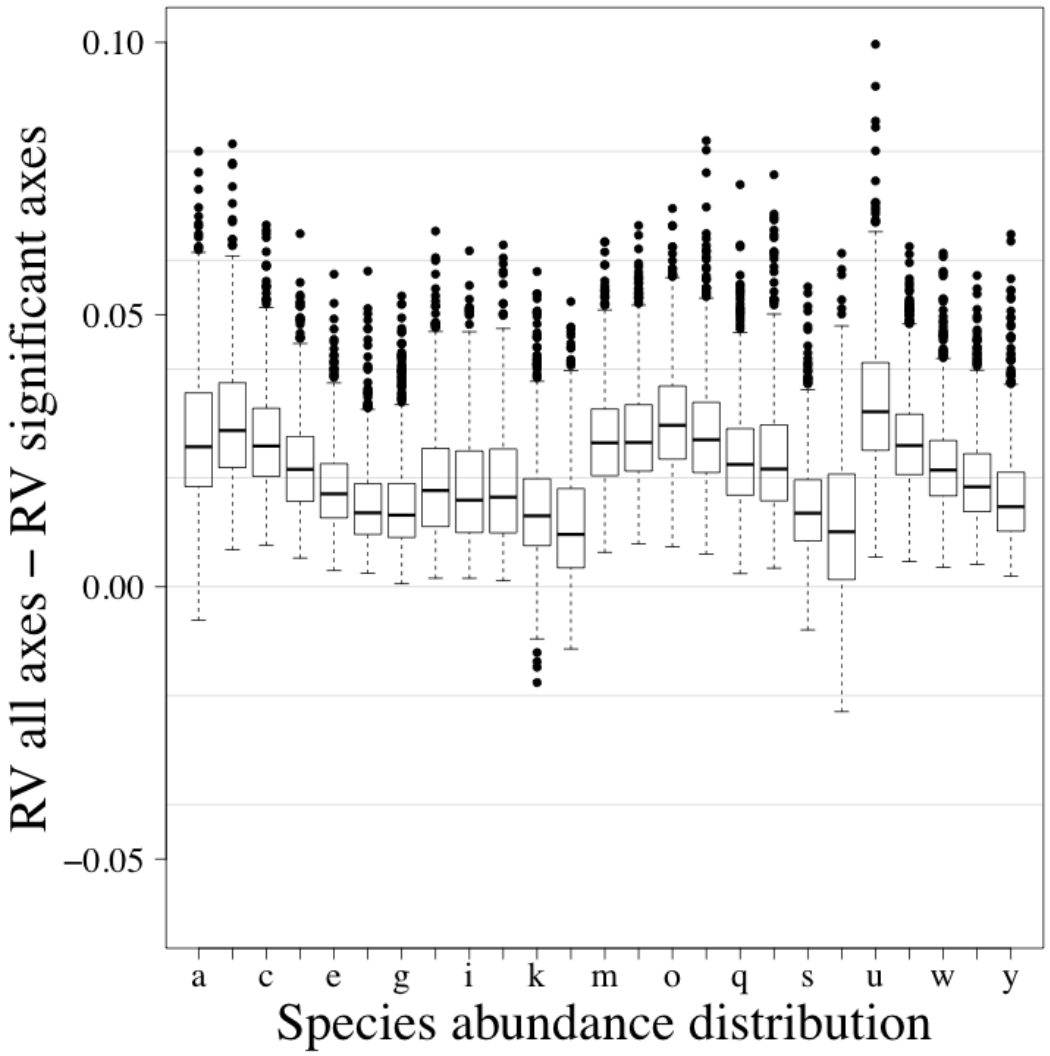

FIGURE 3C7. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from presence-absence data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (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.

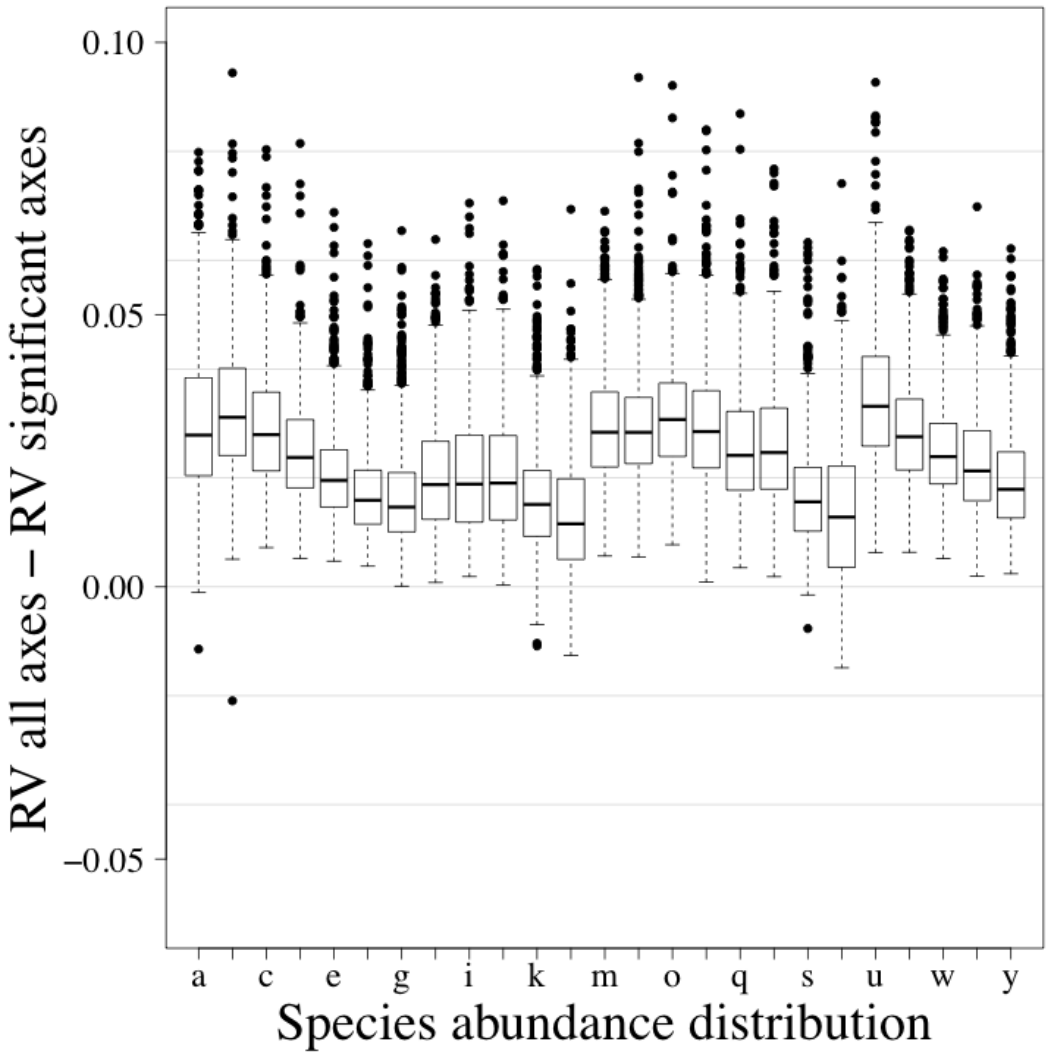

FIGURE 3C8. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from presence-absence data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (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.

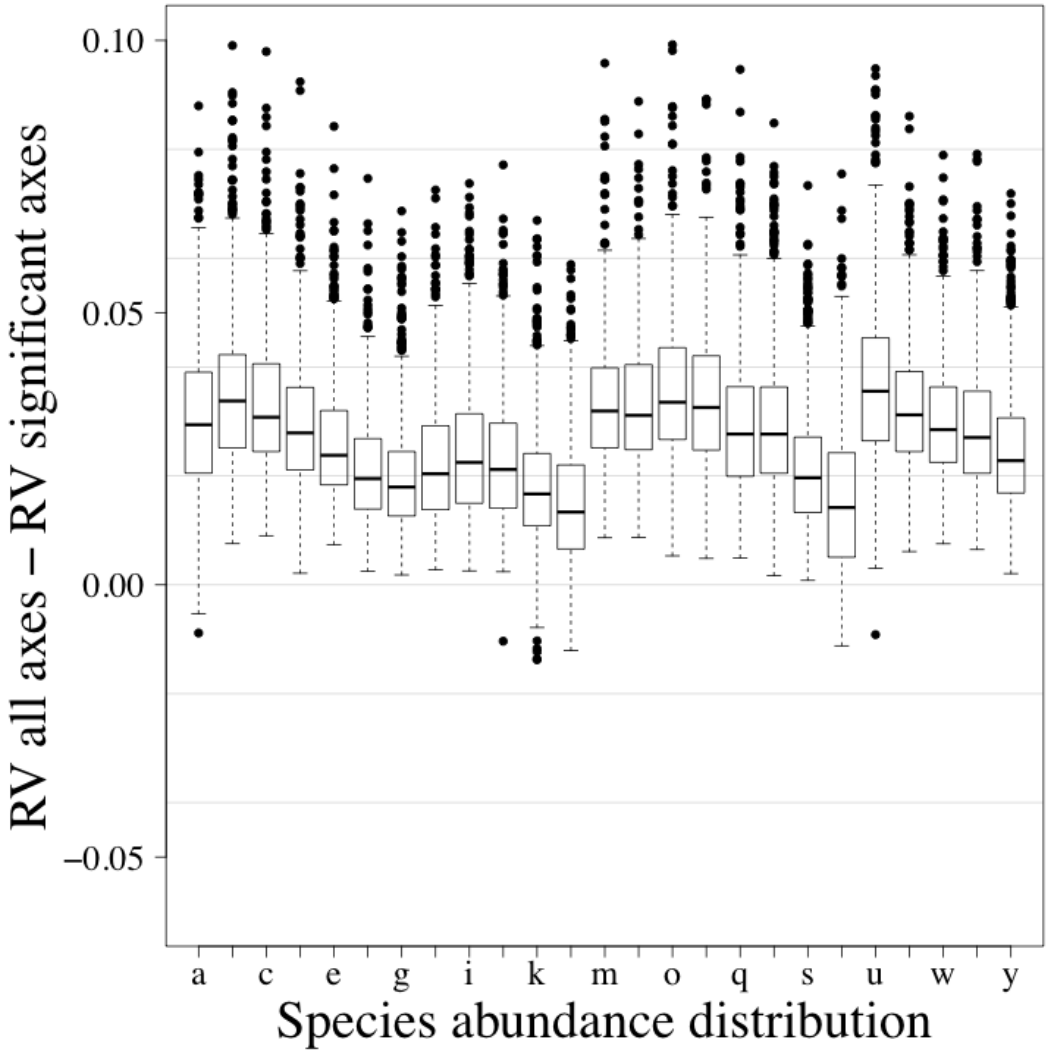

FIGURE 3C9. Comparison of consensus RDAs constructed using all canonical axes with consensus RDAs using only significant canonical axes. The $\mathbf{Z}^*$ matrices calculated from presence-absence data were used in the comparison. Letters along the abscissa refer to the species abundance distribution (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 3D

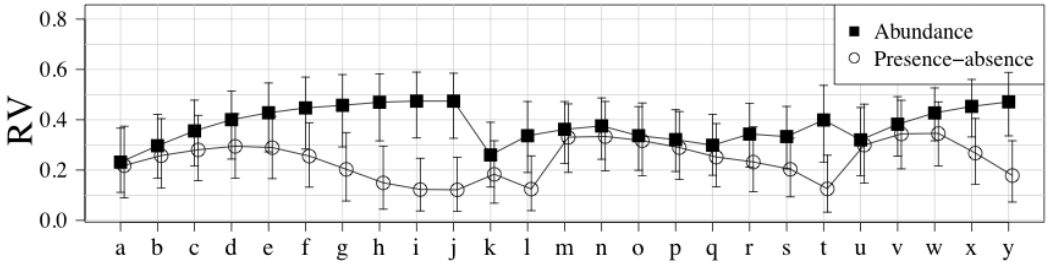

FIGURE 3D1. Comparison between abundance and presence-absence data to know how much of the true species structure (Equation 3.6 without the error term) is modelled by the canonical ordination models. For each data type (abundance and presence-absence), the significant canonical axes for all association coefficients (with the exception of the 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) as presented in Figure 3.1. A line was drawn between each SAD of each association coefficient to ease comparisons between the two data types. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 0.25). A thousand simulations were run for each SAD.

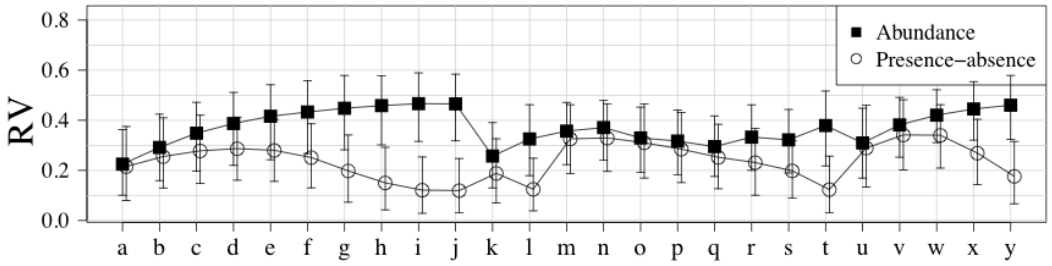

FIGURE 3D2. Comparison between abundance and presence-absence data to know how much of the true species structure (Equation 3.6 without the error term) is modelled by the canonical ordination models. For each data type (abundance and presence-absence), the significant canonical axes for all association coefficients (with the exception of the 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) as presented in Figure 3.1. A line was drawn between each SAD of each association coefficient to ease comparisons between the two data types. Results are based on species simulated with an error term sampled from a Normal distribution (mean = 0, standard deviation = 0.5). A thousand simulations were run for each SAD.

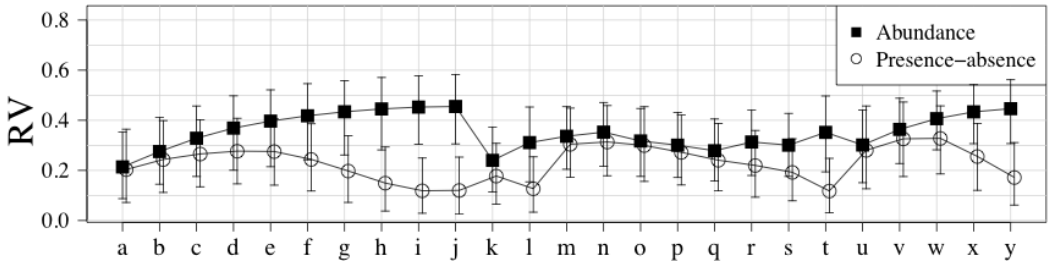

FIGURE 3D3. Comparison between abundance and presence-absence data to know how much of the true species structure (Equation 3.6 without the error term) is modelled by the canonical ordination models. For each data type (abundance and presence-absence), the significant canonical axes for all association coefficients (with the exception of the 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) as presented in Figure 3.1. A line was drawn between each SAD of each association coefficient to ease comparisons between the two data types. 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.

