Combined PDF for the Chinese and English versions of the paper

Legendre, P. 2007. Studying beta diversity: ecological variation partitioning by multiple regression and canonical analysis. *Journal of Plant Ecology (Chinese version)*, [formerly *Acta Phytoecologica Sinica*] 31: 976-981.

Legendre, P. 2008. Studying beta diversity: ecological variation partitioning by multiple regression and canonical analysis. *Journal of Plant Ecology (English version)* 1: 3-8.
[Published by Oxford University Press on behalf of the Institute of Botany, Chinese Academy of Science.]

植物生态学报　2007，31（5）976 ~ 981
Journal of Plant Ecology（*Chinese Version*）

β-多样性的研究：应用多元回归和典范分析研究生态方差的分解

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

摘　要　β-多样性刻画了地理区域中不同地点物种组成的变化，是理解生态系统功能、生物多样性保护和生态系统管理的一个重要概念。该文介绍了如何从群落组成，相关环境和空间数据角度去分析 β-多样性。β-多样性可以通过计算每个地点的多样性指数，进而对可能解释点之间差异的因子所作的假设进行检验来研究。也可以将涵盖所有点的群落组成数据表看作是一系列环境和空间变量的函数，进行直接分析。这种分析应用统计方法将多样性指数或群落组成数据表的方差进行关于环境和空间变量的分解。该文对方差分解进行阐述。方差分解是利用环境和空间变量来解释 β-多样性的一种方法。β-多样性是生态学家用来比较不同地点或同一地点不同生态群落的一种手段。方差分解就是将群落组成数据表的总方差无偏分解成由各个解释变量所决定的子方差。调整的决定系数提供了针对多元回归和典范冗余分析的无偏估计。方差分解后，可以对感兴趣的方差解释部分进行显著性检验，同时绘出基于这部分方差解释的预测图。

关键词　调整的决定系数　β-多样性　生物多样性　典范冗余分析　群落组成　方差分解

STUDYING BETA DIVERSITY：ECOLOGICAL VARIATION PARTITIONING BY MULTIPLE REGRESSION AND CANONICAL ANALYSIS

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

Abstract ***Aims*** 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. This paper describes how to analyze it from community composition and associated environmental and spatial data tables.

Methods Beta diversity can be studied by computing diversity indices for each site and testing hypotheses about the factors that may explain the variation among sites. Or，one can carry out a 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.

Important findings Variation partitioning is a method of choice for the interpretation of beta diversity using tables of environmental and spatial variables. Beta diversity is an interesting“ currency ”for ecologists to compare either different sampling areas，or different ecological communities co-occurring in an area. 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.

Key words adjusted coefficient of determination，beta diversity，biodiversity，canonical redundancy analysis，community composition，variation partitioning

收稿日期：2006-08-04　**接受日期**：2006-09-19

此文将奉献给我的导师 Sokal **教授。**Sokal **教授出生于维也纳，成长于中国上海，在上海接受了高中和大学教育，并学习了汉语。**Sokal **教授和** Sneath **教授一起创建了数量分类学。随后创建了遗传数据空间分析研究领域和相关方法。我于** 1985 **年学年休假时和** Sokal **教授深入探讨了空间遗传分析的原理和方法，受益颇丰。此文是对** Sokal **教授创建多尺度生态系统分析方法所作贡献的证明**

感谢 Borcard **博士（加拿大蒙特利尔大学）、何芳良博士（加拿大阿尔伯塔大学）对本文所提的建设性意见和建议。感谢加拿大自然科学和工程研究委员会（**NSERC**）对我的研究资助（项目号** OGP0007738**）。感谢喻梅博士（中国科学院植物研究所）、何芳良博士（加拿大阿尔伯塔大学）和米湘成博士（中国科学院植物研究所）将此文译为中文**

E-mail：Pierre.Legendre@umontreal.ca

生态学家通过收集研究区内不同地点的群落组成数据(物种出现-不出现或物种多度)来分析和解释 β-多样性，即不同地点间物种组成的变化(Whittaker，1960，1972；Legendre *et al*.，2005)。对综合指标(如物种丰富度或 Shannon 指数)的分析可以通过回归分析，而对于整个群落组成数据的分析则需要进行典范分析。这两种分析的结果是不同的：由于针对整个群落组成数据的分析提供了每个物种对于环境和空间变化的反应信息，因此这种分析得到的结果涵盖了更多的信息。这种研究中用到的非对称型典范分析是典范冗余分析(RDA)(Rao，1964)和典范对应分析(CCA)(ter Braak，1986，1987a，1987b)。一些教材中对这些分析进行了介绍(Legendre & Legendre，1998)，一些软件包中也集成了这些分析，如 Canoco 软件包(ter Braak & Smilauer，2002)和 R 统计软件包(R Development Core Team，2007)的‘vegan’库(Oksanen *et al*.，2007)。

