Submitted to the *Journal of Plant Ecology*

Studying beta diversity: ecological variation partitioning by multiple regression and canonical analysis

Pierre Legendre

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

E-mail: Pierre.Legendre@umontreal.ca

Abstract. — Beta diversity is the variation in species composition among sites in a geographic region. Beta diversity is a key concept for understanding the functioning of ecosystems, for the conservation of biodiversity, and for ecosystem management. It can be studied by computing diversity indices for each site and testing hypotheses about the factors that may explain the variation among sites. It can also be studied by direct analysis of the community composition data table over the study sites, as a function of sets of environmental and spatial variables. These analyses are carried out by the statistical method of partitioning the variation of the diversity indices or the community composition data table with respect to environmental and spatial variables. Variation partitioning is briefly described in this paper. Partitioning must be based upon unbiased estimates of the variation of the community composition data table that is explained by the various tables of explanatory variables. The adjusted coefficient of determination provides such an unbiased estimate in both multiple regression and canonical redundancy analysis. After partitioning, one can test the significance of the fractions of interest and plot maps of the fitted values corresponding to these fractions.

INTRODUCTION

Ecologists collect community composition data (species presence-absence or abundance data) at several sites in a region of interest in order to analyze and interpret beta diversity, which is the variation in species composition among the sites (Whittaker 1960, 1972; Legendre et al. 2005). Analysis of a synthetic descriptor such as species richness or Shannon diversity can be done by multiple regression, whereas the analysis of whole community composition data tables is carried out by canonical analysis. Results from these two types of analyses are not equivalent: analysis of the whole community composition data produces results that are much more informative since they provide information about the reactions of individual species to the environmental and spatial variables. The asymmetrical forms of canonical analysis used for this type of research are canonical redundancy analysis (RDA, Rao 1964) and canonical correspondence analysis (CCA, ter Braak 1986, 1987a, 1987b). These analyses are described in several textbooks, including Legendre and Legendre (1998), and available in computer packages such as Canoco (ter Braak and Smilauer 2002) and the 'vegan' library (Oksanen et al. 2006) of the R statistical language (R Development Core Team 2006).

Variation in species composition among sites is studied by canonical analysis of the species composition data as a function of different types of environmental variables: water or soil chemistry, geology, geomorphology, environmental impact descriptors, and so on. The study of spatial structures involves spatial variables derived from the geographic coordinates of the sampling sites, described below. Variation partitioning is a technique of choice for this type of analysis. In all cases, statistics are used to describe how successful the explanatory variables are at explaining the response variables (community composition data). The choice of an appropriate, unbiased statistical estimator is of great importance for the correct interpretation of the results. This paper will briefly describe partial linear regression and

canonical analysis, the simple and adjusted forms of the coefficient of determination used in regression and canonical analysis, and finally variation partitioning.

PARTIAL LINEAR REGRESSION

The notation **y~X|W** represents the partial linear regression of a response variable **y** (vector of length *n*) on a matrix **X** containing *m* explanatory variables, while controlling for the linear effect of a matrix **W** containing *q* covariables. Partial regression is computed in two steps: (1) regress **X** on **W** and compute the residuals $\mathbf{X}_{\mathrm{res}(\mathbf{W})}$. (2) Regress **y** on $\mathbf{X}_{\mathrm{res}(\mathbf{W})}$ to obtain the partial R^2, the fitted values, the residuals, and so on.

The R^2 statistic of a partial regression that will be used to construct the *F*-statistic for the test of significance (next paragraph) is called the partial R^2. It is the ratio of the sum-of-squares (SS) of the fitted values of the partial regression on the sum (SS of the fitted values + SS of the residuals):

$$R^2_{\mathbf{y}\sim\mathbf{X}|\mathbf{W}} = \mathrm{SS}(\text{fitted values of } \mathbf{y}\sim\mathbf{X}|\mathbf{W}) \,/\, (\mathrm{SS}(\text{fitted values}) + \mathrm{SS}(\text{residuals})) \quad (1)$$

Using the graphical representation of Fig. 1, $R^2_{\mathbf{y}\sim\mathbf{X}|\mathbf{W}}$ = [a] / [a+d].