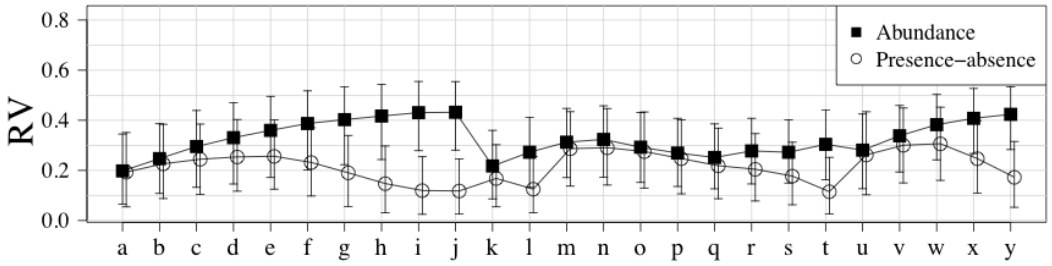

FIGURE 3D4. Comparison between abundance and presence-absence data to know how much of the true species structure (Equation 3.6 without the error term) is modelled by the canonical ordination models. For each data type (abundance and presence-absence), the significant canonical axes for all association coefficients (with the exception of the 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) as presented in Figure 3.1. A line was drawn between each SAD of each association coefficient to ease comparisons between the two data types. 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 3E

Species code and name for Carabidae and tree species

TABLE 3E1: 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 3E2: 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	*Abies balsamea*
Ll	Tamarack	*Larix laricina*
Pb	Balsam poplar	*Populus balsamifera*
Pc	Lodgepole pine	*Pinus contorta*
Pm	Black spruce	*Picea mariana*

Chapter 4 –A new cost-effective approach to survey ecological communities

Introduction

In ecological studies, the data collected in the field or obtained from laboratory experiments are the window through which we look to describe the patterns found in nature and to understand the processes that allow these patterns to emerge. Data collection is undoubtedly the most important step of any ecological study because if data acquisition is badly performed, data analysis cannot yield good results. Deciding how ecological data should be obtained is of crucial importance. This starts with the sampling design, which must be related to the ecological question or to the hypothesis to be tested.

In community ecology, researchers have proposed many different approaches to sample organisms (e.g., Anderson 1965, Martin 1977). The resulting data are usually in the form of either presence-absence or abundance. There are pros and cons for data collection and analysis of either data type. It is usually more time and cost effective to obtain presence-absence data; however, accuracy (the detail of the information the data convey) is lost because the information is only about where a species occurs. In contrast, abundance data may be tedious to obtain, but the data are more informative, and more knowledge about ecological processes can be gained from them.

Spatial or temporal distribution of individuals is an aspect that influences the number of individuals of a species that can be counted within a sampling unit (SU). Generally, individuals of a species are clustered in space or through time. Aggregation of species in space may be the result of animal behaviour, dispersal limitation, or environmental patchiness to which organisms respond (Legendre

and Fortin 1989). Through time, species succession and reproductive cycles may also generate clustered patterns (Legendre and Legendre 2012, Chapter 12). Clustered patterns of species in space or time usually lead to a lower α diversity, compared to species that are randomly or regularly distributed. If species are aggregated, they are generally found in large abundances in some SUs if the size of a cluster is smaller than the span of a SU. When highly aggregated species are sampled, many individuals are found only in one or a few SUs. In that instance, the information lost by recording only presence-absence data can be very large. However, the cost of counting all individuals (abundances) for the same community can be overwhelming. It may also be unethical to count all individuals, for example when species determination requires killing individuals belonging to rare species.

We first examined whether counting all individuals of a species in each SU is necessary to identify the distribution patterns characterizing a community. This was done to study how abundance distributions and aggregation influence the number of individuals of species found in a SU. We then devised a method to determine a counting threshold, which is the maximum number of individuals per species that needs to be counted within a SU to extract sufficient information to represent the multivariate structure of the community (as if all individuals had been counted). Regardless of the species considered, when the counting threshold is reached within a SU, the patterns describing the variance of a community should be as apparent as if all individuals had been counted.

We constructed an example explaining the counting procedure we are proposing. Table 4.1 (top) shows the complete abundance of five fictitious species

at two SUs. Note that all individuals need to be counted to obtain these data. Assuming that the patterns defining this fictitious community are apparent if a counting threshold of 8 individuals is reached, the resulting community matrix would be the one presented in Table 4.1 (bottom). Whenever there are 8 or more individuals for a species in a SU, a count of 8 is recorded. For abundances smaller then 8, the counted abundance is recorded.

The counting method we propose in this paper aims at finding a balance between presence-absence and abundance data that maximizes cost-efficiency when surveying ecological communities. Because abundance and aggregation patterns can vary in many ways, our aim is not to find a universal counting threshold that applies to all communities. Rather, we propose a general procedure to determine the counting threshold for any particular community of interest.

The procedure proposed in this paper is validated using simulations. To illustrate how this procedure can be applied to real ecological data, we implemented it for a community of boreal forest Carabidae sampled in northwestern Alberta, Canada.

FROM PRESENCE-ABSENCE TO ABUNDANCE

Presence-absence and abundance data are extremes of a spectrum of data formats characterizing composition and distribution of communities. Intermediate cases between these two extremes can be found by counting individuals of each species within a SU until either a predefined (user-defined) counting threshold is reached or all individuals of a species within the SU are accounted for, as illustrated in Table 4.1. By sequentially increasing the counting threshold from 1

to the largest number of individuals for a species found within a SU, all intermediate cases can be studied from presence-absence to full abundance data. We will refer to the case where all individuals are counted as "complete-abundance" counts while all cases with counts of fewer individuals will be referred to as "partial-abundance" counts.

In this paper we consider a species as abundant if it is found with high abundance in at least one SU. As a rule of thumb, we consider the abundance of a species to be high if it counts more individuals than there are SUs. Conversely, a rare species can potentially be found in many SUs but its abundance may be low in all SUs. Given these definitions, modest variation in abundance do not generally influence the interpretation of the patterns of variation of abundant species but can importantly impact the interpretation of rare species. Based on this premise, use of partial-abundance counts instead of complete-abundance counts can effectively produce information about rare species while the associated loss of information for common species is largely inconsequential to understanding community variance patterns. The challenge, therefore, is to find the lowest counting threshold that efficiently and accurately allows the description of the variance pattern of a community.

SIMULATING ECOLOGICAL COMMUNITIES

Species abundance distribution (SAD) and patterns of spatial or temporal aggregation of species vary among communities. As we argued in the *Introduction*, these are the two major components of a community that will influence the choice of a counting threshold. To evaluate how these two

components influence the efficiency of partial-abundance data for characterizing patterns of variation in a community, we simulated community matrices comprised of 100 SUs and 50 species. Sample size and species richness should not influence the counting threshold required for community patterns to be accurately characterized because these two components do not impact the spatial or temporal aggregation patterns of species or the positively skewed abundance distribution typical of ecological communities.

In our simulations, the sampling area (i.e., the area where all samples are collected) was a square of unit size. The abundance of each species in the community ranged from 1 to 500. A probability was given to each abundance value following a lognormal distribution (Preston 1948) with a standard deviation of 5. In that lognormal distribution, the probabilities of finding a species with an abundance of 1, 2, or 3 individuals in the sampling area are respectively, 0.18, 0.090, and 0.059. Note that a standard deviation of 5 was the smallest for which a community could be generated where at least one individuals was found at all of the 100 SUs of the community matrix. Because multivariate analyses commonly used in community ecology, such as χ^2-based ordinations (e.g., principal component analysis after χ^2 transformation of the data, and correspondence analysis), have trouble handling situations where SUs where no individuals of any species were counted, we did not simulate these cases.

For each species of a simulated community, the spatial position of each individual was specified using a Matérn cluster process (Illian et al. 2008), in which a homogeneous Poisson process is used to define the position of the cluster centers. The intensity of the Poisson process is used to determine the number of

clusters to generate. In a statistical context, the intensity of the Poisson process is also the average of a Poisson distribution. In our simulations, the intensity of the Poisson process was defined by random selection of an integer value between 1 and the species abundance previously obtained from the lognormal distribution. Using this approach, the number of clusters is on average equal to the randomly selected value. This generates species with spatial patterns ranging from aggregated into one patch (if a single cluster is generated), to randomly dispersed where each individual is a separate spatial cluster (if the number of clusters equals the number of individuals in a species).

In the Matérn cluster process, a parameter defines how many individuals should on average be included in each cluster. We obtained that parameter by dividing the abundance chosen from the lognormal distribution by the intensity of the Poisson process. Individuals within each spatial cluster are uniformly distributed. The radii of spatial clusters are used as surrogates of aggregation levels because it is the zone of influence of a cluster. For example, if a cluster has a radius of one meter, all individuals associated to that cluster will be anywhere within a one-meter radius of the centre of the cluster. Radii of spatial clusters are chosen by the user. The Matérn cluster process also allows clusters to overlap, potentially presenting patches of individuals of different shapes and sizes. A variety of spatial patterns can thus be generated even though the radii of all clusters are the same when individuals of a species are grouped into more than one cluster. The simulated species spatial distribution can thus represent patterns similar to what is found in nature. Unlike the number of clusters and the species

abundance, which are related to each other in the Matérn cluster process, the choice of cluster radii is not influenced by other parameters.

Because the Matérn cluster process is a random process that relies on the three parameters described above (the intensity of the Poisson process that defines the number of clusters, the average number of individuals within each cluster, and the radii of clusters), the total number of individuals of a species generated by the Matérn cluster process may vary around the abundance value defined by the random sampling of the lognormal distribution. We inspected the abundance patterns of all simulated communities to ensure that the random variation resulting from the Matérn cluster process did not make the resulting abundance distribution diverge markedly from the reference lognormal SAD defined in the first step of the simulation. The random variations introduced by the Matérn cluster process had only minor influences on the abundance distribution, which will not affect the following steps of the simulations.

Keeping in mind that the sampling area is a square of unit size, we generated a first set of communities where the range of aggregation was broad (cluster radii varied from 0.01 to 0.5) and another in which individuals were highly aggregated (cluster radii varied from 0.01 to 0.02). For all species in each set of communities, cluster radii were randomly sampled from a uniform distribution within the cluster radii range. For each aggregation level, we simulated 1000 communities. The spatstat package (Baddeley and Turner 2005) was used through the R statistical language (R Development Core Team 2012) to simulate these communities.

Simulations based on the same parameters were also performed, where the broken-stick model (McArthur 1957) was used as the reference SAD instead of the lognormal distribution to define the probability of choosing the abundance of a species. In the broken-stick model, the probability of finding a species with an abundance ranging from 1 to 500 is defined by randomly cutting a conceptual stick of unit length 499 times. The broken stick pieces are then ordered from the longest to the shortest to define the probability of sampling a species with an abundance of 1 through 500. Because the length of the stick pieces in the broken-stick model can vary between iterations, we used the expected stick lengths to choose the abundance of a species (Barton and Davis 1956). The probability of sampling any abundance *n* from 1 to 500 can be obtained from $(\sum_{n=1}^{500} 1/n)/500$. For example, the probability of sampling exactly one individual of a species is 0.0136 whereas the probability of sampling exactly 500 individuals of a species is 0.000004. The broken-stick model is negatively skewed compared to the lognormal distribution. The lognormal distribution and the broken-stick model are commonly used to model SADs, making them relevant choices to define our simulated community abundances.

We also simulated three other sets of communities where the abundance of all species was defined as 500. The first two sets of communities used the Matérn cluster process to distribute individuals in a square sampling area of unit size following the procedure described above where the number of clusters was randomly selected between 1 and 500 and the cluster radii ranged from 0.01 to 0.02 for the first set of communities, and from 0.01 and 0.5 for the second set of

communities. As previously explained, because the Matérn cluster process is a random process, the exact abundance of each species was not necessarily 500; it often diverged slightly from that value. In a third set of communities, the locations of individuals in the unit size sampling area was defined using a uniform distribution (minimum = 0, maximum = 1) for the X and Y coordinates. In this set of simulated communities, the abundance of each species was exactly 500 because the Matérn cluster process was not used to distribute the individuals in the sampling area. This last set of communities differs from the others in that all species were not clustered but randomly distributed in the sampling area. These three sets of communities were used to evaluate the importance of the abundance distribution and aggregation patterns in defining a counting threshold.

Finally, we divided the sampling area into 100 non-overlapping SUs of equal sizes (the SUs completely covered the study area) and counted the number of individuals of each species in each SU for all simulated communities. This count provided the complete-abundance community matrix. Note that although in these simulations the SUs completely covered the study area, this condition is not necessary for the counting approach we propose. In the *Ecological illustration*, we apply our procedure to an experimental research area where the SUs covered only a small fraction of the whole study area.

CORRELATION OF ALL PARTIAL-ABUNDANCE TO THE COMPLETE-ABUNDANCE DATA

To evaluate how much information is included in the increasingly precise partial-abundance data, we used the RV coefficient (Escoufier 1973, Robert and

Escoufier 1976) to correlate the partial-abundance community matrices with the complete-abundance community matrix. The RV coefficient is defined by:

$$RV(\mathbf{X}, \mathbf{Y}) = \frac{\mathrm{tr}(\mathbf{X}\mathbf{X}^t\mathbf{Y}\mathbf{Y}^t)}{\sqrt{\mathrm{tr}((\mathbf{X}\mathbf{X}^t)^2)\mathrm{tr}((\mathbf{Y}\mathbf{Y}^t)^2)}} \tag{4.1}$$

where **X** and **Y** are two column-centered matrices with the same number of rows, *t* is the transpose of a matrix, and tr the trace of a matrix. The RV coefficient is a multivariate extension of the squared Pearson correlation coefficient. It ranges from 0 (no correlation) to 1 (perfect correlation). When comparing species communities defined by partial and complete-abundance counts, the RV coefficient measures the accuracy in the estimation of the spatial (temporal) variation of the data.

Figure 4.1 presents the correlation results between the increasingly accurate partial-abundance data (abscissa) and the complete-abundance data for the seven different sets of simulated communities. In this figure, the first row highlights the results obtained from the two SADs (lognormal distribution and broken-stick model) while the second row presents communities where species are all abundant (~500 individuals) and the aggregation patterns range from random to highly aggregated. The most striking result was that the stronger the aggregation of species, the more individuals had to be counted to reach the same RV coefficient, compared to communities where species had a wider range of aggregation or were randomly distributed; compare panels a and b, c and d, f and g in Figure 4.1. Another noteworthy observation was that in all simulated communities, including those in which individuals were randomly distributed in the study area (Figure 4.1e), the number of individuals that needed to be counted

to reach a high RV coefficient (e.g. > 0.9) was much smaller than the maximum number of individuals per species per SU found in the complete-abundance data. In Figure 4.1, this is illustrated by the long horizontal line showing an RV coefficient = 1 found for all sets of communities.