研究不同地点间物种组成的变化可以将物种组成数据视为一系列环境变量的函数，进而进行典范分析。这些环境变量有水分或土壤化学特性、地质、地形和环境影响因子等。空间结构的研究涉及到从采样点的地理坐标衍生的空间变量。方差分解是这类分析方法中的一种。统计是用来描述解释变量可以成功解释应变量(群落组成数据)变化的程度。选取适当的无偏统计估计对正确阐述分析结果至关重要。本文将简要介绍偏线性回归和典范分析、回归和典范分析中的简单和调整的决定系数，以及方差分解。

1　偏线性回归

y ~ X|W 代表 n 维应变量 y 对矩阵 X(m 个自变量)的偏线性回归，X 含有来自矩阵 W(q 个协变量)的线性影响，|W 代表剔除 W 对 X 的影响。偏回归分为两步：(1)X 对 W 进行回归，得到 X 的残差矩阵 $X_{res(W)}$；(2)y 对 $X_{res(W)}$ 进行回归，得到偏 R^2、拟合值和残差等。

偏回归得到的 R^2(用于构建显著性检验中的 F 统计量)称为偏 R^2。它是偏回归得到的拟合值的平方和(SS)与总平方和(拟合值的平方和 + 残差平方和)之比，即：

$$R^2_{y\sim X|W} = \mathrm{SS}(y\sim X|W\ \text{拟合值})/(\mathrm{SS}(\text{拟合值}) + \mathrm{SS}(\text{残差}))\qquad(1)$$

用图 1 的标识可以表示为：$R^2_{y\sim X|W} = [a]/[a + d]$。

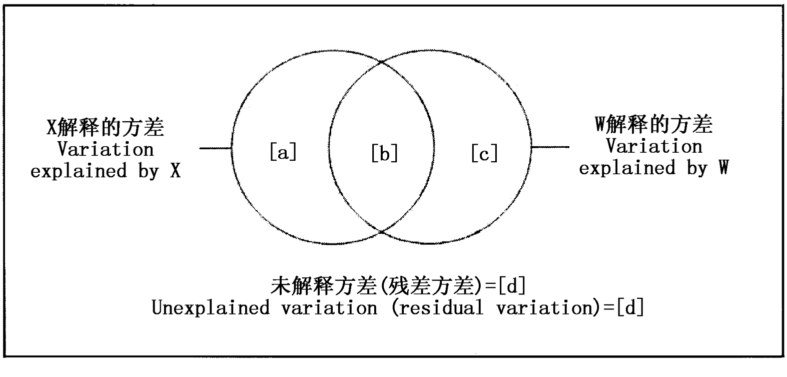

图 1　表示将对应变量 y 或应变量矩阵 Y 的方差分解成自变量矩阵 X 和 W 的 Venn 图(Legendre，1993)

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 (Legendre，1993)

长方形代表 100% y 或 Y 的方差。[b]是由线性模型 X 和线性模型 W 所解释方差的交集(不是交互作用)。改自 Legendre (1993) 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

用来检验偏回归关系显著性的 F 统计量考虑到协变量个数 q 在普通的多元回归中，$q = 0$。利用偏 R^2，F 统计量的计算公式可表示为：

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

也可以直接用平方和表示：

$$F = (\mathrm{SS}(y\sim X|W\ \text{拟合值})/m)/((\mathrm{SS}(\text{残差}))/(n - 1 - m - q))\qquad(3)$$

或用图 1 中标识，

$$F = ([a]/m)/([d]/(n - 1 - m - q))\qquad(4)$$

如果满足残差正态分布（生态数据很难满足此条件），可以利用 F 分布对 F 统计量的显著性进行检验；否则，可采用常用的置换检验（Permutation test）。一些教材中对置换检验进行了介绍，如 Manly（1997）及 Legendre 和 Legendre（1998）。下面介绍的方差分解中将涉及到 y ~ X|W 和 y ~ W|X 的计算和显著性检验。

2 偏典范分析

与前述相类似，Y ~ X|W 代表应变量矩阵 Y（$n \times p$）对含 m 个自变量的矩阵 X 的偏典范冗余分析（偏 RDA），X 含有来自矩阵 W（q 个协变量）的线性影响，|W 代表剔除 W 对 X 的影响。偏典范分析的计算与偏线性回归一样，并且也是用 F 统计量进行显著性检验。方差分解中将涉及 Y ~ X|W 和 Y ~ W|X 的计算和显著性检验。

3 未调整和调整的决定系数

多元决定系数（未调整的 R^2）估计了多元回归方程的预测潜力：

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

其中，“regression SS”是回归方程拟合值的平方和。R^2 度量了 y 相对于其均值的变化中可以被回归方程解释的比例。

多元回归中的另一种决定系数是调整的决定系数，${R_a}^2$（Ezekiel，1930）：

$${R_a}^2 = 1 - \frac{\text{residual mean square}}{\text{total mean square}} = 1 - (1 - R^2)\left(\frac{\text{total df}}{\text{residual df}}\right) \tag{6}$$

