5 约束排序与置换检验(Constrained Ordination and Permutation tests)

在这一章，我们将讨论约束排序及其相关的内容：环境因子的逐步筛选，蒙特卡罗置换检验和变量分解分析。

5.1 线性多元回归模型 (Linear multiple regression model)

首先,我们必须回顾一下传统的线性回归模型，因为这对于我们理解“直接梯度分析”（约束排序）相当重要。

图5-1展示的是最简单的线性回归模型，线性模型可以模拟响应变量Y依赖自变量X的程度。图5-1中不仅有拟合回归线，也展示了模拟值和实测值之间的差别。模拟值^Y_i (回归线上的值)与实测值Y_i之间的差值叫做回归残差（regression residual）,用e表示。

所有的统计模型(statistical models,包括回归模型)有个重要的特征是它们都有两个主要的部分构成:系统组成部分（systematic component）表示响应变量中能被一个或更多的解释变量（模型）解释的部分，这部分用带参数的函数表示。另外一部分就是随机部分（stochastic component），表示不能被目前解释变量（模型）所能解释的部分。随机部分通常用概率和分布特性来定义。

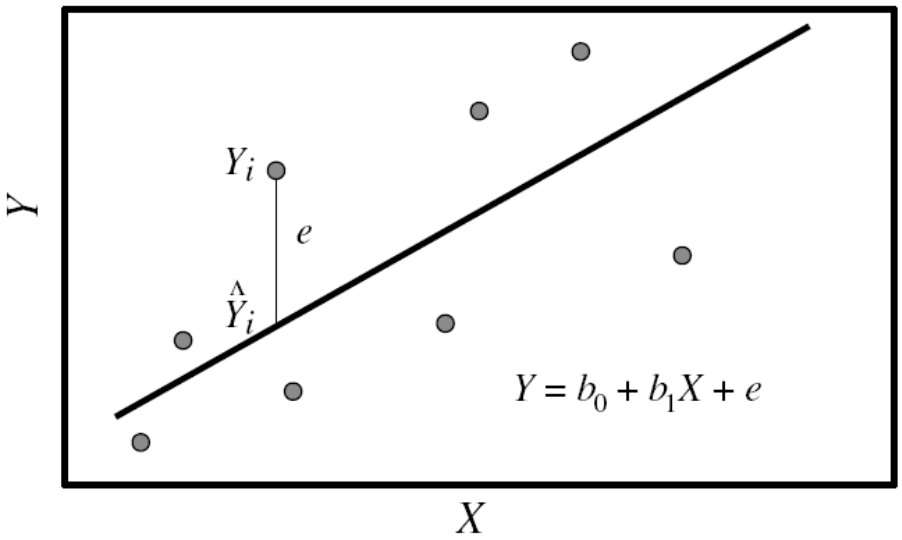

Figure 5-1. Graphical representation of a simple linear regression model.

我们通常通过响应变量有多少能够被系统组成部分解释来评估拟合模型的好坏。也经常将能被解释和未被解释的部分进行比较。目的在于，尽力去构建一个最简约的回归模型来解释最多的变化量，让所有的自变量对于响应变量的解释都有显著贡献。我们可以通过逐步迭代（回归）(stepwise selection)的方式来选择解释变量(环境变量)的子集合，在排序术语中往

往叫预选（forward selection）。预选变量的过程是无响应变量的零和假设开始，零和假设是响应变量中没有可以被解释变量预测,而仅仅由随机变量解释。当我们选择一个解释变量(环境变量)进入分析，可以导致回归模型能解释一部分响应变量。可以根据所加入的变量所能解释部分的大小来确定是否选择的该环境变量。另外，需要用随意置换(randomly swap)环境因子的值来检验这种解释量是偶然的，还是真的为环境变量所解释？如果被检验的变量所能解释的部分被证明是非随机的(统计显著), 我们就可以接受这个变量。可以重复这个过程,再进一步从剩下的变量中选择另外更好的变量，直到选择具有足够的变量为止。