The *F*-statistic used to test the significance of the partial regression relationship takes into account the number of covariables *q*; in ordinary multiple regression, $q = 0$. The *F*-statistic is computed as follows using the partial R^2:

$$F = (R^2_{\mathbf{y}\sim\mathbf{X}|\mathbf{W}} / m) \,/\, ((1-R^2_{\mathbf{y}\sim\mathbf{X}|\mathbf{W}})/(n-1-m-q)) \quad (2)$$

It can also be computed directly from the sums-of-squares:

$$F = (\mathrm{SS}(\text{fitted values of } \mathbf{y}\sim\mathbf{X}|\mathbf{W}) / m) \,/\, ((\mathrm{SS}(\text{residuals}))/(n-1-m-q)) \quad (3)$$

or, using Fig. 1:

$$F = ([\mathrm{a}]/m) \,/\, ([\mathrm{d}]/\,(n-1-m-q)) \quad (4)$$

Significance of the *F*-statistic can be tested with reference to an *F*-distribution if the condition of normality of the residuals is met (this is rarely the case for ecological data), or by a permutation test if they are not (this is the most common case). Permutation tests are described in several textbooks, including Manly (1997) and Legendre and Legendre (1998). In the application to variation partitioning described below, both **y~X|W** and **y~W|X** will be computed and tested for significance.

PARTIAL CANONICAL ANALYSIS

Similarly, the notation **Y~X|W** represents the partial canonical redundancy analysis (partial RDA) of a response data matrix **Y** of size (*n*×*p*) on a matrix **X** containing *m* explanatory variables, while controlling for the linear effect of a matrix **W** containing *q* covariables. Partial canonical analysis is computed in the same way as partial linear regression and uses the same *F*-statistic for significance testing; see below for details. In the application to variation partitioning described below, both **Y~X|W** and **Y~W|X** will be computed and tested for significance.

UNADJUSTED AND ADJUSTED COEFFICIENTS OF DETERMINATION

The coefficient of multiple determination (unadjusted R^2) estimates the forecasting potential of a multiple regression equation:

$$R^2 = \frac{\text{regression SS}}{\text{total SS}} = \frac{\Sigma (\hat{y}_i - \bar{y})^2}{\Sigma (y_i - \bar{y})^2} = 1 - \frac{\text{residual SS}}{\text{total SS}} \qquad (5)$$

where “regression SS” is the sum-of-squares of the fitted values of the regression equation. It measures the proportion of the variation of **y** about its mean that is explained by the regression equation.

In multiple regression, an alternative measure of determination is the adjusted coefficient of multiple determination R_a^2 (Ezekiel 1930):

$$R_a^2 = 1 - \frac{\text{residual mean square}}{\text{total mean square}} = 1 - (1 - R^2)\left(\frac{\text{total d.f.}}{\text{residual d.f.}}\right) \quad (6)$$

The right-hand parenthesis of equation 6 shows that R_a^2 takes into account the numbers of degrees of freedom associated with the numerator and denominator of equation 5. In ordinary multiple regression, the total degrees of freedom of the F-statistic are (n–1) and the degrees of freedom of the residuals are (n–m–1) where n is the number of observations and m is the number of explanatory variables in the model. In multiple regression through the origin, where the intercept is forced to zero, the total degrees of freedom of the F-statistic are n and the residual degrees of freedom are (n–m). R_a^2 is a suitable measure of goodness-of-fit for comparing regression equations fitted to different data sets, with different numbers of objects and explanatory variables. Using simulated data with normal error, Ohtani (2000) has shown that R_a^2 is an unbiased estimator of the contribution of a set of explanatory variables **X** to the explanation of **y**. The R_a^2 statistic cannot be directly computed for partial linear regression because the number of degrees of freedom to use in the correction is unknown.

In canonical redundancy analysis (RDA), the canonical R^2 is called the bimultivariate redundancy statistic (Miller and Farr 1971) or the canonical coefficient of determination. It is computed in the same way as in multiple regression: it is the ratio of the sum of each response variable's regression (or fitted values) SS to the sum of all response variables' total SS. In canonical analysis, the significance of the F-statistic is always tested by permutation, except in the very restrictive case where the variables in **Y** are standardized and the residuals are multi-normal. These conditions are almost never met with ecological data; in the rare cases where they are, the F-statistic is tested using the Fisher-Snedecor F-distribution with (m×p) and p(n–m–1) degrees of freedom (Miller 1975). Using numerical simulations, Peres-Neto et

al. (2006) have shown that, for normally distributed data or Hellinger-transformed species abundances in RDA, the adjusted bimultivariate redundancy statistic R_a^2, obtained by applying equation 6 to the canonical R^2, produced unbiased estimates of the real contributions of the variables in **X** to the explanation of a response matrix **Y**. The Hellinger transformation is one of five transformations that make community composition data containing many zeros suitable for analysis by linear methods such as principal component analysis (PCA) or canonical redundancy analysis (RDA) (Legendre and Gallagher 2001).