${R_a}^2$ 进一步考虑了式 5 中分子和分母的自由度。常规的多元回归中，F 统计量的总自由度为 $n - 1$，残差自由度为 $n - m - 1$，n 为样本个数，m 为模型中自变量个数。基于原点的多元回归中，截矩项置为 0，F 统计量的总自由度为 n，残差自由度为 $n - m$。两种情形下，对 R^2 所作的修正都考虑了样本数 n 和自变量数 m，当 m 相对于 n 较小时，修正作用较弱。${R_a}^2$ 适用于对不同样本数和自变量数的数据进行回归的拟合结果优劣程度的比较。利用具正态误差的模拟数据，Ohtani（2000）指出 ${R_a}^2$ 是自变量矩阵 X 对应变量 y 贡献量的一个无偏估计。在偏线性回归中，由于用于修正的自由度未知，不能直接计算 ${R_a}^2$ 统计量。

典范冗余分析（RDA）中的典范 R^2 称为双多元冗余统计量（Bimultivariate redundancy statistic）（Miller & Farr，1971）或典范决定系数，其计算公式与多元回归类似，为所有应变量的回归平方和之和与所有应变量的总平方和之和的比值。典范分析中，除 Y 中的变量是标准化的，并且残差服从多元正态分布的极个别情形，F 统计量的显著性检验总是用置换（Permutation）方法。生态学数据一般很难满足这种极个别情形，但在满足的情况下，可以用自由度分别为（$m \times p$）和 $p(n - m - 1)$ 的 Fisher-Snedecor F 分布进行 F 统计量的检验（Miller，1975）。Peres-Neto 等（2006）用数值模拟的方法指出，对于正态分布或经过 Hellinger 转换的物种多度数据进行的 RDA，将公式（6）应用到典范 R^2 的调整的双多元冗余统计量 ${R_a}^2$，是自变量矩阵 X 对应变量矩阵 Y 贡献量的无偏估计。Hellinger 转换是 5 种对群落组成数据进行转换的方法之一。通过转换，数据中出现大量零值，从而可以应用线性方法，如主成分分析（PCA）或 RDA 进行分析（Legendre & Gallagher，2001）。

多元回归和典范分析中调整的决定系数有时会出现负值。对大样本数据，当自变量不比随机正态变量解释的更多时，${R_a}^2$ 取零值。出现负值的 ${R_a}^2$ 可解释为零，对应自变量比随机正态变量解释更少的情形。

4 方差分解

方差分解可用来分析用两个或更多个互补的假设来解释生态响应变量的变化方差。比如说，物种多度可能随着生物和非生物因素的改变而变化。研究 β-多样性时，群落组成数据的总方差（SS（Y））可分解为由两部分因素引起：一个或一组环境变量的变化和采样点之间空间关系变化。将群落组成数据对空间变量做拟合可以用来分析数据中是否存在不同尺度上的显著空间格局。由于环境数据常常具有空间结构，显著空间格局的存在既可以用来支持中性模型（Bell，2001；Hubbell，2001；He，2005），也可以用来支持环境控制假说。而物种和环境变量显著性关系的存在将有力地支持环境控制假说，但正如 Legendre 等（2005）所讨论的，这并没有和中性过程假说相矛盾。

Borcard 等（1992）及 Borcard 和 Legendre（1994）首先介绍了将生态方差分解成两部分的方差分解方法。下面将针对群落组成数据 Y 的分析介绍方差

分析。由于偏线性回归和偏典范分析的代数计算一致，这种方法也可以用于单个应变量 y 的情形。

应变量数据表 Y 关于两个自变量矩阵 X 和 W 的方差分解包括如下对应于不同研究目的的 3 个步骤：

1. 求取各方差组分

表 1 中列出了基于 3 个多元回归（单变量 y）或 3 个典范分析（多元应变量 Y）的计算过程：

表 1　计算图 1 所示调整的方差组分[a] 到 [d]的方法（需使用 3 个多元回归或典范分析）

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

典范分析 Canonical analyses	计算 R^2 Compute R^2(eq. 5)	计算 R_a^2 和各方差组分 Compute R_a^2(eq. 6) and fractions of variation	显著性检验可行性 Tested for significance
Y ~ X	R^2 of Y ~ X	[a + b] = R_a^2 of Y ~ X	可行 Yes
Y ~ W	R^2 of Y ~ W	[b + c] = R_a^2 of Y ~ W	可行 Yes
Y ~ (X,W)	R^2 of Y ~ (X,W)	[a + b + c] = R_a^2 of Y ~ (X,W)	可行 Yes
		[a] = [a + b] − [b]	可行 Yes
		[b] = [a + b] + [b + c] − [a + b + c]	不可行 No
		[c] = [b + c] − [b]	可行 Yes
		Residuals = [d] = 1 − [a + b + c]	不可行 No

1）Y ~ X 的典范分析。应用方程(5)和(6)计算 R^2和 R_a^2。假设图 1 中长方形的面积为 1，R_a^2对应左边圆的面积，含调整的组分 [a] 和 [b]。

2）Y ~ W 的典范分析。计算 R^2和 R_a^2。R_a^2对应图 1 中右边圆的面积，含调整的组分 [b] 和 [c]。

3）Y ~ (X, W) 的典范分析。计算 R^2 和 R_a^2，R_a^2对应图 1 中两个圆合起来占的面积，含调整的组分[a]、[b] 和 [c]。

4）计算组分[b]：[b] = [a + b] + [b + c] − [a + b + c]。

5）计算组分[a]：[a] = [a + b] − [b]