These results confirm our hypothesis that the effort placed on counting all individuals is not necessary to reconstruct and analyze the variance of communities. Finally, although the abundance distribution used to construct communities influenced the number of individuals needed to obtain a good representation of the complete-abundance data, a positively skewed abundance pattern (typical of SAD found in nature) is not necessary. Regardless of the SAD used, it is not required to count all individuals to identify the variance patterns defining a community. This is an important observation because it implies that the counting approach we are proposing is not limited to community data, it can be used on virtually any type of multivariate count data sets. However, for the remainder of this paper we will continue to present our results and interpretation in terms of ecological communities, species, and sampling units.

In this paper we use the lower bound of the 99% confidence interval exclusively to elaborate conclusions. By using only these extreme scenarios of our simulation results, we minimize the impact of losing important information by counting too few individuals.

If we assume that the minimum RV coefficient required between partial and complete-abundance count needs to be at least 0.9 to give an acceptable representation of the community patterns, we can evaluate the cost of counting partial-abundance data with a good level of accuracy. In the most optimistic case,

where all species are composed of 500 individuals uniformly distributed in the sampling area (Figure 4.1e), a 0.9 RV coefficient is reached with a counting threshold of 7 individuals, an RV coefficient of 0.95 is attained with a counting threshold of 8 individuals, and a 0.99 RV coefficient requires a counting threshold of 11 individuals. These results are interesting because they show that with randomly distributed individuals in space or time, it is possible to be very cost-effective when counting.

At the other end of our simulation spectrum, when species are all abundant (i.e. composed of ~500 individuals) but highly aggregated (Figure 4.1g), to reach a 0.9, 0.95, and 0.99 RV coefficient between partial and complete-abundance data, a counting thresholds of 293, 343, and 430 individuals are needed, respectively (lower bound of the 99% confidence interval).

By comparing the results presented in Figure 4.1, we show that for the ecological situations that we simulated, aggregation is the dominant factor increasing the number of individuals that need to be counted to reach a predefined RV coefficient. Thus, because species are known to aggregate in space and time, we must ask whether this procedure is interesting for community data?

CORRELATING PARTIAL TO COMPLETE-ABUNDANCE DATA USING ECOLOGICALLY MEANINGFUL DISTANCES

Multivariate analyses of communities are rarely carried out on raw count data because using raw count data is equivalent to performing an ordination (or a clustering) analysis based on Euclidean distances. Euclidean distance is appropriate to answer ecological questions focusing on phenomena that cause

changes in total abundances, such as disturbances or predation, but it is ill-adapted for other types of ecological questions (Anderson et al. 2011). Numerous other distances have been proposed to study patterns in community data resulting from habitat variation. Legendre and Legendre (2012, Chapter 7) described many distances specifically designed for modeling a variety of ecological data.

Currently, community data are almost always analyzed with tools that use distances other than the Euclidean to extract ecological patterns. A typical example is the widespread use of correspondence analysis (CA, e.g., Greenacre 2007) and its canonical counterpart CCA (ter Braak 1986), which involves the χ^2 distance. Other distances are commonly used to analyze the variation of community composition data. In these instances, it becomes relevant to study how partial-abundance counts can accurately characterize the community patterns defined by complete-abundance data using distances other than the Euclidean distance. We can then evaluate if by using different distances, the information in the complete-abundance data can be recovered by counting a smaller number of individuals of each species.

Six distances commonly used with community composition data were considered: the Hellinger (Rao 1995), chord (Orlòci 1967, Cavalli-Sforza and Edwards 1967), the distance between species profiles (Legendre and Gallagher 2001), χ^2 (Lebart and Fénelon 1971), percentage difference (Odum 1950), and modified Gower using a base 2 logarithm (Anderson et al. 2006). All these distances are well adapted to the analysis of community composition data (Legendre and Legendre 2012, Chapter 7). The χ^2 distance is widely used in ecology because it is the basis for CA and CCA. The percentage difference (also

known as the Odum or, incorrectly, the Bray-Curtis distance) has been shown by Faith et al. (1987) to be well adapted to extract ecological patterns. Anderson et al. (2006) applied their transformation to the asymmetrical form of the Gower distance coefficient and called the combination the modified Gower dissimilarity (or distance). Anderson et al. (2006) transformed all abundances in a community matrix using a logarithm of the abundance + 1, with the exception 0s, which remain unchanged. When calculating the modified Gower distance, an increase in the base of the logarithm decreases the emphasis on abundances. For this reason, we chose the modified Gower using a base 2 logarithm because any larger base of logarithm would give less importance to abundant species and thus make it easier to find a higher correlation between partial and complete-abundance data.

When applied directly to a community matrix, each of the distances presented in the previous paragraph yields a symmetric distance matrix. However, it is also possible to transform a community matrix in such a way that a distance other than the Euclidean distance is preserved. Two different approaches were followed to transform the community data (partial and complete), depending on the distance used (Figure 4.2). (1) Legendre and Gallagher (2001) have shown that the Hellinger, chord, species profiles, and χ^2 transformations can be applied directly to a community matrix using pre-transformations without calculating a distance. A pre-transformation is a transformation applied to a community matrix before any analyses are carried out; the transformation changes the distance preserved between SUs in the analysis using a linear model such as principal component analysis (PCA), redundancy analysis (RDA), or *K*-means partitioning. Calculating the Euclidean distance of a pre-transformed community matrix yields

a symmetric distance matrix where the distance between each pair of SUs is the distance corresponding to the pre-transformation. We thus pre-transformed all partial-abundance community data and correlated them to the pre-transformed complete-abundance community matrix using the RV coefficient (Figure 4.2a). (2) The percentage difference and modified Gower using a base 2 logarithm distances cannot be obtained by pre-transforming a community matrix. To compare partial and complete-abundance for these two distances, we first calculated the distance matrices for all partial-abundance community matrices and the complete-abundance community matrix. We then performed a principal coordinate analysis (PCoA, Gower 1966) independently on each distance matrix (partial and complete) and used all the eigenvectors of each partial-abundance community data and correlated them with all the eigenvectors from the complete-abundance data using RV coefficients (Figure 4.2b). The PCoA is not used here as a dimension reduction tool, it is used to transform a distance matrix into a matrix with the same format as the original community matrix but where the species (columns of the community matrix) are replaced by eigenvectors (Figure 4.2). Performing a PCoA on non-Euclidean distance matrices may generate complex eigenvectors (Legendre and Legendre 2012, Subsection 9.3.4), which are difficult to handle. To ensure that no complex eigenvectors are generated when the percentage difference is used, we square-rooted all percentage difference distance matrices. This makes the percentage difference a metric and the resulting distances have the Euclidean property (Legendre and Legendre 2012, Subsection 7.4.2), which ensures that no complex eigenvectors are generated when PCoA is applied to a square-rooted percentage difference distance matrix. Applying a

square-root transformation to a modified Gower distance matrix, however, may not make it Euclidean. For this reason, we added a constant equal to the largest positive eigenvalue to all values of each modified Gower distance matrix to ensure that was Euclidean and that no complex eigenvectors were generated from the PCoA (Gower and Legendre 1986, Legendre and Legendre 2012, Subsection 9.3.4). This procedure is known as the Cailliez correction (Cailliez 1983).

All of the calculations presented in the previous paragraph were performed with the vegan package (Oksanen et al. 2012), with the exception of the PCoA, which was carried out with the stats package (R Development Core Team 2012). All calculations were performed within the R statistical language.

To compare the different distances, we focused on simulated communities where the abundances of all species were large (~500 individuals) and species were highly aggregated. We chose to focus on this set of communities instead of any other set because they required the most individuals to reach the same RV coefficients between partial and complete-abundance data than when raw count data were used (Figure 4.1g).

The results in show that regardless of the transformations used, the 0.9, 0.95, and 0.99 RV coefficients between partial and complete-abundance are reached with many fewer individuals than when raw community data are used (Figure 4.3). Of the distances compared, the distance between species profiles required the largest number of individuals to reach the same RV coefficients between partial and complete-abundance (Figure 4.3c). To reach a 0.99 RV coefficient between partial and complete-abundance, a counting threshold of at least 350 individuals was needed. Although it is the worst of the distances

compared, it is still much more efficient than using raw count data (compare to Figure 4.1g).

The χ^2 (Figure 4.3d) and the Hellinger (Figure 4.3a) distances also required many individuals to reach a predefined RV coefficient. To attain the 0.99 RV coefficient value, a counting threshold of 224 individuals was needed for the χ^2 distance and of 207 for the Hellinger distance. The percentage difference (Figure 4.3e) and chord (Figure 4.3b) distances were more efficient since they required counting thresholds of 160 and 102 individuals, respectively, to reach a 0.99 RV coefficient between partial and complete-abundance data. For these simulated data, the most striking result was obtained from the modified Gower (Figure 4.3f) distance: a counting threshold of merely 9 individuals was needed to reach a 0.99 RV coefficient between partial and complete-abundance data.

The results found with the other sets of simulated communities are presented in Appendix 4A. They yield the same conclusion as discussed in this paragraph but a lower counting threshold was needed when any of the other six sets of simulated communities were used.

The results that stem from Figures 4.1 and 4.3 are interesting because they show that few individuals need to be counted in a community for the variance patterns to be identified. However, these results are not helpful when planning a survey because they are not useful to suggest a counting threshold that needs to be reached before all sites have already been sampled.

PILOT STUDY: THE BASIS FOR A NEW SAMPLING PROCEDURE

The generality of the results presented in Figures 4.1 and 4.3 makes it possible to apply the same procedure on a reduced number of randomly selected sampling units, in a pilot study, to estimate the counting threshold required for a sample that provides a good representation of the actual community. We propose to use pilot studies as a tool to improve the cost in time and money of a study. In this paper, a pilot study is a study used as a reference to estimate a counting threshold. It can result from data collected in a previous sampling year or it can include SUs selected at random or in a systematic design in the sampling area of an ongoing study.

The size of the pilot study is important to infer a counting threshold. A pilot study that includes two SUs will likely not yield the same counting threshold as one that comprises ten SUs. From the results in Figures 4.1 and 4.3, we know that it is possible to estimate community patterns by counting a fraction of all the individuals. In this section, our goal is to evaluate the minimum number of SUs that needs to be randomly sampled in a pilot study to ensure that the counting threshold associated to a particular RV coefficient can be reached.

To ensure that our simulation results can be used as a reference for studies on real communities, we estimated the counting threshold of the pilot study data by constructing a 99% empirical confidence interval from the RV coefficient correlating partial and complete-abundance of the full survey data. We consistently referred to the lower bound of the confidence interval. In other words, we referred to extreme cases where the number of individuals to count is large. Also, the choice of the SUs in the pilot study may influence considerably

the estimation of the counting threshold, especially when the number of SUs within a pilot study is small. In practice, the SUs in the pilot study should always be selected randomly throughout the sampling area. However, for our simulation to be exactly reproducible and to represent the worst-case scenario that can be designed, we selected the SUs to be included in the pilot study to form an abundance gradient, from the ones that contained the lowest maximum counts of individual for any one species to the ones that presented the largest counts. For example, referring to the complete-abundance in Table 4.1 (top), sampling unit 2 would be considered first because the maximum count of individuals for any one species is 500 (species A). Sampling unit 1 would follow because the maximum count for any one species is larger (900 individuals for species E). This simulation procedure has an additional advantage over randomly choosing the SUs, it makes the number of individuals that are needed to reach a predefined RV coefficient larger than for all other situations. This constraining approach to consider SUs in a pilot study was chosen to ensure that the counting threshold was not underestimated in the pilot study.

Using the communities previously simulated, we carried out a pilot study that included 3% of the SUs. We then correlated the increasingly accurate partial-abundance data with the complete-abundance data using RV coefficients, but referring only to the data collected within the pilot study. This is the same procedure as used in the previous sections. Note that the number of species in the pilot study will not affect the estimation of the counting threshold because the procedure we propose focuses on variation at the individual level. We then compared the 0.9, 0.95 0.99, 0.999, and 0.9999 RV coefficients calculated from

the pilot data with the 0.9, 0.95, 0.99, 0.999, and 0.9999 RV coefficients computed from the full-survey data. We repeated the same procedure using pilot studies that included 4%, 5%, …, up to 100% (the whole sampling area) of the SUs. This procedure was carried out for each set of simulated communities and using all the distances considered in the previous section.

In our simulations, we know the abundance and aggregation patterns of the sampled species because these parameters formed the basis of our artificial communities. However, such patterns are difficult to evaluate using only data obtained from a pilot study. For this reason, through our interpretation of Figure 4.4 and Figures 4B1-4B7, we decided to consistently select the number of SUs where the RV coefficient calculated with all SUs was exceeded. This ensured that the survey-wide RV coefficients between partial and complete-abundance were reached even in the most difficult scenarios.

Figure 4.4 presents the results calculated for the set of communities where the abundance of each species was large (~500 individuals for each species) and species had a broad range of aggregation levels. With this set of communities, a pilot study generally required more SUs to reach a predefined survey-wide RV coefficient. This set of communities presents the worst-case scenario we simulated for all the distances we compared, except for the modified Gower distance, which performed poorly when the abundance pattern of the communities followed a lognormal distribution (Figure 4B7a). Focussing on the worst cases makes our interpretation of these results more conservative because more individuals need to be counted in this set of communities.

From these results, the modified Gower distance has a clear advantage over the other dissimilarities for the type of communities that we simulated. It is the only distance to produce a survey-wide 0.95 RV coefficient calculated between partial and complete-abundances based on 3% of the SUs. For this 0.95 RV coefficient to be reached, the RV coefficient calculated between partial and complete-abundances within the pilot study data needs to be at least 0.999. However, we favor using a 0.9999 RV coefficient between partial and complete-abundances to reach the survey-wide 0.95 RV coefficient because Figure 4B7a shows that with a 0.999 RV coefficient, the pilot study barely meets the criterion that defined the 0.95 RV coefficient calculated using all SUs. If higher accuracy is required, the modified Gower is the only distance that is worth using. To reach a survey-wide 0.99 RV coefficient, at least 47% of the study area needed to be surveyed with a pilot study where a 0.9999 RV coefficient was used as a reference.

The percentage difference is the next best choice of distance after the modified Gower distance. We can expect that by using 37% of the study area it is possible to reach a survey-wide 0.9 RV coefficient if we refer to the pilot study 0.9999 RV coefficient.

All other distances required a pilot study to cover at least 85% of the study area to reach a survey-wide 0.9 RV coefficient, and to reach this survey-wide RV coefficient value the RV coefficient of the pilot study needs to be at least of 0.9999. If the Euclidean, chord, χ^2, and Hellinger distances and the distance between species profiles need to be used to extract community patterns, it is preferable to use a pilot study that includes at least as many SUs as what is

planned for the complete study. For example, if the same group of organisms has been sampled in previous years, such community data could serve as a reference pilot study. In the *Ecological illustration*, we show how data from a previous year can be used as a pilot study to estimate a counting threshold.