5.2 约束排序模型 （Constrained ordination model）

在第三章，非约束的排序（PCA 和CA）被定义为寻找潜在的梯度代表最优的解释变量(预测器)来拟合物种的回归模型。

约束排序跟非约束的排序有一个很大的区别，非约束排序是虚拟的（潜在的）梯度、而约束排序的梯度是明确给出的。这些梯度（排序轴）是参与排序的环境变量的线性组合。因此我们通过合成变量（排序轴）尽力解释物种的多度变化，这些合成变量是实测环境变量的线性组合。

因此，约束排序方法（RDA和CCA）类似于多元多重回归。但在多元多重回归中，如果有m个响应变量，p个环境因子，我们必须估计出m×p个的参数（回归系数）（每个方程需要p个参数，m个方程自然是需要m×p个的参数）。然而，在约束排序里面，不必这么麻烦，环境因子对于响应变量的影响被集中在几个合成的梯度（排序轴），也叫典范轴(canonical axes)。典范轴的数量是跟独立解释变量的数量一样多，但是我们经常使用前面几轴。

在CANOCO里面，如果有协变量(covariables)，我们经常使用偏分析（partial analyses）。有协变量情况，表示我们要将这些协变量的所能解释的部分先剔除出去。协变量在方差分析中也有相同的用法，通常是把量化的协变量作为一种因子处理。而在传统的回归中，协变量的概念是不常用的，协变量与真变量没有什么不同，叫法不同而已。

5.3 RDA :约束的 PCA（ RDA: constrained PCA）

上一节提到关于RDA（redundancy analysis）的概念，其实RDA就是PCA的约束排序。下面以两个环境变量（Z_1和Z_2）跟第一排序轴（第一主分量）来说明RDA的运算过程。

PCA和RDA排序的目的均是寻找新的变量作为最好的预测器来预测物种（响应变量）分布。我们设立这个新变量为X（假设是第一轴）。跟实测的环境变量一样，X在每个样方里面有个对应的值。假定这个新变量在第i个样方的值为X_i，那么第k个物种在第i样方的值可以通过下面的公式来预测。

$$Y_{ik} = b_{0k} + b_{1k}X_i + e_{ik}$$

在这里，无论是PCA还是RDA，都必须估计两套参数：X_i和b_{1k}。X_i的值是样方在第一轴

的坐标。每个物种回归系数b_{1k},代表物种在第一轴的坐标。另外一个参数b_{0k}代表回归拟合线的截距，可以通过原始数据的中心化将它归零（详见4.4节）。

其实，到这里为止，PCA和RDA样方坐标X_i算法是相同的。但后来，RDA样方的坐标值X_i是经过约束的，是环境因子的线性组合。在这里举例说明计算过程，假设有两个实测的环境变量Z_1和Z_2，新变量X_i可以表示为环境变量Z_1和Z_2线性组合

$$X_i = c_1 z_{i1} + c_2 z_{i2}$$

注意这里的参数C_1和C_2并不是环境因子在第一轴的坐标，而是Canoco分析结果里面回归系数，仅仅表示环境因子与排序轴之间的相关性。

我们可以组合上面两个等式到一个等式里面，实际变成一个多重多元回归方程组：

$$Y_{ik} = b_{0k} + b_{1k} c_1 z_{i1} + b_{1k} c_2 z_{i2} + e_{ik}$$

在这个表达式里面，系数$b_{ik}C_j$代表多元多重回归模型中真正的回归系数(actual coefficients)，这个回归系数描述着k物种的多度取决于j环境因子的程度。如果有m个物种，p个环境因子，我们需要去估计m*p个回归系数。在RDA里面，我们仅仅需要估计的是被约束的那些回归系数：假设只有一个典范轴，我们仅仅需要估计m+p个参数（b_{ik}和C_j参数，b_{ik}是物种与轴之间的回归系数，后者是轴与环境因子自己的关系）。如果是两个轴，仅仅是2(m+p)个参数，相比多元多重回归m*p的系数，的确简单不少。【这是为什么要做约束排序的原因，也说明了如果环境因子和物种的数量很少的时候，是不必做约束排序的，做普通回归即可，m和p比较多的时候，约束排序的优势才能展示出来】