Adjusted coefficients of determination in multiple regression and canonical analysis can, on occasion, take negative values. For large data sets, R_a^2 is zero when the explanatory variables explain no more variation than random normal variables would. Negative values of R_a^2 are interpreted as zeros; they correspond to cases where the explanatory variables explain less variation than random normal variables would.

VARIATION PARTITIONING

The technique of variation partitioning is used when two or more complementary sets of hypotheses can be invoked to explain the variation of an ecological variable. For example, the abundance of a species could vary as a function of biotic and abiotic factors. In the study of beta diversity, the community composition data table can be partitioned among one or more sets of environmental variables and a table describing the spatial relationships among the sampling sites. Fitting the community composition data to spatial variables, as described below, allows researchers to establish that there are significant spatial patterns, perhaps at various scales, present in the species data. The presence of significant spatial patterns in the response data can be invoked as support for either a neutral model (Bell 2001, Hubbell 2001, He 2005) or for environmental control since environmental data are often spatially structured. The presence of significant relationships between the species and environmental variables

would strongly support the hypothesis of environmental control, which is not in opposition with a hypothesis of neutral process, as discussed by Legendre et al. (2005).

Variation partitioning among environmental and spatial components was first described by Borcard et al. (1992) and Borcard and Legendre (1994). Variation partitioning will be presented in the context of the analysis of a response community composition data table **Y**. It can be applied as well to a single response variable **y** since the algebra of partial linear regression is the same as that of partial canonical analysis.

Variation partitioning of a response data table **Y** with respect to two matrices of explanatory variables **X** and **W** involves the following three steps, which correspond to different research objectives.

1. Obtaining the fractions of variation. — The calculations, based upon three multiple regressions (for a single variable **y**) or three canonical analyses (for a multivariate response table **Y**), are summarized in Table 1:

• Compute the canonical analysis of **Y** with respect to the first table of explanatory variables **X**. Compute the R^2 and R_a^2 using equations 5 and 6. Assuming that the rectangle has a surface area normalized to 1, the R_a^2 corresponds to the surface area of the left-hand circle in Fig. 1. It contains the adjusted fractions [a] and [b].

• Compute the canonical analysis of **Y** with respect to the second table of explanatory variables **W**. Compute the R^2 and R_a^2 using equations 5 and 6. The R_a^2 corresponds to the surface area of the right-hand circle in Fig. 1. It contains the adjusted fractions [b] and [c].

• Compute the canonical analysis of **Y** with respect to the union of tables **X** and **W**. Compute the R^2 and R_a^2 using equations 5 and 6. The R_a^2 corresponds to the union of the two circles in Fig. 1. It contains the adjusted fractions [a], [b] and [c].

• From these first results, compute fraction [b] by subtraction: [b] = [a+b] + [b+c] – [a+b+c].

• Compute fraction [a] by subtraction: [a] = [a+b] – [b]

• Compute fraction [c] by subtraction: [c] = [b+c] – [b]

• Compute fraction [d], which represents the residual variation, by subtraction: [d] = 1 – [a+b+c].

These values can be added to a Venn diagram such as the one shown in Fig. 1. Because they are based on adjusted coefficients of determination, the fractions can, on occasion, take negative values. These are interpreted as zeros, as explained in the previous section of this paper.

When **X** is a matrix of environmental variables and **W** contains descriptors of the spatial relationships among the sampling sites, the Venn diagram (Fig. 1) provides the following information:

• The circle containing [a+b] shows how much of the variation of **Y** is explained by the environmental variables. Of that, [b] is the variation explained jointly by **X** and **W**, or the fraction of the environmentally-explained variation that is spatially structured. [a] is the environmentally-explained variation that is not explained by the spatial variables found in **W**.

• The circle containing [b+c] shows how much of the variation of **Y** is explained by the spatial variables found in **W**. Of that, [c] is the variation explained uniquely by a linear model of the spatial variables found in **W** and not by a linear effect of the environmental variables **X**. This component may be due to spatially-structured environmental variables that are not present in table **X** or to nonlinear effects of the environmental variables **X** on **Y**. That variation may also be due to processes, such as competition or dispersal, in the ecological community depicted by table **Y**. In that case, it cannot be related to environmental variables.