The results discussed above represent the worst scenario of our simulations. However, because species abundance distributions are always positively skewed for ecological communities, one can refer to the results obtained from simulated communities where the species abundance distributions follow a lognormal distribution or a broken-stick model (Figure 4B1-4B7). However, these simulations should only be referred to if all species sampled in a community are used. In any case, it is preferable to refer to the scenario where the number of SUs to sample is the largest, as we did in this section.

ECOLOGICAL ILLUSTRATION

To illustrate how this new method can be applied to real ecological data, we used data about boreal Carabidae data. In that study, 196 sites were sampled using pitfall traps (Spence and Niemelä, 1994) in a near-regular grid of 70 km^2 of mature boreal forest at the Ecosystem Management Emulating Natural Disturbances (EMEND) research site located in northwestern Alberta, Canada (see Bergeron et al. 2011, 2012 and in Chapter 2 of this thesis). The data include 9869 individuals defining 45 carabid species. *Calathus advena* was most abundant at any single site with 128 individuals.

By using data from all the sites, we first estimated the counting threshold for future studies with the modified Gower distance using base 2 logarithms.

Following, we estimated the counting threshold that would be needed to extract the patterns found in this carabid assemblage at different levels of accuracy, again with the modified Gower distance.

When sampling is carried out using pitfall traps, it is common to correct for disturbances (e.g., flooding of the trap) by dividing the abundance of each species by the number of days a trap was active. The procedure proposed in this paper is unaffected by that normalization because the time for which a trap was active remains constant regardless of the number of individuals of a species counted in a trap. In other words, the normalization does not affect the calculation of the counting threshold. For this reason, we can omit any normalizing procedure applied on the SUs when estimating the counting threshold.

Using all beetles sampled at the 196 sites as a pilot study, we estimated that with a counting threshold of 4 individuals, a 0.9 RV coefficient can be reached with the modified Gower distance using base 2 logarithms. This counting threshold of 4 individuals can be used for any future carabid study carried out on the EMEND landscape that includes up to 196 sites. This counting threshold of 4 individuals is valid as long as the modified Gower using a base 2 logarithm is used and that a level of correlation between partial and complete-abundance data of 0.9 (RV coefficient) is considered reasonable. As can be logically expected, a more constraining RV coefficient requires more individuals to be counted. For example, a counting threshold of 8 individuals would be needed to attain a 0.95 RV coefficient whereas a 0.99 RV coefficient would require a counting threshold of 19 individuals.

These counting thresholds can be translated into cost-effectiveness by evaluating the number of individuals that would be counted in total to reach the 0.9, 0.95, and 0.99 RV coefficient if the same 196 sites were sampled. To reach a 0.9 RV coefficient, a total of 3513 individuals would have to be counted. This is 35.6% of all the individuals counted for the full survey. To reach a 0.95 RV coefficient we would need to count 5211 individuals (52.8% of all individuals), and 7525 individuals (76.2% of all individuals) needed to be counted to reach a 0.99 RV coefficient. This evaluation of cost-effectiveness shows it is possible with real ecological data to be more efficient by counting only a subset of all the individuals.

If we were to sample carabids at 1000 sites across the EMEND landscape, what counting threshold should be reached to analyze the future data assuming the modified Gower distance with base 2 logarithm was used? If we refer to the 196 sites already sampled as the pilot study and use the simulation results presented in Figure 4B7a, we can answer that question. Because 196 sites = 19.6% of the 1000 sites we plan to sample, to reach at least a survey-wide 0.95 RV threshold between the partial and complete-abundance, a 0.999 RV threshold must be attained in the pilot study. With these parameters, we estimated that with a counting threshold of 38 individuals, we expect to reach at least a 0.95 RV coefficient between partial and complete-abundance.

Assuming now that we want to plan a survey where 200 sites are sampled to study the carabid assemblage on the EMEND landscape, as was originally the plan for the carabid study (Bergeron et al. 2011), but that no previous data is available to evaluate a counting threshold. To estimate the counting threshold, we

first have to decide the particular distance measure in which all analyses need to be performed. As for the other examples, we chose the modified Gower distance using a base 2 logarithm.

Referring to the simulation results in Figure 4B7a, we know that 6 randomly selected sites (3% of the sampling area) are required to estimate the counting threshold for a survey-wide 0.95 RV coefficient between partial and complete-abundance data by referring to the pilot study 0.9999 RV coefficient calculated between partial and complete-abundance data. Because the results in Figure 4.4 and Figure 4B7 present extreme cases, it is highly unlikely that the 6 sites chosen in the pilot study will not reach the minimum number of individuals required to reach a survey-wide 0.95 RV coefficient. In fact, it is likely that the pilot study will present a number of individuals larger than the minimum required.

As an example, if we randomly choose six sites in the study area 1000 times and evaluate the counting threshold for all iterations, we can estimate that 45 would be the average number of individuals necessary as counting threshold. This counting threshold is much larger than the 8 individuals required when we have information from the whole dataset. Furthermore, it is comforting that the pilot study tends to propose a number of individuals much larger than the minimum number required if the survey-wide data was considered.

Using all the carabid assemblage data, the probability to count too few individuals after sampling 6 sites can be calculated. From the example presented above, we know that to reach a 0.95 RV coefficient between partial and complete-abundance, a counting threshold of 8 individuals is needed. For a pilot study to underestimate the counting threshold, all six sites in the pilot study need to have

fewer than 8 individuals per species. In the actual carabid assemblage data, 35 sites out of 196 have fewer than 8 individuals. The probability of randomly selecting six of these sites is (35/196) × (34/195) × (33/194) × (32/193) × (31/192) × (30/191) or 0.000022. If any of the other 161 sites is considered in the pilot study, the number of individuals estimated by our procedure will be equal or higher than 8, thus reaching at least a survey-wide 0.95 RV coefficient.

DISCUSSION

Counting partial-abundance data is a cost-effective approach for sampling ecological communities. Because of its flexibility, the approach proposed in this paper makes it possible for researchers to decide the accuracy they want to have in the data they collect and then to reduce the sampling and identification effort to achieve this accuracy. In addition to saving cost, this would help overcome the taxonomic impediment to biodiversity studies of arthropods.

In our simulations, we showed that it was possible to estimate a counting threshold by using as few as 3% of the SUs. Although our results lead us to believe that, up to a certain extent, a larger pilot study points to a smaller counting threhold, it is left at the discretion of the researcher to consider a larger number of SUs in the pilot study. However, for surveys where the number of SUs to be sampled is small, a pilot study should include a minimum of 5 SUs to ensure that the chance of sampling too few individuals is low.

Pilot studies are at the core of the procedure we proposed in this paper. If in the pilot study the counting threshold calculated seems too low, considering more SUs in the pilot study should improve the estimation of the counting

threshold. Because the information gained in the pilot study may be included in the full study, the cost of considering additional SUs in the pilot study is not as important as it would be if it was carried out independent from the survey-wide study. Moreover, if the SUs considered in the pilot study present a surprising low number of individuals, they should be quick to count compared to SUs with larger abundances, making the effort to include new SUs in the pilot study less of a constraint.

Counting partial-abundance data in a pilot study may be easy or difficult depending on the survey design and the organisms sampled. For example, if individuals are collected in the field and brought to a laboratory for sorting, identification, and counting, it is easy to reevaluate the counting threshold with a minimum of effort because all pilot study SUs are readily available. Insects, mites, and spiders are typical examples of organisms that allow such flexibility because they are usually sample in traps, and when sorting and identification is carried out all traps are easily accessible. Conversely, if organisms need to be recorded in the field (e.g., trees or birds), a pilot study would need to be carried out before the full-scale survey begins. However, as explained in the previous paragraph, the time spent on the pilot study is not usually lost because the data collected while carrying out the pilot study can often be included in the final dataset. Moreover, as we have shown, the pilot study will make it possible to be much more cost-effective when surveying.

In the context of this procedure, community data can be organized in two distinct groups: (1) Where the sampling is carried out blindly (e.g., fish in a lake, birds in a forest, mites in soil cores) or (2) where the number of individuals within

a SU can be coarsely assessed (e.g., trees in a forests, carabids in pitfall traps). When sampling is carried out blindly, we have to rely on the knowledge gained from the pilot study to estimate a counting threshold. As we have shown, this is by no mean constraining because it is highly unlikely that one will consistently sample SUs with very low abundance.

If coarse evaluation of the number of individuals can be made, the SUs used to carry out a pilot study do not need to be randomly selected; in fact, they can be chosen by considering the ones with large numbers of individuals. This would prevent researchers from estimating a counting threshold that is too low.

We have also shown that the distance used to analyze the data can have tremendous impact on the number of sites to consider in a pilot study. In that instance, it becomes important to choose the distance with which all our analysis will be carried out before sampling the community. For simple and canonical ordinations, it is common for researcher to be ambivalent about the choice of dissimilarity to use. This confusion is justified at least for canonical ordinations where the differences for using one distance over another are usually minor (Blanchet et al. unpublished). With the result found here, we show that the modified Gower distance using a base 2 logarithm is very efficient. Although the modified Gower distance has been given more emphasis in this paper, it does not mean that the other distances should not be used. However, the counting threshold should be expected to be higher than when the modified Gower distance is used.

The counting procedure we propose in this paper applies to a broad range of ecological surveys, but not to all of them. If each individual needs to be handled separately for taxonomic identification, it is irrelevant to use the proposed

counting approach because all individuals sampled would need to be manipulated anyways. For example, the counting procedure cannot be applied to zooplankton collected in vials because each animal needs to be handled individually for identification. Note, however, that our counting procedure can be used for vegetation quadrats, where percentage of cover can be readily estimated, as well as for any group of organisms where individuals can be easily identified and counted.

In this paper, we showed that using prior information from a pilot study to evaluate community patterns can be useful to increase cost-effectiveness while minimizing the loss of information. The proposed counting procedure has the potential to be applied to numerous types of studies within and outside the scope of ecology. In ecology, it can be valuable for large-scale monitoring studies such as the Alberta Biodiviersity Monitoring Institute project (Boutin et al. 2009). In some studies involving organisms whose sizes differ greatly, abundance data may be transformed to biomass data. It is possible to evaluate a counting threshold on biomass data using the approach proposed in this paper. Our approach is also applicable to landscape genetics where gene (or marker, etc.) frequencies in local populations are used instead of species frequencies. Since the lab work is costly in genetic studies, our approach could lead to important savings of technician time and materials. Although our counting approach has been presented mainly in the context of terrestrial surveys, it can be applied as well to aquatic communities. Outside the scope of ecology, it can be applied to any situations where many count variables are measured for a series of SUs.

The conclusions reached in this paper rely on simulated data. Although we tried to make these simulations as general as possible, we knew it was impossible to simulate all possible cases found in nature (Milligan 1996).

LITERATURE CITED

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.

Anderson, R. M. 1965. Methods of collecting and preserving vertebrate animals. Department of the Secretary of State.

Baddeley, A., and R. Turner. 2005. Spatstat: an R package for analyzing spatial point patterns. Journal of Statistical Software **12**:1–42.

Barton D. E., and F. N. Davis. 1956. Some notes on ordered random intervals. Journal of the Royal Statistical Society. Series B (Methodological) **18**:79–94.

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

Bergeron, J. A. C., J. R. Spence, and W. J. A. Volney. 2011. Landscape patterns of species-level association between ground-beetles 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.

Boutin, S., D. L. Haughland, J. Schieck, J. Herbers, and E. Bayne. 2009. A new approach to forest biodiversity monitoring in Canada. Forest Ecology and Management **258**:S168–S175.

Cailliez, F. 1983. The analytical solution of the additive constant problem. Psychometrika **48**:305–308.

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

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

Faith, D., P. Minchin, and L. Belbin. 1987. Compositional dissimilarity as a robust measure of ecological distance. Vegetatio **69**:57–68.

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

Greenacre, M. 2007. Correspondence analysis in practice. Chapman & Hall.

Illian, J., A. Penttinen, H. Stoyan, and D. Stoyan. 2008. Statistical analysis and modelling of spatial point patterns. Wiley.

Lebart, L., and F. Jean-Pierre. 1971. Statistique et Informatique Appliquées. Dunod.

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. 2012. Numerical Ecology. 3rd English edition. Elsevier, Amsterdam.

MacArthur, R. H. 1957. On the relative abundance of bird species. Proceedings of the National Academy of Sciences of the United States of America **43**:293–295.

Martin, J. E. H. 1977. The Insects and Arachnids of Canada Part 1. Agriculture Canada.

Milligan, G. W., 1996. Clustering validation: results and implications for applied analyses. Pages 341–375 in P. Arabie, L. J. Hubert, and G. De Soete, editors. Clustering and Classification. World Scientific.

Odum, E. P. 1950. Bird populations of the Highlands (North Carolina) plateau in relation to plant succession and avian invasion. Ecology **31**:587– 605.

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

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

Preston, F. W. 1948. The commonness, and rarity, of species. Ecology **29**:254–283.

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

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

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

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

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

TABLE 4.1: Fictitious example illustrating the counting procedure proposed in this paper.

	Complete abundance				
	Species A	Species B	Species C	Species D	Species E
Sampling unit 1	0	2	10	100	900
Sampling unit 2	500	100	9	0	3
	After reaching a counting threshold of 8 individuals				
	Species A	Species B	Species C	Species D	Species E
Sampling unit 1	0	2	8	8	8
Sampling unit 2	8	8	8	0	3

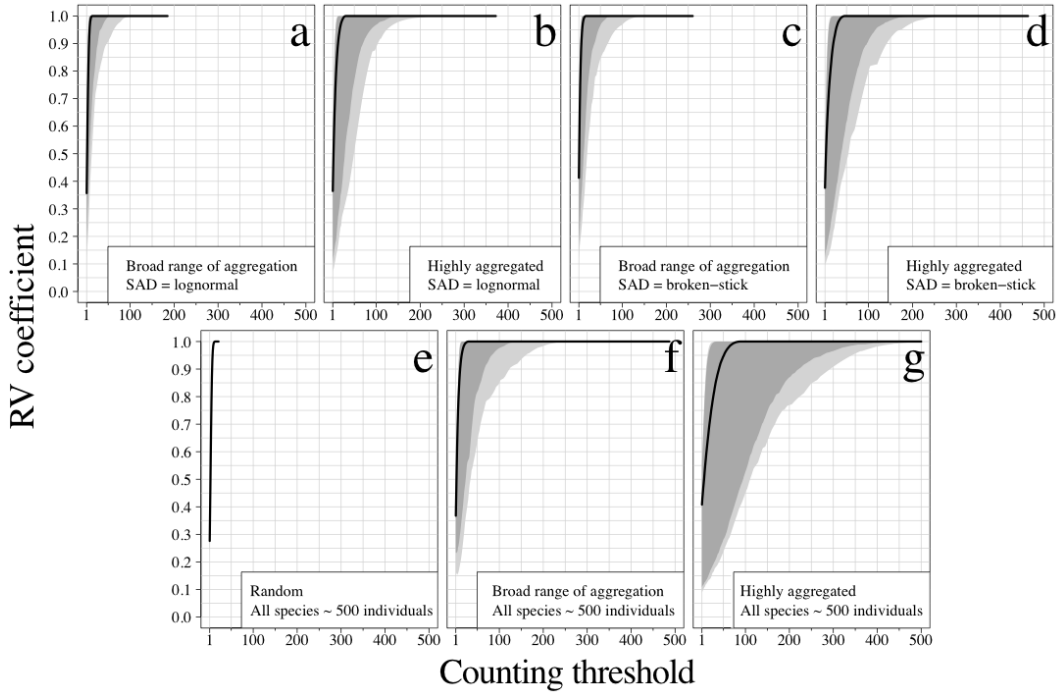

FIGURE 4.1. Simulation results of the multivariate correlation (using RV coefficient) between increasingly precise partial-abundance community data and the complete-abundance community data using raw data. The counting threshold (abscissa) is the maximum number of individuals counted for a species within a sampling unit. The ordinate represents the RV coefficients. Each panel presents the results of a set of simulated communities. Light grey areas represent the 99% empirical confidence intervals of the simulation results (constructed using the 5th and 995th largest RV coefficients associated with each increasingly precise partial-abundance data), the dark grey areas the 95% empirical confidence intervals (constructed using the 25th and 975th largest RV coefficients associated with each increasingly precise partial-abundance data), and the black lines are the medians per count threshold value.