6）计算组分[c]：[c] = [b + c] − [b]

7）计算代表残差方差的组分[d]：[d] = 1 − [a + b + c]。

可以将这些值带入如图 1 所示的 Venn 图中进行表示。由于计算是基于调整的决定系数，这些值有时可能为负值，可将其解释为 0。

当 X 代表环境变量，W 代表采样点的空间关系时，图 1 所示的 Venn 图反映了如下信息：

1）包含[a + b]的圆表示 Y 的方差中可以被环境变量解释的部分。其中，[b]是被 X 和 W 共同解释的部分，或具有空间结构的环境变量解释的方差；[a]是环境解释方差中未被空间关系解释的部分。

2）包含[b + c]的圆表示 Y 的方差中可以被 W 中空间关系解释的部分。其中，[c]是唯一被 W 中关于空间变量的线性模型所解释，而没有被环境变量 X 的线性模型所解释的部分。这部分可能是由未被 X 包括的具空间结构的环境变量引起或是由环境变量 X 对 Y 的非线性作用引起。也可能是由 Y 描述的生态群落的某些过程，如竞争或扩散过程所引起。此情况下无法将这部分方差和环境变量联系起来进行解释。

Borcard 等(1992)及 Borcard 和 Legendre (1994)采用采样点坐标的三次多项式函数作为矩阵 W 作方差分解来进行只考虑大尺度空间格局的研究。近年来，Borcard 和 Legendre (2002)及 Borcard 等(2004)介绍了 PCNM 分析，这种分析建立的矩阵 W 中空间指标反映了采样点空间关系的谱分解。PCNM 分析使得研究者可以在所有空间尺度上建立模型来分析这些关系。这种分析也被称为‘基于距离的特征向量图”(DBEM)或 Moran 特征向量图(MEM)(Dray *et al.*, 2006)。

2. 方差各组分的显著性检验

为支持本节第一段中的推论，需要对方差各组分的显著性进行检验。可以用参数检验或置换(Permutation)检验对关于组分[a + b]、[b + c]和[a + b + c] (表 1) 的 3 个回归或典范分析的 F 统计量进行检验。但这种方法不适用于对[a]和[c]进行检验。如表 1 所示，无法对[b]进行显著性检验。残差方差[d]与其自由度构成对其它方差组分进行检验的 F 统计量的分母。

对组分[a]和[c]的显著性检验需要分别进行Y ~ X|W 和 Y ~ W|X 的偏典范分析。依据方程(2) ~ (4)可以计算 F 统计量的值。对这些 F 统计量的检验需要用到一种特殊的置换方法——残差置换(Legendre & Legendre, 1998；Anderson & Legendre, 1999)。

3. 拟合值作图

对对应于[a + b]、[b + c]、[a + b + c]、[a]和[c]

的拟合值作图将有助于对他们进行解释。对于单应变量 y 情形，对应于这些组分的多元和偏多元回归的拟合值可用来作图。对于多元应变量矩阵 Y 情形，如群落组成数据表，拟合值存在于一系列由典范和偏典范分析形成的多元样点得分数据表中。每一个表中的前几个轴对应于最大的几个典范特征值，可用来做图。对应于不具有空间结构的组分[a]需要使用点图，如 Bubble plots，这种图显示在每一个采样点的局部特征。图形插值方法，如 kriging，用于具有空间相关的其它组分。

对应变量矩阵 Y 的方差分解还可以对于 3 个或 4 个自变量矩阵进行。其算法包括更多的步骤，在 R 统计软件的‘ vegan ’包中进行了介绍（Oksanen *et al*.，2007）。

5　讨　论

群落组成数据表的方差分析是群落生态学中广泛使用的一种方法。正如前言中所述，群落组成数据表中的总方差 SS（Y）是 β-多样性的度量，反映了研究区采样点之间的多样性。自 20 世纪 70 年代以来，排序方法，如主成分分析（Principal component analysis，PCA）、对应分析（Correspondence analysis，CA）和主坐标分析（Principal coordinate analysis，PCoA），被用来将群落组成数据表的方差分解到各正交轴，进而作出排序图或与潜在的解释变量相联系。20 世纪 80 和 90 年代，生态学家最初通过使用 Canoco 软件，开始广泛应用典范排序法（ter Braak，1988；ter Braak & Smilauer，2002）。典范排序使得有可能直接将感兴趣的环境变量作为约束条件考虑到排序分析中，因此，也可称为‘ 约束下的排序方法 ’。生态学家迅速利用这一先进的方法并应用到物种-环境关系的分析中[1)]。Legendre（1990）提出用典范分析来建立群落组成数据的空间结构的模型，其中，采样点的空间关系用其地理坐标的多项式函数表示。这进而发展成在环境和空间组分间进行方差分解的方法。