To model broad-scale spatial patterns only, Borcard et al. (1992) and Borcard and Legendre (1994) used a third-degree polynomial function of the geographic coordinates of the sampling sites as matrix **W** in variation partitioning. More recently, Borcard and Legendre (2002) and Borcard et al. (2004) described PCNM analysis, which generates a matrix **W** containing spatial descriptors that represent a spectral decomposition of the spatial relationships among the sampling sites. PCNM analysis allows researchers to model these relationships at all spatial scales. That form of analysis can also be called "distance-based eigenvector maps" (DBEM) or Moran's eigenvector maps (MEM) (Dray et al. 2006).

2. Testing the significance of the fractions. — The fractions must be tested for significance in order to fully support the reasoning described in section (1). The *F*-statistics of the three regressions or canonical analyses giving rise to the adjusted fractions [a+b], [b+c], and [a+b+c] (Table 1) can be tested directly by parametric or permutation tests. Individual fractions [a] and [c] cannot be tested in that way (next paragraph), while fraction [b] cannot be tested at all, as shown in Table 1. [d] is the residual variation. Fraction [d], together with its degrees of freedom, forms the denominator of the *F*-statistics used in testing the other fractions.

The partial canonical analyses **Y~X|W** and **Y~W|X** have to be computed to test the significance of fractions [a] and [c], respectively. The *F*-statistics are computed following equation 2, 3, or 4. These *F*-statistics are tested using special permutation methods, called "permutation of the residuals", described in Legendre and Legendre (1998) and Anderson and Legendre (1999).

3. Mapping the fitted values of the fractions. — The fitted values corresponding to fractions [a+b], [b+c], [a+b+c], [a], and [c] can be computed in order to draw maps that will help in interpreting them. In the case of a single response variable **y**, the fitted values of the

multiple and partial multiple regressions giving rise to these fractions provide the values that can be mapped. In the case of a multivariate response table **Y**, e.g. a community composition table, the fitted values are contained in multivariate tables of site scores produced by the canonical and partial canonical analyses. The first few axes of each of these tables, which correspond to the largest canonical eigenvalues, can be used for mapping. Point maps, such as bubble plots, should be produced for fraction [a] because that fraction is not spatially structured; the map will display the "local innovation" at each sampling site. Interpolation mapping techniques, such as kriging, can be used for the other fractions, which contain spatially correlated values.

Variation partitioning of **Y** can be computed with respect to three or four tables of explanatory variables. The algebra, which involves more steps, will not be explained in detail here. It is described in one of the documentation files of the package 'vegan' (Oksanen et al. 2006) of the R statistical language.

CONCLUSION

Statistical analysis of community composition data must not be taken lightly. For proper tests of hypotheses concerning the factors responsible for the creation and maintenance of beta diversity in ecosystems, it is important to use tests of significance that do not rely on unrealistic assumptions, such as multivariate normality, when the data do not support these assumptions. Tests of significance must have correct type I error rates and good power to detect effects, natural or anthropogenic, when these effects are present. When significant effects are identified, one should use unbiased statistics (R_a^2) to report their magnitude. The conclusions reached during ecological analysis will be used by practitioners to take important decisions about the management of ecosystems, so they must be grounded in good science.

This paper described the method of variation partitioning, which took many years to develop. Variation partitioning allows researchers to test precise hypotheses about the origin of beta diversity in ecosystems and determine how much of the spatial variation is controlled by environmental variables and how much remains unexplained. The latter fraction may be under the influence of unmeasured environmental variables, or else it may be determined by community processes such as competition or dispersal that need to be explored. In any case, the use of appropriate statistics is of foremost importance during ecological variation partitioning.

ACKNOWLEDGEMENTS

The author is grateful to Daniel Borcard (Université de Montréal) and Fangliang He (University of Alberta, Edmonton) for judicious comments on a preliminary version of this paper. This research was supported by NSERC grant no. OGP0007738 to P. Legendre.

REFERENCES

Anderson, M. J. and P. Legendre. 1999. An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *Journal of Statistical Computation and Simulation* 62: 271-303.

Bell, G. 2001. Neutral macroecology. *Science* 293: 2413-2418.

Borcard, D. and P. Legendre. 1994. Environmental control and spatial structure in ecological communities: an example using oribatid mites (Acari, Oribatei). *Environmental and Ecological Statistics* 1: 37-53.

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

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

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

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.

Ezekiel, M. 1930. *Methods of correlation analysis.* John Wiley and Sons, New York.

He, F. 2005. Deriving a neutral model of species abundance from fundamental mechanisms of population dynamics. *Functional Ecology* 19: 187-193.

Hubbell, S. P. 2001. *The unified neutral theory of biodiversity and biogeography.* Princeton University Press, Princeton, NJ.