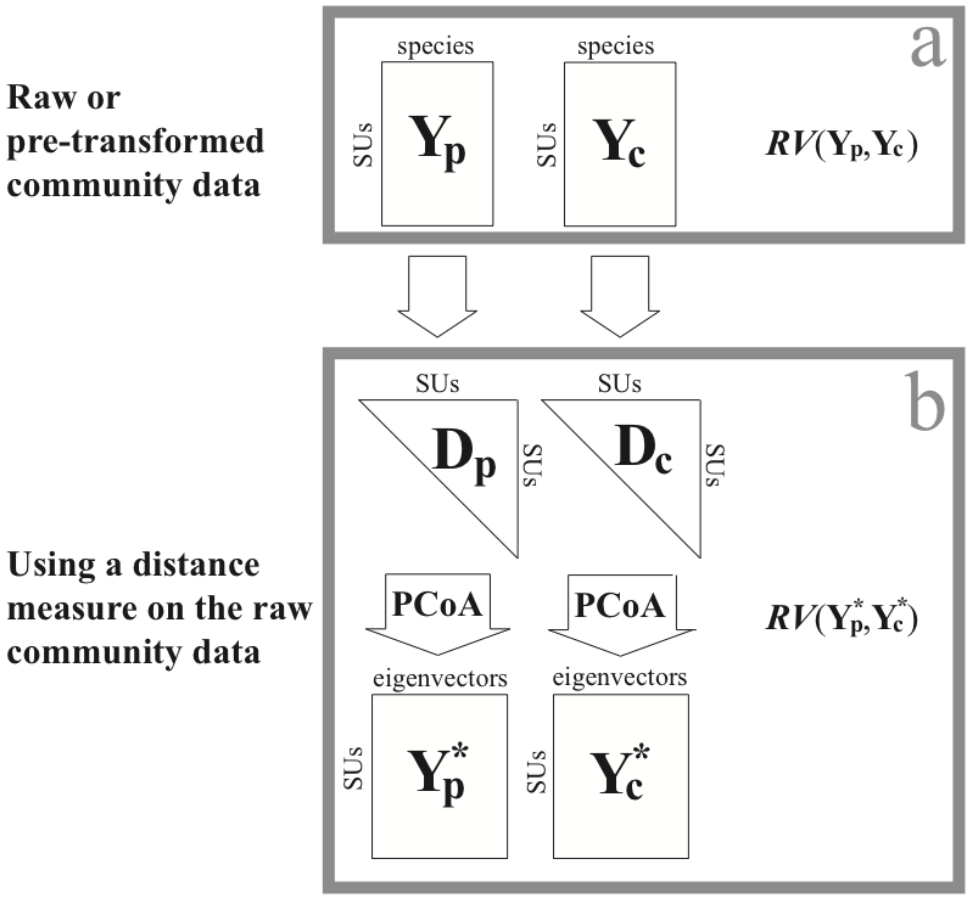

FIGURE 4.2. Explanation of the procedure used to correlate (using the RV-coefficient) partial ($\mathbf{Y_p}$) and complete-abundance ($\mathbf{Y_c}$) community matrices using different distances. (a) If the distance can be applied to the community data using pre-transformation, or if the raw data are used, partial and complete-abundance community matrices can be correlated after the pre-transformation is carried out. (b) If the distance cannot be applied to the community data using pre-transformation, symmetric distance matrices must be computed for partial ($\mathbf{D_p}$) and complete-abundance ($\mathbf{D_c}$) data. Because the RV-coefficient can only be used to correlate rectangular matrices with the same number of rows, a principal coordinates analysis (PCoA) is calculated on the symmetric distance matrices to obtain rectangular matrices $\mathbf{Y_p}^*$ and $\mathbf{Y_c}^*$ where the number of rows equals the number of SUs and the columns are eigenvectors. It then becomes possible to use the RV-coefficient to correlate $\mathbf{Y_p}^*$ and $\mathbf{Y_c}^*$.

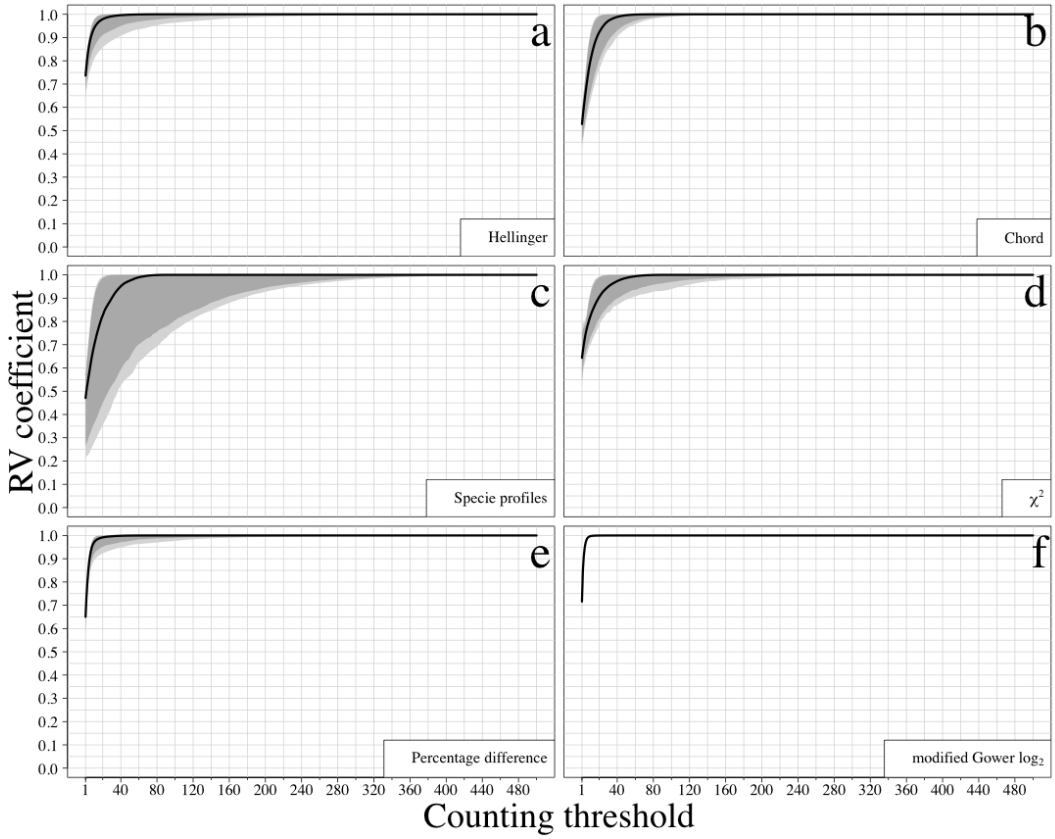

FIGURE 4.3. Simulation results of the multivariate correlation (using RV coefficient) between increasingly precise partial-abundance community data and the complete-abundance community data. All species in the simulated communities were composed of ~500 individuals and species were highly aggregated in the sampling area. The counting threshold (abscissa) is the maximum number of individuals counted for a species within a sampling unit. The ordinate represents the RV coefficients. Each panel presents the results for one distance measure. Light grey areas represent the 99% empirical confidence intervals of the simulation results (constructed using the 5th and 995th largest RV coefficients associated with each increasingly precise partial-abundance data), the dark grey areas the 95% empirical confidence intervals (constructed using the 25th and 975th largest RV coefficients associated with each increasingly precise partial-abundance data), and the black lines are the medians per count threshold value.

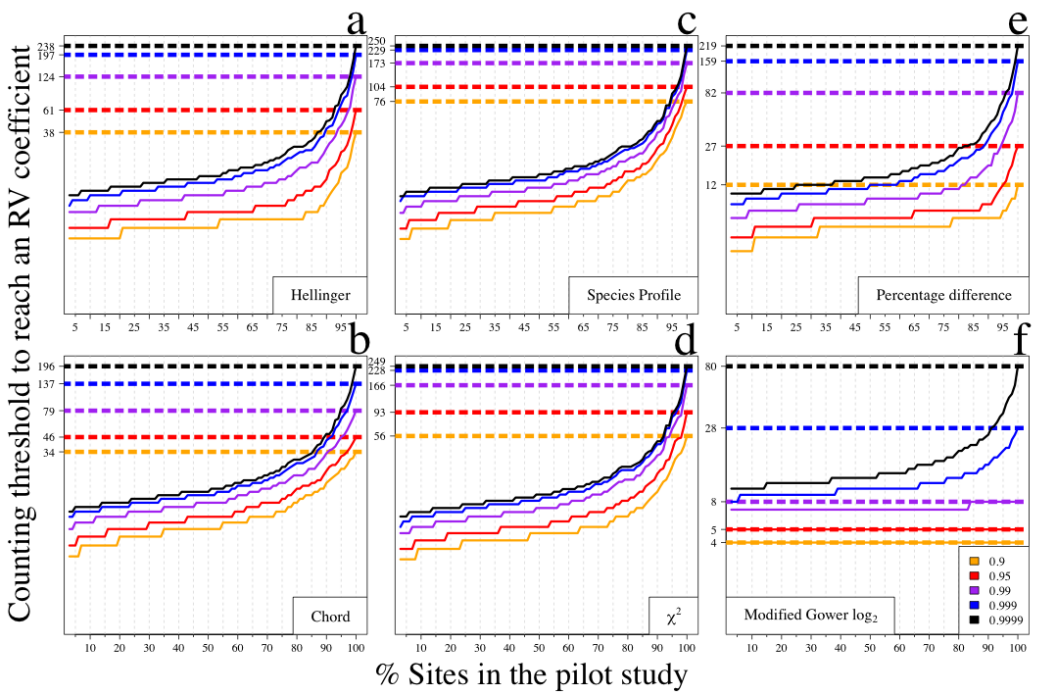

FIGURE 4.4. Percentage of sites required in a pilot study to accurately estimate the number of individuals that needs to be counted when sampling partial-abundances. In this figure, we focus on the differences between six distances for 0.9, 0.95, 0.99, 0.999, and 0.9999 RV coefficient calculated between partial and complete-abundance to be met. The species in the simulated communities were all composed of ~500 individuals and the range of their spatial aggregation level was broad. The survey-wide RV coefficients are represented by dotted lines. They are the lower bounds of the 99% confidence intervals of the simulations results presented in Figure 4.2. The full lines represent the RV coefficient between partial and complete-abundance calculated using pilot studies data. To obtain the pilot study RV coefficient, the sampling units were ordered to form an abundance gradient, from the ones that contained the lowest maximum counts of individual for any one species to the ones that presented the largest counts. Following this order, the sites were sequentially included in the pilot study. The values on the ordinates represent the number of individuals that need to be counted to reach an RV coefficient between partial and complete-abundance data. The ordinates were log-transformed for visual clarity. The counting threshold is the maximum number of individuals counted for a species within a sampling unit.

APPENDIX 4A

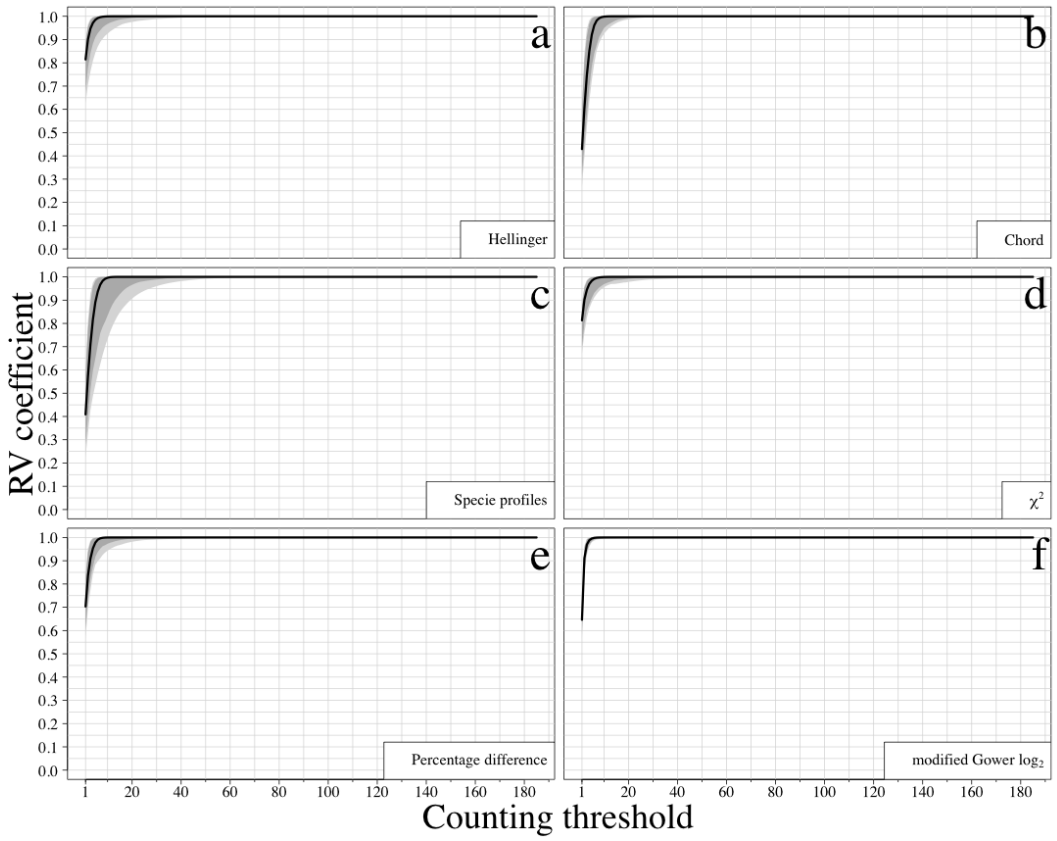

FIGURE 4A1. Simulation results of the multivariate correlation (using RV coefficient) between increasingly precise partial-abundance community data and the complete-abundance community data. The species abundance distribution of the simulated communities used to obtain this result follow a lognormal distribution and the range of aggregation of the species in the sampling area was broad. The counting threshold (abscissa) is the maximum number of individuals counted for a species within a sampling unit. The ordinate represents the RV coefficient. Each panel presents the results of one distance. Light grey areas represent the 99% empirical confidence intervals of the simulation results (constructed using the 5th and 995th largest RV coefficients associated with each increasingly precise partial-abundance data), the dark grey areas the 95% empirical confidence intervals (constructed using the 25th and 975th largest RV coefficients associated with each increasingly precise partial-abundance data), and the black lines are the medians per count threshold value.