方差分解已成为利用环境和空间变量解释 β-多样性的一种方法。在最近的一次检索中，*ISI Web of Knowledge* of the *Institute for Scientific Information* 列出了 603 篇使用或引用这种方法的论文。这些应用实例涵盖了绝大多数的生物群体。比如，泥炭沼泽中甲螨虫空间变化的研究中，对采自 70 个土壤钻孔中的 35 个甲螨虫种的数据依据一系列环境和空间变量信息通过方差分解进行了分析。Borcard 等（1992）及 Borcard 和 Legendre（1994）的研究中利用采样点地理坐标的多项式函数来表示土壤钻孔间的空间关系。Borcard 和 Legendre（2002）和 Borcard 等（2004）中 PCNM 空间函数的使用更好地解释了土壤钻孔间物种组成的空间变化（β-多样性）。

β-多样性是生态学家用来比较不同研究区或同一研究区共存的不同生态群落的一种手段。1）只有在研究区面积相同、采样方法一致的情况下，对不同研究区的比较才是有意义的。比如对构成纬度带梯度的中国森林生物多样性监测网络中 5 个 24 hm^2 样地的比较研究。若所有被比较的样地都被类似划分为 20 m × 20 m、40 m × 40 m 或其它的格点进行研究，这样的比较就是有意义的。方差分解中 SS（Y）是度量每个研究区内 β-多样性的一个便利的指标。总 β-多样性变化可以被分解到一个或一系列环境变量和空间变量中。进而可以对 5 个研究区方差分解的结果进行比较。2）对每个样地，我们可以进行乔木层 β-多样性和其它植被层 β-多样性的比较。方差分解方法使得研究者可以将每个群落的 β-多样性变化分解到环境和空间变量上并进一步研究确定对于不同的生物群体控制空间结构的因素是否一样。

对群落组成数据进行统计分析并不轻松。为正确检验有关生态系统 β-多样性形成和维持因素的假说，使用的显著性检验不能依赖于不切合实际的假设。如当数据不满足多元正态假设时，不能使用依赖于多元正态假设的显著性检验。显著性检验必须具有正确的 I 类错误比例和足以检测出存在的自然或人为影响的功效（Power）。当确定存在显著性影响时，应使用无偏统计量（$R_a{}^2$）进行表示。生态分析的结论将会被管理实践者用来作出生态系统管理中的重要决策，因而这些结论必须有严格的科学性。

本文介绍了历经多年发展的方差分解方法。方差分解使得研究者可以检验关于生态系统 β-多样性起源的精确假说，并确定空间变化中有多少是由环境变量所控制，有多少尚无法解释。后一部分可能是受未测量的环境变量的影响，也可能是由需要进一步研究的诸如竞争或扩散等群落过程所决定。任一情况下，生态方差分解中统计学的正确使用是最重要的。

[1)] Two bibliographies on the applications of canonical analysis to ecology，covering together the period 1986 to 1996，contain a total of 804 entries. They are available from H. J. B. Birks，Botanical Institute，University of Bergen，Allégaten 41，N-5007 Bergen，Norway and also on the URL http ://www.bio.umontreat.ca/casgrain/cca_bib1.

参 考 文 献

Anderson MJ, Legendre P (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, Legendre P (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, Legendre P (2002). All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecological Modelling*, 153, 51 – 68.

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

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

Dray S, Legendre P, Peres-Neto PR (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 SP (2001). *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton University Press, Princeton, New Jersey.

Legendre P (1990). Quantitative methods and biogeographic analysis. In: Garbary DJ, South RG eds. *Evolutionary Biogeography of the Marine Algae of the North Atlantic*. NATO ASI Series, Vol. G 22. Springer-Verlag, Berlin, 9 – 34.

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

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

Legendre P, Gallagher ED (2001). Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129, 271 – 280.

Legendre P, Legendre L (1998). *Numerical Ecology* 2nd English edn. Elsevier Science BV, Amsterdam.

Manly BJF (1997). *Randomization, Bootstrap and Monte Carlo methods in biology* 2nd edn. Chapman and Hall, London.

Miller JK (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 JK, Farr SD (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, Kindt R, Legendre P, O'Hara RB (2007). Vegan: community ecology package version 1.8 – 5. URL http://cran.r-project.org/.

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

Rao CR (1964). The use and interpretation of principal component analysis in applied research. *Sankhyaá, Series A, Indian Journal of Statistics*, 26, 329 – 358.

R Development Core Team (2007). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

ter Braak CJF (1986). Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology*, 67, 1167 – 1179.

ter Braak CJF (1987a). The analysis of vegetation-environment relationships by canonical correspondence analysis. *Vegetatio*, 69, 69 – 77.

ter Braak CJF (1987b). Ordination. In: Jongman RHG, ter Braak CJF, van Tongeren OFR eds. *Data Analysis in Community and Landscape Ecology*. Pudoc, Wageningen (reissued in 1995 by Cambridge University Press, Cambridge), 91 – 173.

ter Braak CJF (1988). *CANOCO — a FORTRAN program for Canonical Community Ordination by [partial] [detrended] [canonical] Correspondence Analysis, Principal Component Analysis and Redundancy Analysis (version 2.1)*. Agricultural Mathematics Group, Ministry of Agriculture and Fisheries, Wageningen.

ter Braak CJF, Smilauer P (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 RH (1960). Vegetation of the Siskiyou mountains, Oregon and California. *Ecological Monographs*, 30, 279 – 338.