5.4 蒙特卡罗置换检验引论（Monte Carlo permutation test: an introduction）

CANOCO通过蒙特卡罗置换检验有能力去检验约束排序模型的显著性。这个统计检验基于普通的零和假设，这里的零和假设就假设物种与环境因子之间是相互独立，不相关的。置换检验的主要原则均在第三章的3.10节和3.11节已经提过，那里举例的完全随机的简单置换。而在CANOCO里面提供了丰富的置换方法，有空间上，有时间上或是逻辑内部结构限制（如图5-2），这些都是与实验设计与样方设计相关的。

图5-2 展示的是这些置换方法的选择窗口的首页。接下来的4小节将会比较详细怎么选择合适的置换方法。本书后面的研究案例也会对这些置换方法进行实践。

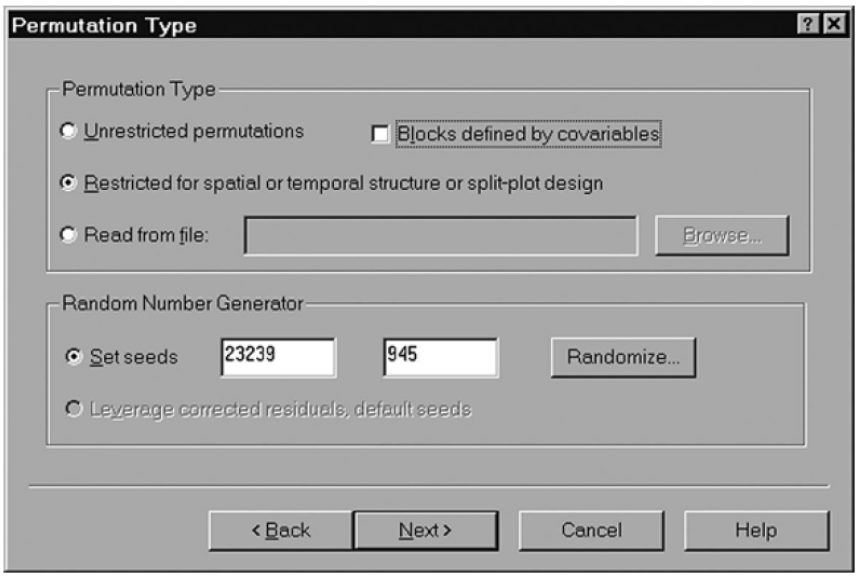

Figure 5-2. Introductory page of the Project setup wizard for selecting the Mon Carlo permutation type.

5.5 零和假设模型（Null hypothesis model）

在CANOCO里面，零和假设是物种矩阵与环境矩阵之间是独立的，简单讲，可以任意调换环境矩阵中各个样方直接的位置，对于约束排序的结果并没有显著影响。

具体的检验过程的算法并不在这里展示。我们仅仅要说明一下置换检验的基本过程：

- 我们先开始随机置换环境矩阵中样方的位置，然后保持物种的位置不变。这样每置换一次，物种和环境因子之间的组合就发生变化，每次都组成新的组合。
- 对每对新的组合，我们重新计算约束排序的过程（回归的过程），每次的算法都是一致，但是每个排序模型的优劣程度并不相同。这里我们可以用F-统计中F统计值来代表每个排序模型的优劣（将在下一节介绍）。
- 如果我们进行了N次置换，将有N个排序模型，也拥有N个F值，做N个F值频度分布（如图5-3）。我们可以找到如果是在不置换的情况下的回归方程的F_{data}值位置。如果F_{data}的处于图的右边区域，即处于低概率区域（比如说F_{data}值比95%个F值都大，也就是P<0.05）,此时我们可以拒绝零和假设，说明物种和环境因子直接是存在显著关系的，样方的位置不能随便调换。相反，如果F_{data}值比95%个F值都大，也就是P>0.05，则不能拒绝零和假设。