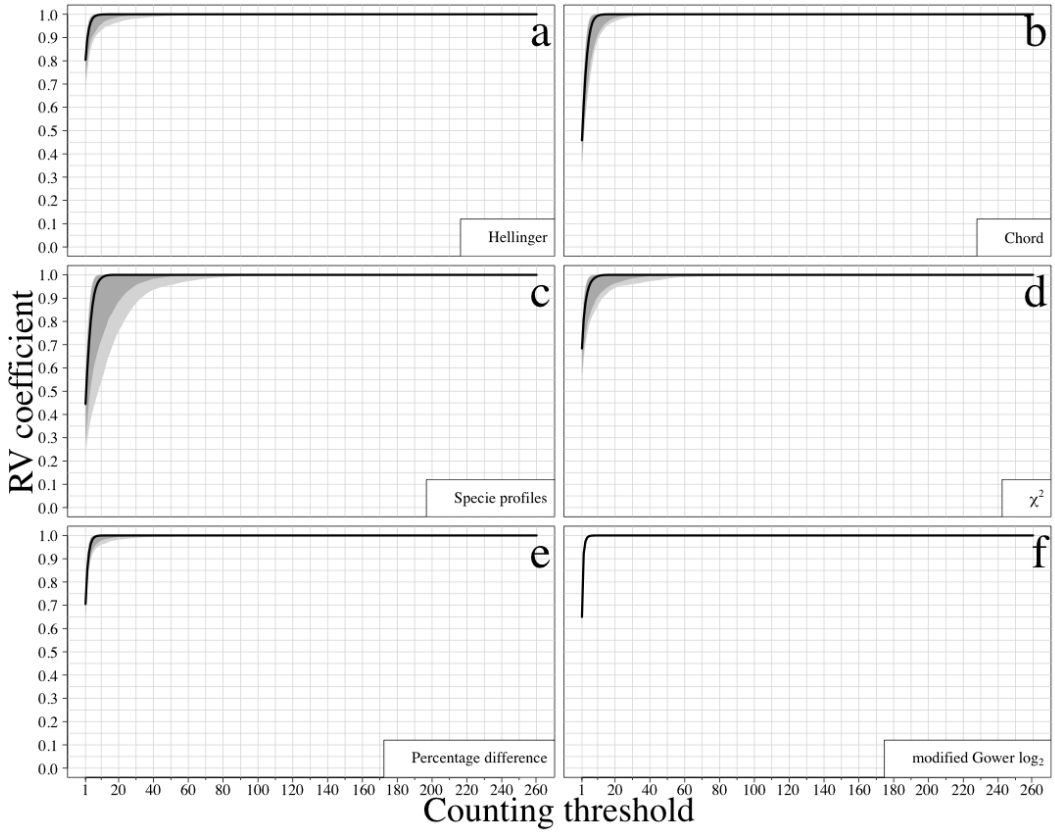

FIGURE 4A2. Simulation results of the multivariate correlation (using RV coefficient) between increasingly precise partial-abundance community data and the complete-abundance community data. The species abundance distribution of the simulated communities used to obtain this result follow a broken stick model and the range of aggregation of the species in the sampling area was broad. The counting threshold (abscissa) is the maximum number of individuals counted for a species within a sampling unit. The ordinate represents the RV coefficient. Each panel presents the results of one distance. Light grey areas represent the 99% empirical confidence intervals of the simulation results (constructed using the 5th and 995th largest RV coefficients associated with each increasingly precise partial-abundance data), the dark grey areas the 95% empirical confidence intervals (constructed using the 25th and 975th largest RV coefficients associated with each increasingly precise partial-abundance data), and the black lines are the medians per count threshold value.

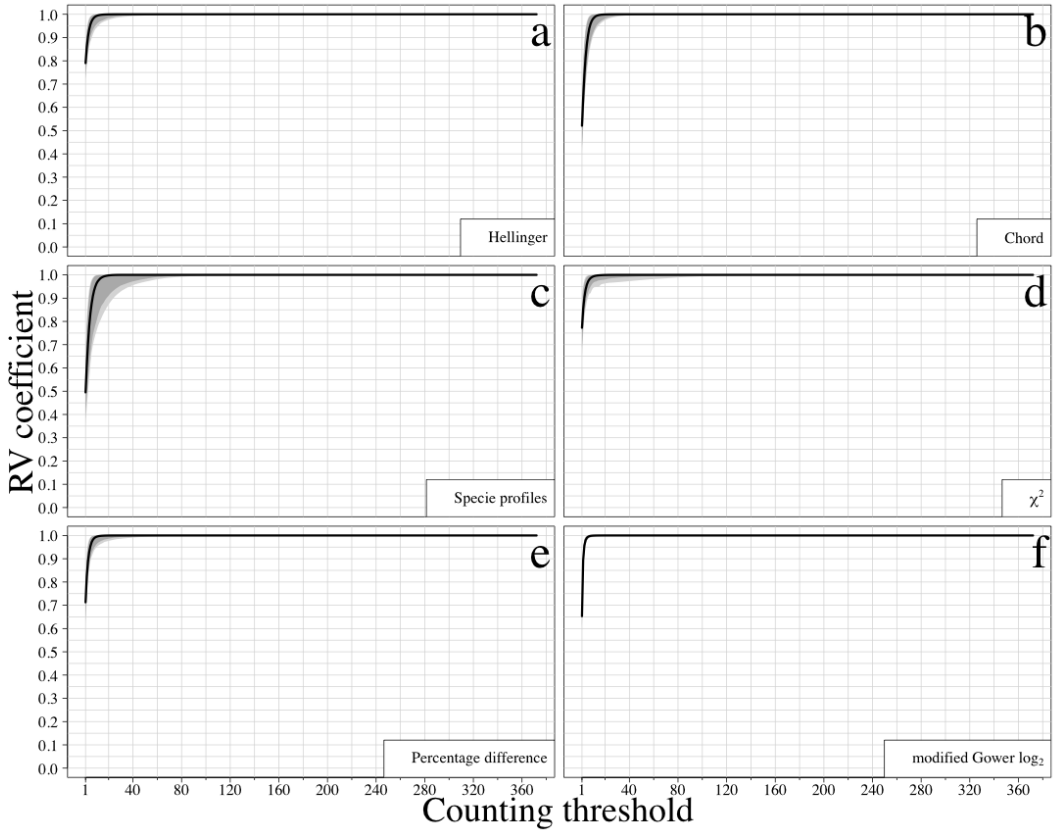

FIGURE 4A3. Simulation results of the multivariate correlation (using RV coefficient) between increasingly precise partial-abundance community data and the complete-abundance community data. The species abundance distribution of the simulated communities used to obtain this result follow a lognormal distribution and species were highly aggregated. The counting threshold (abscissa) is the maximum number of individuals counted for a species within a sampling unit. The ordinate represents the RV coefficient. Each panel presents the results of one distance. Light grey areas represent the 99% empirical confidence intervals of the simulation results (constructed using the 5^{th} and 995^{th} largest RV coefficients associated with each increasingly precise partial-abundance data), the dark grey areas the 95% empirical confidence intervals (constructed using the 25^{th} and 975^{th} largest RV coefficients associated with each increasingly precise partial-abundance data), and the black lines are the medians per count threshold value.

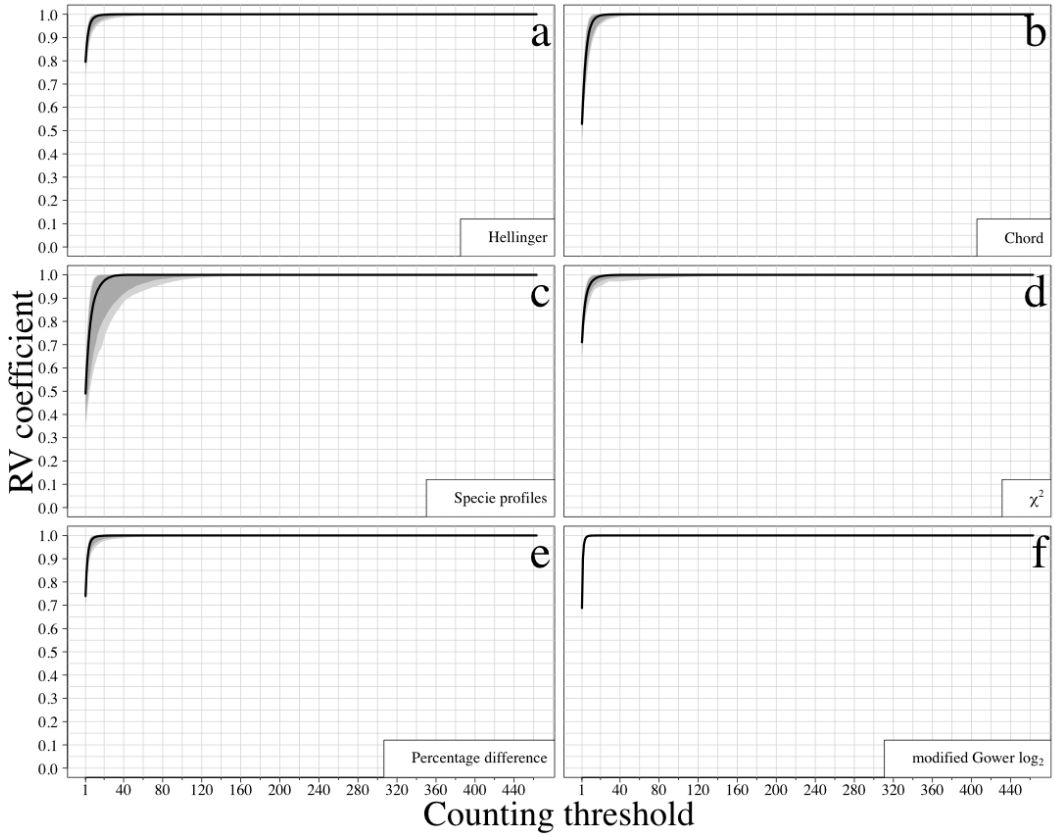

FIGURE 4A4. Simulation results of the multivariate correlation (using RV coefficient) between increasingly precise partial-abundance community data and the complete-abundance community data. The species abundance distribution of the simulated communities used to obtain this result follow a broken-stick model and species were highly aggregated. The counting threshold (abscissa) is the maximum number of individuals counted for a species within a sampling unit. The ordinate represents the RV coefficient. Each panel presents the results of one distance. Light grey areas represent the 99% empirical confidence intervals of the simulation results (constructed using the 5^{th} and 995^{th} largest RV coefficients associated with each increasingly precise partial-abundance data), the dark grey areas the 95% empirical confidence intervals (constructed using the 25^{th} and 975^{th} largest RV coefficients associated with each increasingly precise partial-abundance data), and the black lines are the medians per count threshold value.

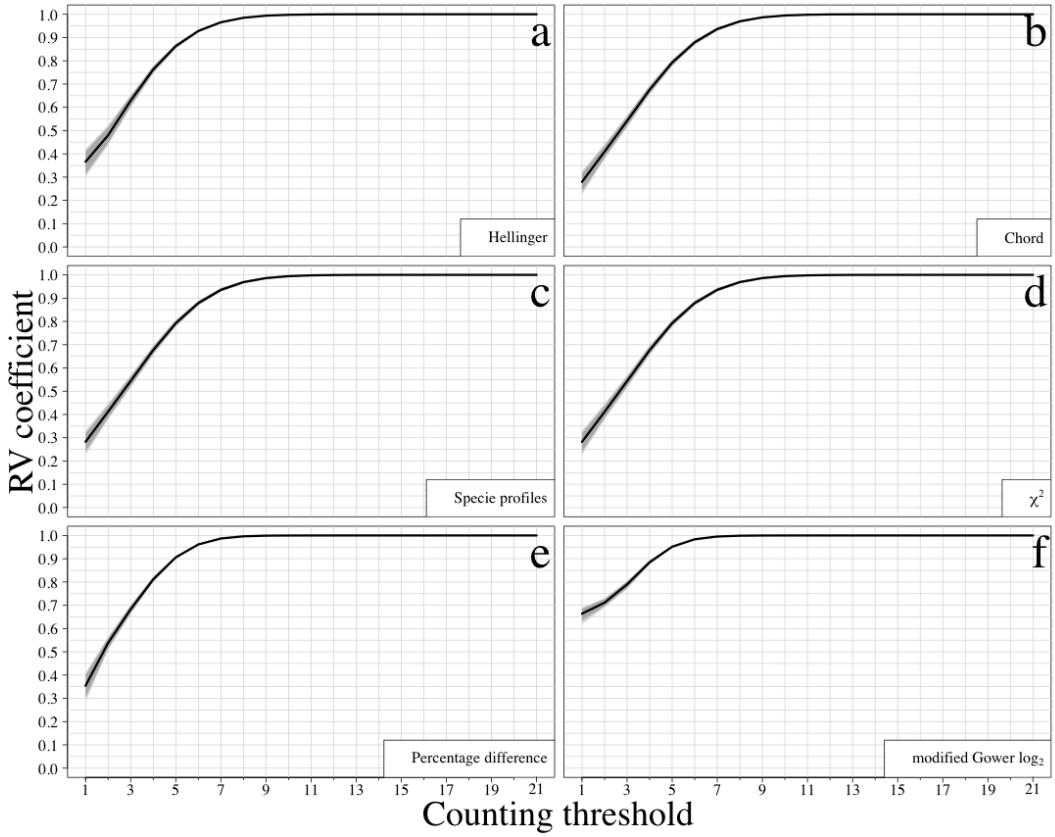

FIGURE 4A5. Simulation results of the multivariate correlation (using RV coefficient) between increasingly precise partial-abundance community data and the complete-abundance community data. All species in the simulated communities were composed of ~ 500 individuals and species were randomly distributed in the sampling area. The counting threshold (abscissa) is the maximum number of individuals counted for a species within a sampling unit. The ordinate represents the RV coefficient. Each panel presents the results of one distance. Light grey areas represent the 99% empirical confidence intervals of the simulation results (constructed using the 5^{th} and 995^{th} largest RV coefficients associated with each increasingly precise partial-abundance data), the dark grey areas the 95% empirical confidence intervals (constructed using the 25^{th} and 975^{th} largest RV coefficients associated with each increasingly precise partial-abundance data), and the black lines are the medians per count threshold value.