Whittaker RH (1972). Evolution and measurement of species diversity. *Taxon*, 21, 213 – 251.

责任编辑：姜联合

Journal of
Plant Ecology

VOLUME 1, NUMBER 1,
PAGES 3–8

MARCH 2008

doi: 10.1093/jpe/rtm001

Advanced Access published
on 17 July 2007

available online at
www.jpe.oxfordjournals.org

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

Pierre Legendre

Département de sciences biologiques, Université de Montréal, CP 6128, succursale Centre-ville, Montréal, Québec, Canada H3C 3J7
E-mail: Pierre.Legendre@umontreal.ca

Abstract

Aims
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. The present report describes how to analyse beta diversity from community composition and associated environmental and spatial data tables.

Methods
Beta diversity can be studied by computing diversity indices for each site and testing hypotheses about the factors that may explain the variation among sites. Alternatively, one can carry out a 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 herein.

Important findings
Variation partitioning is a method of choice for the interpretation of beta diversity using tables of environmental and spatial variables. Beta diversity is an interesting 'currency' for ecologists to compare either different sampling areas or different ecological communities co-occurring in an area. 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.

Keywords: Adjusted coefficient of determination • beta diversity • biodiversity • canonical redundancy analysis • community composition • variation partitioning

Received: 4 August 2006 Accepted: 4 August 2006

Introduction

Ecologists collect community composition data (species presence–absence or abundance data) at several sites in a region of interest in order to analyse 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, 1987*a*, *b*). These analyses are described in several textbooks, including Legendre and Legendre (1998). They are implemented in computer packages such as Canoco (ter Braak and Smilauer, 2002) and the 'vegan' library (Oksanen et al., 2007) of the R statistical language (R Development Core Team, 2007).

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 report 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 $\mathbf{y}\sim\mathbf{X}|\mathbf{W}$ represents the partial linear regression of a response variable $\mathbf{y}$ (vector of length n) on a matrix $\mathbf{X}$ containing m explanatory variables, while controlling for the linear effect of a matrix $\mathbf{W}$ containing q covariables. Partial regression is computed in two steps: (i) regress $\mathbf{X}$ on $\mathbf{W}$ and compute the residuals $\mathbf{X}_{\mathrm{res}(\mathbf{W})}$; and (ii) regress $\mathbf{y}$ 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):

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

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:

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

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

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

or, using Fig. 1:

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

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 it is 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 $\mathbf{y}\sim\mathbf{X}|\mathbf{W}$ and $\mathbf{y}\sim\mathbf{W}|\mathbf{X}$ will be computed and tested for significance.

Partial canonical analysis

Similarly, the notation $\mathbf{Y}\sim\mathbf{X}|\mathbf{W}$ represents the partial canonical redundancy analysis (partial RDA) of a response data matrix $\mathbf{Y}$ of size ($n \times p$) on a matrix $\mathbf{X}$ containing m explanatory variables, while controlling for the linear effect of a matrix $\mathbf{W}$ 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 $\mathbf{Y}\sim\mathbf{X}|\mathbf{W}$ and $\mathbf{Y}\sim\mathbf{W}|\mathbf{X}$ will be computed and tested for significance.

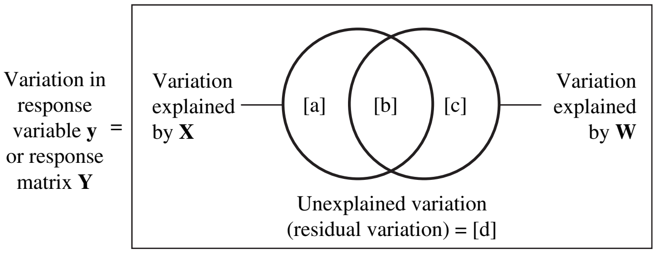
Variation in response variable y or response matrix Y =
Variation explained by X
[a]
[b]
[c]
Variation explained by W
Unexplained variation (residual variation) = [d]

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

Unadjusted and adjusted coefficients of determination

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

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

where 'regression SS' is the sum-of-squares of the fitted values of the regression equation. It measures the proportion of the variation of $\mathbf{y}$ 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^2_{a} (Ezekiel, 1930):

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

The right-hand parentheses of equation 6 shows that R^2_a 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 ($n - m$). In both cases, the correction takes into account the number of objects *n* and the number of explanatory variables *m*; the correction is light when *m* is small when compared with *n*. 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 then unknown.

In 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 multinormal. 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 ($m \times p$) 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 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 response 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 total variation of the community composition data table, denoted SS(**Y**), 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 either for 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 to 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 also be applied 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.

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.

(i) 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].

(ii) 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].

(iii) 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].

(iv) From these first results, compute fraction [b] by subtraction: [b] = [a + b] + [b + c] − [a + b + c].