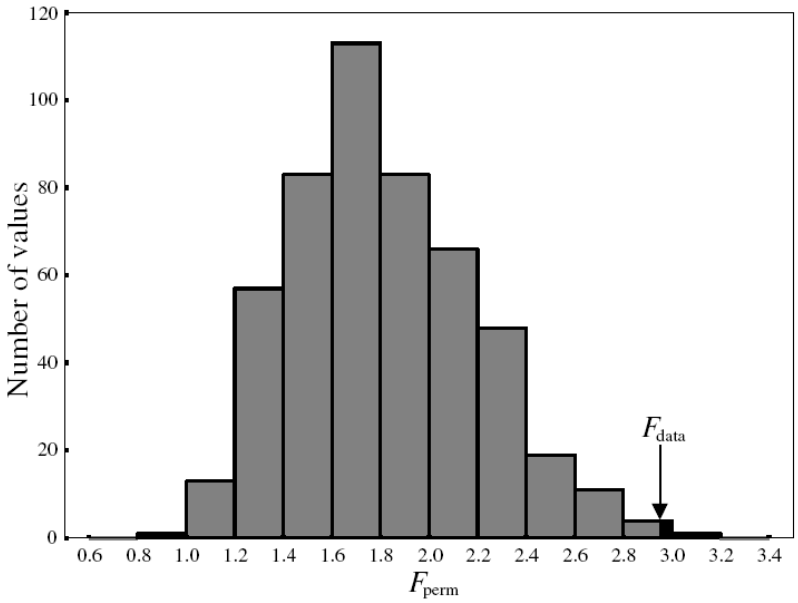

Figure 5-3. The distribution of the *F* -statistic values from a Monte Carlo permutation test compared wiht the *F* -statistic value of the 'true' data sets. The black area corresponds to permutation based *F* -ratio values exceeding the *F* -ratio value based on the 'true data'.

5.6 检验统计 Test statistics

前面一节描述了置换检验的基本过程，也提到在排序模型的质量检验是用类似于回归模型参数显著性F-统计检验。但因为排序模型的多维性，这个统计在约束排序的确很难定义。一般来说，能被环境因子解释物种的变量，要很多轴一起表示。但是每个典范轴的相对重要性（解释量）从第一轴到最后一轴是逐渐降低的，但我们很少忽略除了第一轴以为的其他轴。因此，我们既关注所有轴的累计解释量，又关注一个轴，通常是第一轴的解释量。在CANOCO 4.5里面有两种对应的置换检验：

- 第一轴检验（Test of the first canonical axis）使用F-统计的算法如下：

$$F_1 = \frac{\lambda_1}{\mathrm{RSS}/(n-p-q)}$$

这的λ_1代表第一轴的特征根，也代表第一轴所能解释的变化量。而RSS是残差平方和（the residual sum of square）缩写，代表不能被第一轴所能解释的物种变化量。*n*是轴的数量，*p*代表主环境变量的数量，*q*代表协环境变量的个数。

- 所有轴的检验（Test of the sum of the canonical eigenvalues）,也就是检测*p*个解释变量的整体效果。此时F值应该按照这么计算：

$$F_{\text{trace}} = \frac{\sum_{i=1}^{p} \lambda_i / p}{\text{RSS}/(n-p-q)}$$

此时的RSS，变成是目前不能被所有的典范轴所解释的物种变化量。

跟上一节讲得一样，我们可以首先计算用原始数据来算出F_{data}值，然后，随机置换环境矩阵中样方的位置m次，可以算出m个F值，作出m个F值的频度分布，看第一次用原始数据算出的F_{data}值在在这些m个值中的位置，图5-3所示。

当然，从图5-3我们也可以算出，用原始数据得到F_{data}值的拒绝零和假设后犯错误的概率：

$$P = \frac{n_x + 1}{N + 1}$$

这里的n_x表示所有的F值中，大于等于F_{data}值的个数，N表示所有F值的数量（置换的次数）。为什么分子和分母都要加1，是因为第一次求得的F_{data}也当作F值零和分布中的一员。为什么在CANOCO里面默认置换的数量一般是99，199或499，就是得加上1这个原因。