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

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

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

Legendre, P. and L. Legendre. 1998. *Numerical ecology, 2nd English edition.* Elsevier Science BV, Amsterdam.

Manly, B. J. F. 1997. *Randomization, bootstrap and Monte Carlo methods in biology. 2nd edition.* Chapman and Hall, London.

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

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

Ohtani, K. 2000. Bootstrapping R^2 and adjusted R^2 in regression analysis. *Economic Modelling* 17: 473-483.

Oksanen, J., R. Kindt, P. Legendre and R. B. O'Hara. 2006. *vegan: Community Ecology Package version 1.8-2.* URL http://cran.r-project.org/.

Peres-Neto, P., P. Legendre, S. Dray and D. Borcard. 2006. Variation partitioning of species data matrices: estimation and comparison of fractions. *Ecology* (in press).

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

R Development Core Team. 2006. *R: a language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

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. 1987a. The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio* 69: 69-77.

ter Braak, C. J. F. 1987b. Ordination. 91-173 in: R. H. G. Jongman, C. J. F. ter Braak & O. F. R. van Tongeren [eds.] *Data analysis in community and landscape ecology.* Pudoc, Wageningen, The Netherlands. Reissued in 1995 by Cambridge Univ. Press, Cambridge, England.

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

Whittaker, R. H. 1960. Vegetation of the Siskiyou mountains, Oregon and California. *Ecological Monographs* 30: 279-338.

Whittaker, R. H. 1972. Evolution and measurement of species diversity. *Taxon* 21: 213-251.

Table 1. Method for calculating the adjusted fractions of variation [a] to [d] depicted in Fig. 1. Three multiple regressions or canonical analyses are needed.

Canonical analyses	Compute R^2 (eq. 5)	Compute R_a^2 (eq. 6) and fractions of variation	Can be tested for significance
Y~X	R^2 of **Y~X**	[a+b] = R_a^2 of **Y~X**	Yes
Y~W	R^2 of **Y~W**	[b+c] = R_a^2 of **Y~W**	Yes
Y~(X,W)	R^2 of **Y~(X,W)**	[a+b+c] = R_a^2 of **Y~(X,W)**	Yes
		[a] = [a+b] – [b]	Yes
		[b] = [a+b] + [b+c] – [a+b+c]	No
		[c] = [b+c] – [b]	Yes
		Residuals = [d] = 1 – [a+b+c]	No

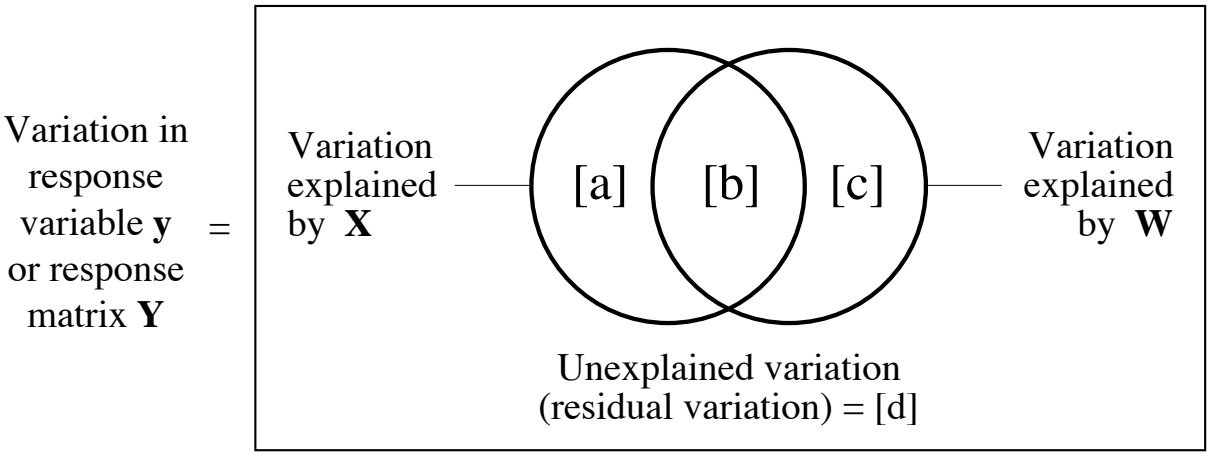

Fig. 1. Venn diagram representing the partition of the variation of a response variable **y** or a response matrix **Y** between two sets of explanatory variables **X** and **W**. The rectangle represents 100% of the variation in **y** or **Y**. Fraction [b] is the intersection (not the interaction) of the amounts of variation explained by linear models of **X** and **W**. Adapted from Legendre (1993).