(v) Compute fraction [a] by subtraction: [a] = [a + b] − [b].

(vi) Compute fraction [c] by subtraction: [c] = [b + c] − [b].

(vii) 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

Table 1 Method for calculating the adjusted fractions of variation [a] to [d] depicted in Fig. 1

Canonical analyses	Compute R^2 (eq. 5)	Compute R_a^2 (eq. 6) and fractions of variation	Can be tested for significance
$\mathbf{Y} \sim \mathbf{X}$	R^2 of $\mathbf{Y} \sim \mathbf{X}$	[a + b] = R_a^2 of $\mathbf{Y} \sim \mathbf{X}$	Yes
$\mathbf{Y} \sim \mathbf{W}$	R^2 of $\mathbf{Y} \sim \mathbf{W}$	[b + c] = R_a^2 of $\mathbf{Y} \sim \mathbf{W}$	Yes
$\mathbf{Y} \sim (\mathbf{X},\mathbf{W})$	R^2 of $\mathbf{Y} \sim (\mathbf{X},\mathbf{W})$	[a + b + c] = R_a^2 of $\mathbf{Y} \sim (\mathbf{X},\mathbf{W})$	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

Three multiple regressions or canonical analyses are required.

coefficients of determination, the fractions can, on occasion, take negative values. These are interpreted as zeros, as explained in the previous section.

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:

(i) 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**.

(ii) 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 non-linear 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 (principal coordinate analysis of neighbour matrices) 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. PCNM geographic functions are a type of 'distance-based eigenvector maps' (DBEMs), which belong to a general class called 'Moran's eigenvector maps' (MEMs) (Dray et al., 2006).

Testing the Significance of the Fractions

The fractions must be tested for significance in order to support fully the reasoning described in the first paragraph of this section. 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 (see below), 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 $\mathbf{Y} \sim \mathbf{X}|\mathbf{W}$ and $\mathbf{Y} \sim \mathbf{W}|\mathbf{X}$ 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).

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., 2007) of the R statistical language.

Discussion

Analysis of the variation of a community composition data table is a widely used approach in community ecology. As stated in the Introduction, the total variation in a community composition table, denoted SS($\mathbf{Y}$), is a measure of beta diversity, which is the diversity among sites in the study area. Ordination methods such as PCA, correspondence analysis (CA) and principal coordinate analysis (PCoA) have been used since the 1970s to partition the variation of community composition data tables into orthogonal axes, which can be used to produce ordination plots or can be related to potentially explanatory variables. In the years 1980 and 1990, canonical ordination methods were made widely available to ecologists, firstly through the program Canoco (ter Braak, 1988; ter Braak and Smilauer, 2002). Canonical ordination offers the possibility of directly incorporating the environmental variables of interest in the analysis as constraints for the ordination, hence the expression 'constrained ordination methods'. Ecologists quickly took advantage of this improved methodology and applied it to all problems of species–environment relationships. (Two bibliographies on the applications of canonical analysis to ecology, covering together the period 1986 to 1996, contain a total of 804 entries. They are available from H. J. B. Birks, Botanical Institute, University of Bergen, Allégaten 41, N-5007 Bergen, Norway, and also on the URL http:/www.bio.umontreal.ca/casgrain/cca_bib/.) In 1990, Legendre proposed to use canonical analysis to model the spatial structure of community composition data, representing the spatial relationships among the sampling sites by a polynomial function of their geographic coordinates. That development led to the method of variation partitioning among environmental and spatial components, described in the previous section.

Variation partitioning has become a method of choice for the interpretation of beta diversity using tables of environmental and spatial variables. At the last count, the ISI Web of Knowledge of the Institute for Scientific Information listed 603 papers that had used the method or were referring to it. The published examples concern most groups of organisms. An example is the analysis of the spatial variation of a community of oribatid mites in the peat carpet of a peat bog. Thirty-five mite species collected in 70 soil cores were analysed by variation partitioning with respect to a set of environmental and spatial variables. In the papers of Borcard et al. (1992) and Borcard and Legendre (1994), a polynomial function of the geographic coordinates was used as the spatial representation of the spatial relationships among the soil cores. In Borcard and Legendre (2002) and Borcard et al. (2004), PCNM spatial base functions were used instead, providing a much better explanation of the spatial variation in species composition among the cores (beta diversity).

Beta diversity is an interesting 'currency' for ecologists to compare either different sampling areas, or different ecological communities co-occurring in an area. (i) For the comparison of different study areas to be meaningful, the areas must be of the same size and sampled in the same way. An example would be the comparative study of the five 24 ha forest plots that are presently monitored under the auspices of the Chinese Forest Biodiversity Monitoring Network, forming a latitudinal gradient through China. The comparison would be meaningful if all compared plots are similarly divided into cells of 20 m × 20 m, or 40 m × 40 m, etc. In the framework of variation partitioning, SS($\mathbf{Y}$) is a convenient measure of beta diversity within each area. The total beta variation can be partitioned among one or several sets of environmental variables, as well as a table of spatial variables. The resulting partitions of the five separate areas can be compared using the results of these analyses. (ii) In each of these forest plots, one could compare the beta diversity of trees with that of other vegetation strata, for example, after dividing the plot into cells of equal sizes. The method of variation partitioning would allow researchers to partition the beta variation of each community among environmental and spatial variables and determine if the factors controlling the spatial organization are the same for the different groups of organisms.