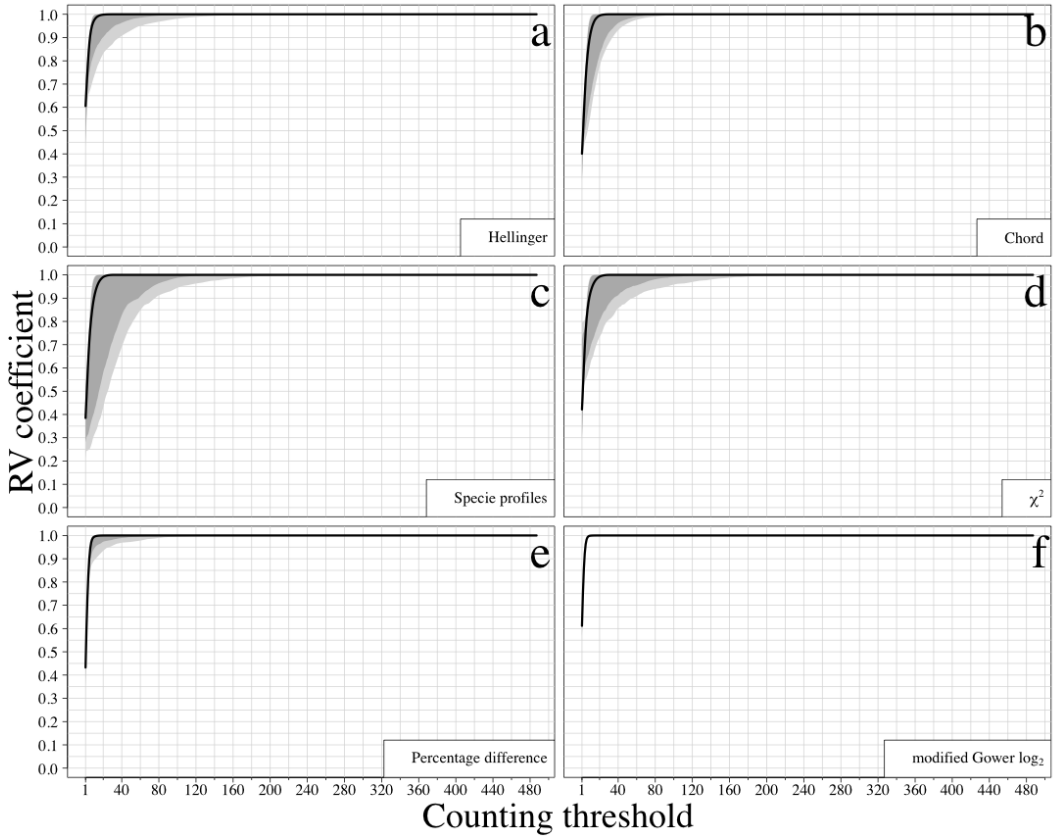

FIGURE 4A6. Simulation results of the multivariate correlation (using RV coefficient) between increasingly precise partial-abundance community data and the complete-abundance community data. All species in the simulated communities were composed of ~ 500 individuals and the range of aggregation of the species in the sampling area was broad. The counting threshold (abscissa) is the maximum number of individuals counted for a species within a sampling unit. The ordinate represents the RV coefficient. Each panel presents the results of one distance. Light grey areas represent the 99% empirical confidence intervals of the simulation results (constructed using the 5^{th} and 995^{th} largest RV coefficients associated with each increasingly precise partial-abundance data), the dark grey areas the 95% empirical confidence intervals (constructed using the 25^{th} and 975^{th} largest RV coefficients associated with each increasingly precise partial-abundance data), and the black lines are the medians per count threshold value.

APPENDIX 4B

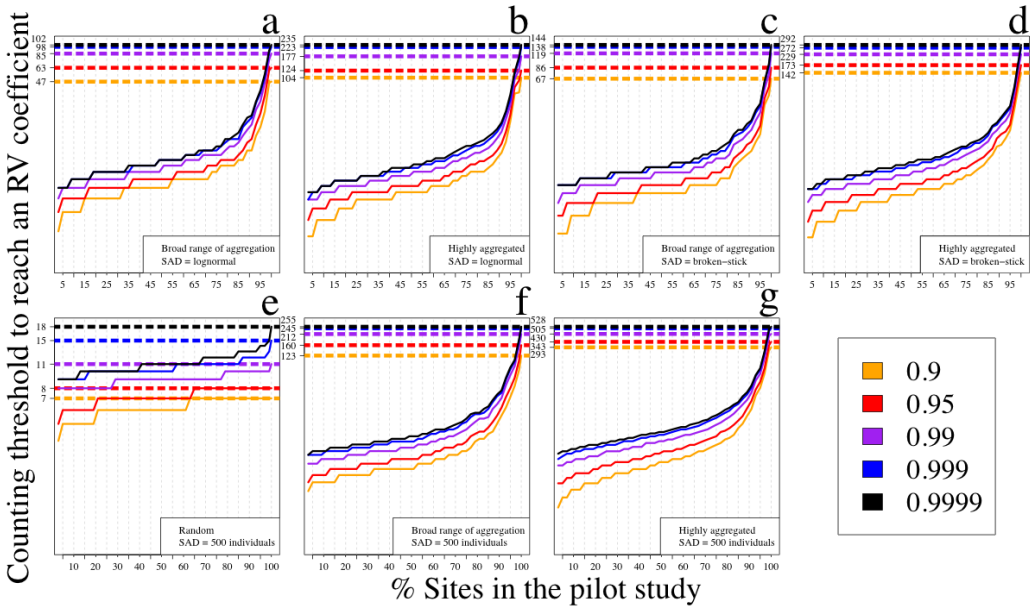

FIGURE 4B1. Percentage of sites required in a pilot study to accurately estimate the number of individuals that needs to be counted when sampling partial-abundances. In this figure, we focus on the differences between the seven types of simulated communities for 0.9, 0.95, 0.99, 0.999, and 0.9999 RV coefficient calculated between partial and complete-abundance to be met when the Euclidean distance is used. The survey-wide RV coefficients are represented by dotted lines. They are the lower bounds of the 99% confidence intervals of the simulations results presented in Figure 4.2. The full lines represent the RV coefficient between partial and complete-abundance calculated using pilot studies data. To obtain the pilot study RV coefficient, the sampling units were ordered to form an abundance gradient, from the ones that contained the lowest maximum counts of individual for any one species to the ones that presented the largest counts. Following this order, the sites were sequentially included in the pilot study. The values on the ordinates represent the number of individuals that need to be counted to reach an RV coefficient between partial and complete-abundance data. The ordinates were log-transformed for visual clarity. The counting threshold is the maximum number of individuals counted for a species within a sampling unit.

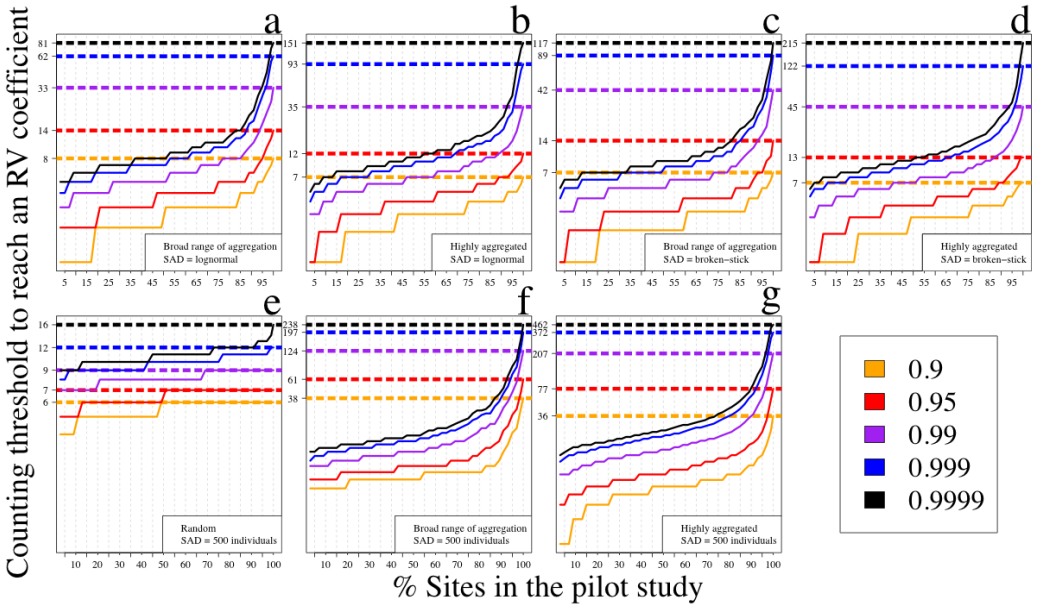

FIGURE 4B2. Percentage of sites required in a pilot study to accurately estimate the number of individuals that needs to be counted when sampling partial-abundances. In this figure, we focus on the differences between the seven types of simulated communities for 0.9, 0.95, 0.99, 0.999, and 0.9999 RV coefficient calculated between partial and complete-abundance to be met when the Hellinger distance is used. The survey-wide RV coefficients are represented by dotted lines. They are the lower bounds of the 99% confidence intervals of the simulations results presented in Figure 4.2. The full lines define the RV coefficient between partial and complete-abundance calculated using pilot studies data. To obtain the pilot study RV coefficient, the sampling units were ordered to form an abundance gradient, from the ones that contained the lowest maximum counts of individual for any one species to the ones that presented the largest counts. Following this order, the sites were sequentially included in the pilot study. The values on the ordinates represent the number of individuals that need to be counted to reach an RV coefficient between partial and complete-abundance data. The ordinates were log-transformed for visual clarity. The counting threshold is the maximum number of individuals counted for a species within a sampling unit.

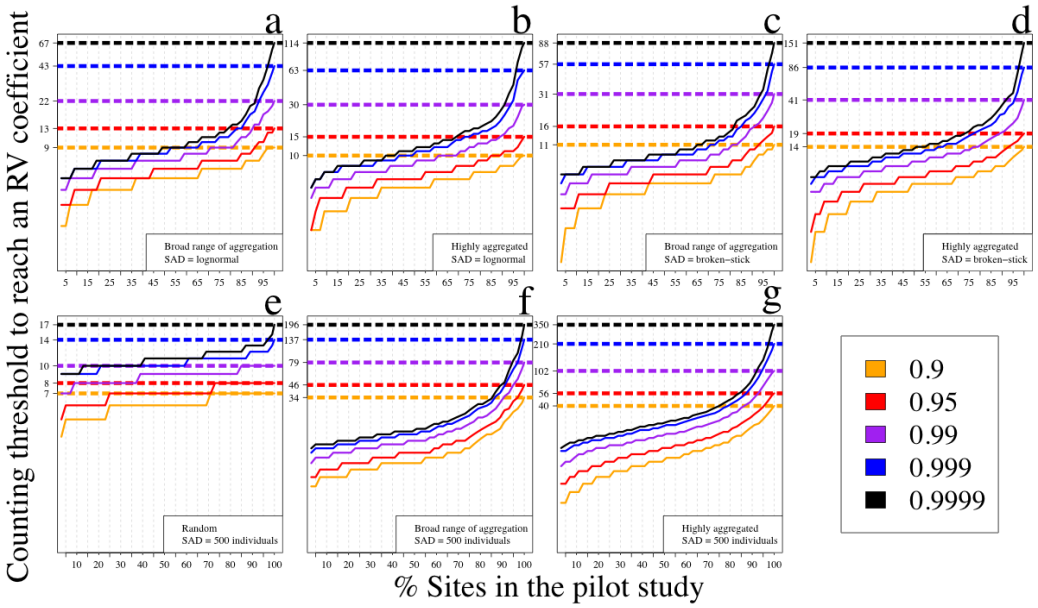

FIGURE 4B3. Percentage of sites required in a pilot study to accurately estimate the number of individuals that needs to be counted when sampling partial-abundances. In this figure, we focus on the differences between the seven types of simulated communities for 0.9, 0.95, 0.99, 0.999, and 0.9999 RV coefficient calculated between partial and complete-abundance to be met when the chord distance is used. The survey-wide RV coefficients are represented by dotted lines. They are the lower bounds of the 99% confidence intervals of the simulations results presented in Figure 4.2. The full lines define the RV coefficient between partial and complete-abundance calculated using pilot studies data. To obtain the pilot study RV coefficient, the sampling units were ordered to form an abundance gradient, from the ones that contained the lowest maximum counts of individual for any one species to the ones that presented the largest counts. Following this order, the sites were sequentially included in the pilot study. The values on the ordinates represent the number of individuals that need to be counted to reach an RV coefficient between partial and complete-abundance data. The ordinates were log-transformed for visual clarity. The counting threshold is the maximum number of individuals counted for a species within a sampling unit.

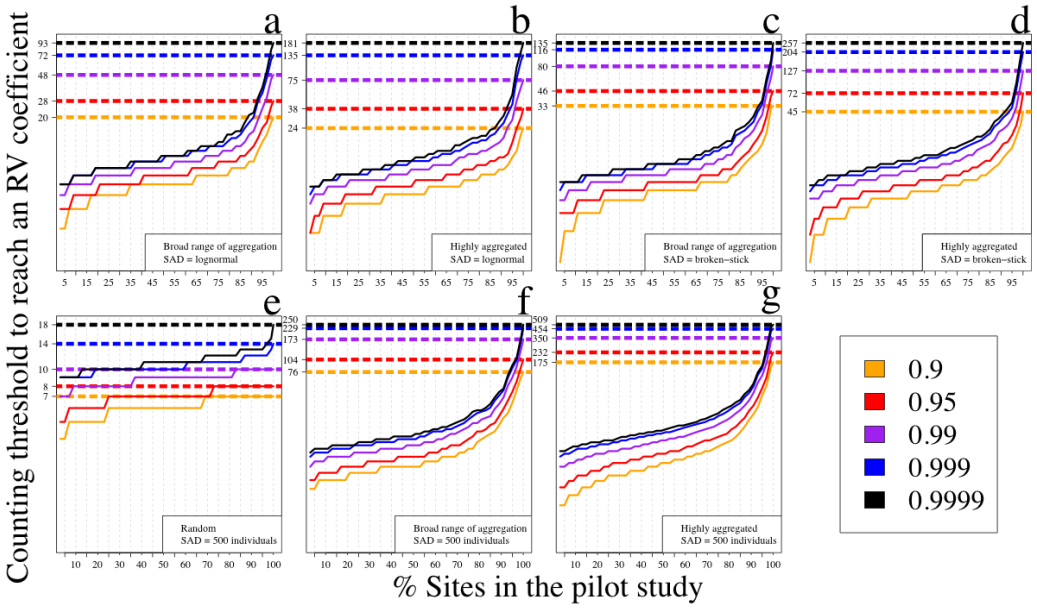

FIGURE 4B4. Percentage of sites required in a pilot study to accurately estimate the number of individuals that needs to be counted when sampling partial-abundances. In this figure, we focus on the differences between the seven types of simulated communities for 0.9, 0.95, 0.99, 0.999, and 0.9999 RV coefficient calculated between partial and complete-abundance to be met when the distance between species profile is used. The survey-wide RV coefficients are represented by dotted lines. They are the lower bounds of the 99% confidence intervals of the simulations results presented in Figure 4.2. The full lines define the RV coefficient between partial and complete-abundance calculated using pilot studies data. To obtain the pilot study RV coefficient, the sampling units were ordered to form an abundance gradient, from the ones that contained the lowest maximum counts of individual for any one species to the ones that presented the largest counts. Following this order, the sites were sequentially included in the pilot study. The values on the ordinates represent the number of individuals that need to be counted to reach an RV coefficient between partial and complete-abundance data. The ordinates were log-transformed for visual clarity. The counting threshold is the maximum number of individuals counted for a species within a sampling unit.