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, whether 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 report 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

This paper is dedicated to my mentor, Robert R. Sokal. Born in Vienna, Professor Sokal spent his formation years in Shanghai, where he received high-school and college education and learned Chinese.

Professor Sokal founded the field of Numerical Taxonomy with Professor Peter H. A. Sneath. Later, he developed the field and methods of spatial analysis of genetic data. I learned a great deal from him during a sabbatical in 1985, where we intensively discussed the principles and methods of spatial genetic analysis. This paper is a testimonial of Professor Sokal's contribution to the development of the ideas behind the methods of multiscale analysis of ecosystems. The author is grateful to Drs. Daniel Borcard (Université de Montréal, Montréal, Québec) and Fangliang He (University of Alberta, Edmonton, Alberta), as well as two anonymous reviewers, for judicious comments on the manuscript of this paper. Funding was provided by Natural Sciences and Engineering Research Council of Canada (NSERC) grant no. OGP0007738 to P.L.

References

Anderson MJ, Legendre P (1999) An empirical comparison of permutation methods for tests of partial regression coefficients in a linear model. *J Stat Comp Simul* **62**:271–303.

Bell G (2001) Neutral macroecology. *Science* **293**:2413–2418.

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

Borcard D, Legendre P (2002) All-scale spatial analysis of ecological data by means of principal coordinates of neighbour matrices. *Ecol Modell* **153**:51–68.

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

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

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

Ezekiel M (1930) *Methods of Correlation Analysis*. New York: John Wiley and Sons.

He F (2005) Deriving a neutral model of species abundance from fundamental mechanisms of population dynamics. *Funct Ecol* **19**: 187–193.

Hubbell SP (2001) *The Unified Neutral Theory of Biodiversity and Biogeography*. Princeton, NJ: Princeton University Press.

Legendre P (1990) Quantitative methods and biogeographic analysis. In: Garbary DJ, South RG (eds). *Evolutionary biogeography of the marine algae of the North Atlantic*. NATO ASI Series, Vol. G 22. Berlin: Springer-Verlag, 9–34.

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

Legendre P, Borcard D, Peres-Neto PR (2005) Analyzing beta diversity: partitioning the spatial variation of community composition data. *Ecol Monogr* **75**:435–450.

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

Legendre P, Legendre L (1998) *Numerical Ecology*, 2nd English edn. Amsterdam: Elsevier Science BV.

Manly BJF (1997) *Randomization, Bootstrap and Monte Carlo Methods in Biology*, 2nd edn. London: Chapman and Hall.

Miller JK (1975) The sampling distribution and a test for the significance of the bimultivariate redundancy statistic: a Monte Carlo study. *Multivar Behav Res* **10**:233–244.

Miller JK, Farr SD (1971) Bimultivariate redundancy: a comprehensive measure of interbattery relationship. *Multivar Behav Res* **6**: 313–324.

Ohtani K (2000) Bootstrapping R^2 and adjusted R^2 in regression analysis. *Econom Modell* **17**:473–483.

Oksanen J, Kindt R, Legendre P, O'Hara RB (2007) *vegan: Community Ecology Package Version 1.8–5*. URL http://cran.r-project.org/.

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

Rao CR (1964) The use and interpretation of principal component analysis in applied research. *Sankhyaá, Ser A, Indian J Stat* **26**: 329–358.

R Development Core Team (2007) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria. URL http://www.R-project.org.

ter Braak CJF (1986) Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis. *Ecology* **67**:1167–1179.

ter Braak CJF (1987*a*) The analysis of vegetation–environment relationships by canonical correspondence analysis. *Vegetatio* **69**:69–77.

ter Braak CJF (1987*b*) Ordination. In: Jongman RHG, ter Braak CJF, van Tongeren OFR (eds). *Data Analysis in Community and Landscape Ecology*. Pudoc, Wageningen [reissued in 1995 by Cambridge: Cambridge University Press], 91–173.

ter Braak CJF (1988) *CANOCO—A FORTRAN Program for Canonical Community Ordination by [Partial] [Detrended] [Canonical] Correspondence Analysis, Principal Component Analysis and Redundancy Analysis (Version 2.1)*. Wageningen: Agricultural Mathematics Group, Ministry of Agriculture and Fisheries.

ter Braak CJF, Smilauer P (2002) *Canoco Reference Manual and CanoDraw for Windows User's Guide: Software for Canonical Community Ordination (Version 4.5)*. Ithaca, NY: Microcomputer Power.

Whittaker RH (1960) Vegetation of the Siskiyou mountains, Oregon and California. *Ecol Monogr* **30**:279–338.

Whittaker RH (1972) Evolution and measurement of species diversity. *Taxon* **21**:213–251.