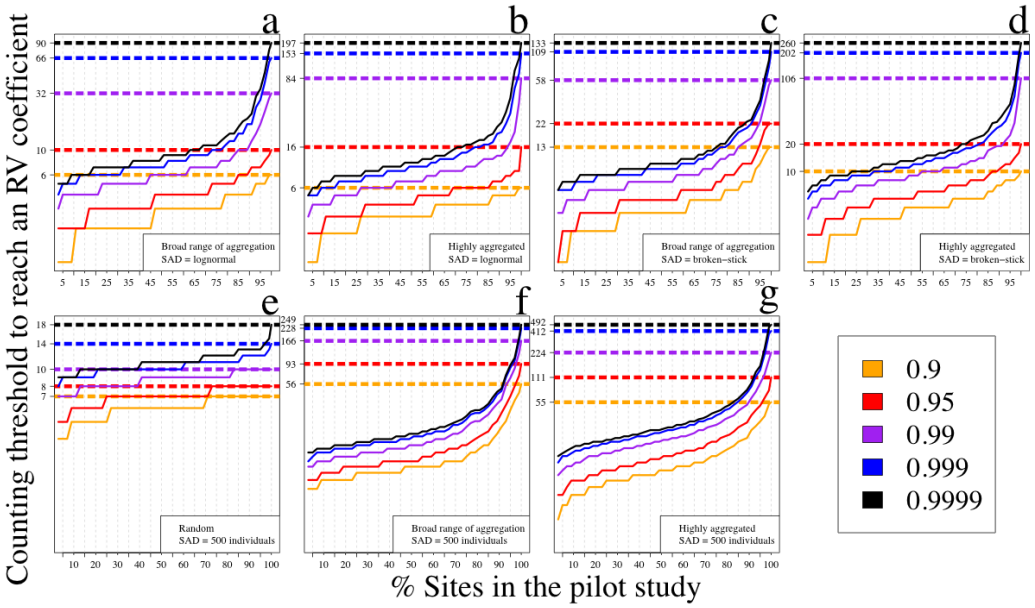

FIGURE 4B5. Percentage of sites required in a pilot study to accurately estimate the number of individuals that needs to be counted when sampling partial-abundances. In this figure, we focus on the differences between the seven types of simulated communities for 0.9, 0.95, 0.99, 0.999, and 0.9999 RV coefficient calculated between partial and complete-abundance to be met when the χ^2 distance is used. The survey-wide RV coefficients are represented by dotted lines. They are the lower bounds of the 99% confidence intervals of the simulations results presented in Figure 4.2. The full lines define the RV coefficient between partial and complete-abundance calculated using pilot studies data. To obtain the pilot study RV coefficient, the sampling units were ordered to form an abundance gradient, from the ones that contained the lowest maximum counts of individual for any one species to the ones that presented the largest counts. Following this order, the sites were sequentially included in the pilot study. The values on the ordinates represent the number of individuals that need to be counted to reach an RV coefficient between partial and complete-abundance data. The ordinates were log-transformed for visual clarity. The counting threshold is the maximum number of individuals counted for a species within a sampling unit.

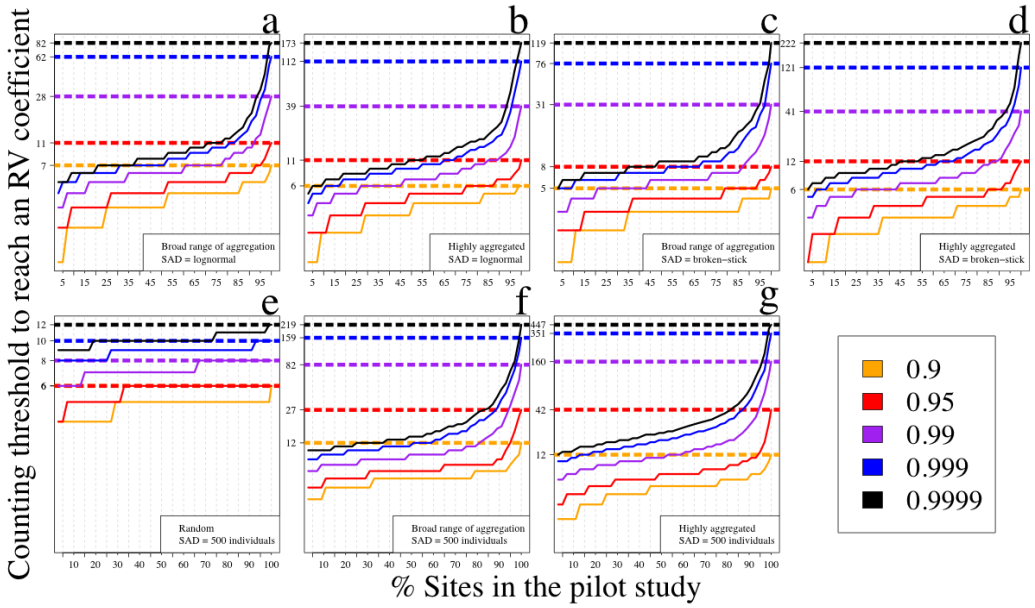

FIGURE 4B6. Percentage of sites required in a pilot study to accurately estimate the number of individuals that needs to be counted when sampling partial-abundances. In this figure, we focus on the differences between the seven types of simulated communities for 0.9, 0.95, 0.99, 0.999, and 0.9999 RV coefficient calculated between partial and complete-abundance to be met when the percentage difference distance is used. The survey-wide RV coefficients are represented by dotted lines. They are the lower bounds of the 99% confidence intervals of the simulations results presented in Figure 4.2. The full lines define the RV coefficient between partial and complete-abundance calculated using pilot studies data. To obtain the pilot study RV coefficient, the sampling units were ordered to form an abundance gradient, from the ones that contained the lowest maximum counts of individual for any one species to the ones that presented the largest counts. Following this order, the sites were sequentially included in the pilot study. The values on the ordinates represent the number of individuals that need to be counted to reach an RV coefficient between partial and complete-abundance data. The ordinates were log-transformed for visual clarity. The counting threshold is the maximum number of individuals counted for a species within a sampling unit.

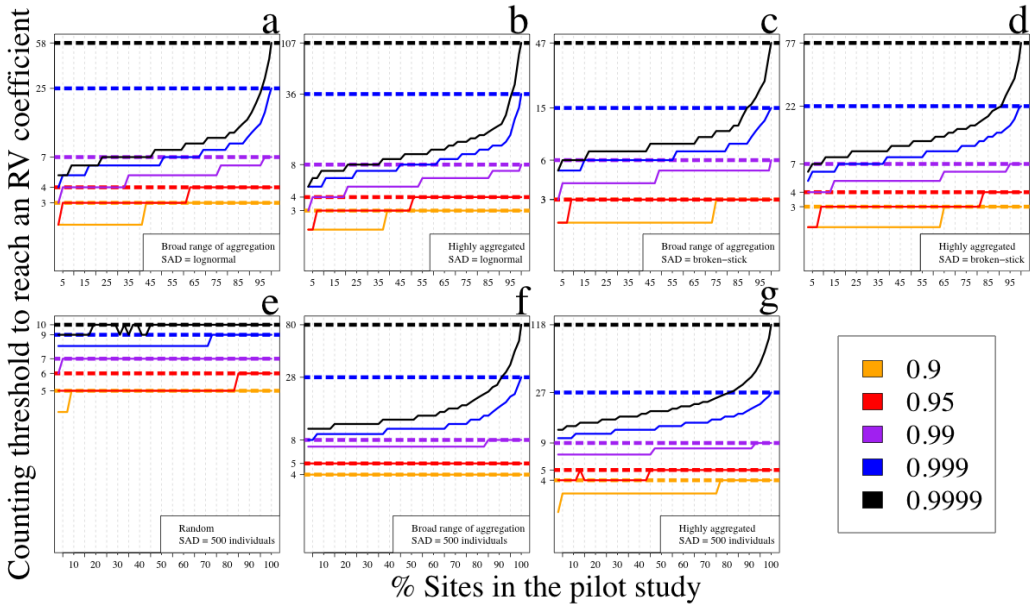

FIGURE 4B7. Percentage of sites required in a pilot study to accurately estimate the number of individuals that needs to be counted when sampling partial-abundances. In this figure, we focus on the differences between the seven types of simulated communities for 0.9, 0.95, 0.99, 0.999, and 0.9999 RV coefficient calculated between partial and complete-abundance to be met when the modified Gower distance calculated with a log base 2 is used. The survey-wide RV coefficients are represented by dotted lines. They are the lower bounds of the 99% confidence intervals of the simulations results presented in Figure 4.2. The full lines define the RV coefficient between partial and complete-abundance calculated using pilot studies data. To obtain the pilot study RV coefficient, the sampling units were ordered to form an abundance gradient, from the ones that contained the lowest maximum counts of individual for any one species to the ones that presented the largest counts. Following this order, the sites were sequentially included in the pilot study. The values on the ordinates represent the number of individuals that need to be counted to reach an RV coefficient between partial and complete-abundance data. The ordinates were log-transformed for visual clarity. The counting threshold is the maximum number of individuals counted for a species within a sampling unit.

Conclusion

This thesis presents a detailed study of the use of space ad habitat by boreal forest Carabidae (Chapter 2) and two methodological developments (Chapter 3 and 4) to improve the use of resemblance measures in analysis of ecological communities. In chapter 2, I use a broad scale study of Carabidae to more fully understand how anthropogenic disturbances, habitat heterogeneity, and spatial autocorrelation influence the distribution of these insects. In chapter 3, I present a new approach to canonical ordination that makes a consensus of RDAs performed on the same data but with different resemblance measures. These resemblance measures capture similar patterns of a dataset when used in a canonical analysis and a consensus amplifies the most constant patterns. In chapter 4, I propose a new procedure that allows researchers to obtain the same conclusions from a study but using only a fraction of the data that would otherwise be required.

The second chapter of this thesis revealed that forest floor cover, soil drainage, and tree composition were the main factors explaining the boreal forest ground beetle assemblages in northwestern Alberta. It also showed that roads, seismic lines, and forest harvest viewed at the landscape scale had little impact on the spatial distribution of Carabidae. Although effects were more notable at the stand scale, the results provided encouragements and showed that it is possible to manage boreal forest in the presence of anthropogenic disturbances in a way that conserves insect biodiversity. At the landscape scale, however, it appears that

presence of large interconnected patches of mature boreal forest is important to the maintenance of ground beetle diversity. This research focuses on broad scale patterns of ground beetles. This study says little about how these insects are spatially structured at a fine scale. However, together with the literature focusing on carabid communities at a scale of less than a few hectares (for reviews see Thiele 1977 and Lövei and Sunderland 1996), the results obtained can be useful to generate novel hypotheses about the factors that influence the distribution of carabids across scales.

The third chapter showed through extensive simulations that the species abundance distribution (SAD) of a community cannot serve as a tool to select a resemblance measure. I could however show that most resemblance measures are equivalent and the performance of the measures is not affected by variations in SADs. These conclusions are valid for both abundance and presence-absence data. This result prompted the development of a mathematical method where a consensus among RDAs is constructed across resemblance measures. This new approach was developed because resemblance measures may influence the interpretation of community data. A consensus of RDAs across a group of resemblance measures has the advantage to prevent the bias imposed by resemblance measure from occurring. By using this approach, more emphasis is given to the commonality among resemblance measure in term of the information on sites, species, and explanatory variables that they provide. In this chapter, I also show that when a community is composed of many rare species, abundance data brings little new information about community structure not already given by presence-absence data. Based on these findings, I propose to use species presence

distribution to evaluate the importance of the information lost by sampling presence-absence. The research presented in this chapter gives only rudimentary insights about SAD comparisons and suggests that the information lost from converting the multivariate community data into SADs is important. However, because this chapter focuses only on the SADs presented in Figure 3.1, further work needs to be carried out to develop a robust theory to compare all SADs, as proposed by McGill et al. (2007).

The fourth chapter of this thesis showed that patterns defining a species community (spatial, environmental, and/or others) can be effectively elaborated even when not all individuals of a community are counted. This result suggests it is possible to estimate a counting threshold, the maximum number of individuals per species that need to be counted within a sampling unit, to extract a minimum set of data that would sufficiently represent the true community. When this counting threshold is reached, the community patterns revealed will present very high correlation with the community where all individuals are counted. We show that when this counting approach is used in association with a resemblance measure, it becomes much more efficient and has the potential to be applied to almost any type of multivariate count datasets - ecological or other. The new counting method proposed in this chapter will increase cost-efficiency when sampling is done to construct multivariate datasets. However, this new counting approach is not useful in situations where all individuals need to be independently examined to identify them. Also, although this approach is well adapted to multivariate datasets, this chapter does not show if the same counting approach could be generalized to univariate data.

The conclusion reached and the analytical development proposed in chapters 3 and 4 are based on simulated data. Although all efforts were put in place to make these simulations as general as possible, it is impossible to simulate all types of ecological data (Milligan 1996). Further empirical tests of the developed methods are warranted.

LITERATURE CITED

Lövei, G. L., and K. D. Sunderland. 1996. Ecology and behavior of ground beetles (Coleoptera: Carabidae). Annual Review of Entomology **41**:231–256.

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

Milligan, G. W., 1996. Clustering validation: results and implications for applied analyses. Pages 341–375 in P. Arabie, L. J. Hubert, and G. De Soete, editors. Clustering and Classification. World Scientific.

Thiele, H.-U. 1977. Carabid Beetles in their Environments. Springer-Verlag, Berlin.