5.7 空间和时间约束 （Spatial and temporal constraints）

上一节讲的是样方之间并没有潜在空间相关结构情况（样方是随机抽取，相互独立）下的置换检验。在这种情况下，我们可以任意置换样方的位置，因为在零和模型条件下，每个环境因子的值与样方的之间的配比是等概率的。

当然，如果样方直接存在内在的联系，当样方不能随机置换时，这种等概率的配比是不存在的。CANOCO里面设立三种基本的样方的内部结构情况，见图5-4.

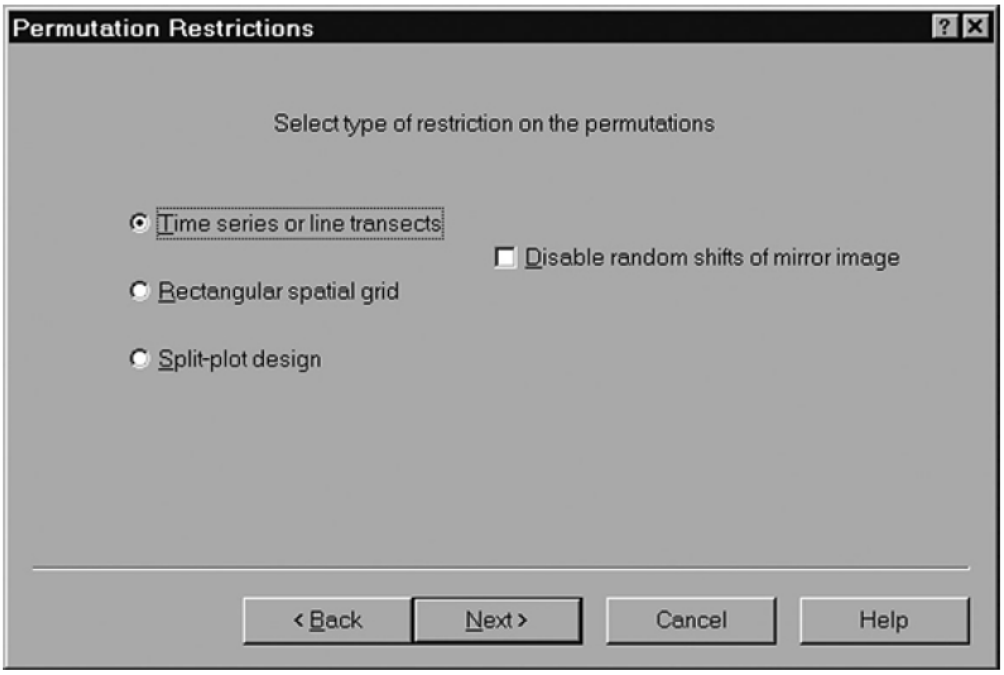

Figure 5-4. Project setup wizard page for selecting restrictions for a permutation test.

样方可能是按照空间或是时间系列来排列的。在这种情况下，样方的位置不能随便置换，因为我们认为在这种情况下，样方之间存在自相关，包括物种之间自相关，或是环境因子之间自相关。因此，在置换检验过程，我们不能随便打破样方之间的分布格局，因为我们检验所关注的焦点在于物种与环境因子的关系，而不是数据本身的关系。考虑到这种自相关的样方结构，CANOCO里面采取了旋转办法来解决这个置换的问题：将样方的头尾链在一起，形成一个圆筒，此时可以通过旋转，来改变样方与环境因子的配比，而不是随机的置换。具体更详细的内容可以参考CANOCO的手册。

同样，一个相同的自相关也存在与一个具有空间梯度的样带中。

最普通的置换限制是裂区设计（split-plot design）的数据，也是图5-4最后一个对话框所指的类型。这种类型的置换限制详见下一节和第15、16章的案例研究。

所有的这些置换限制，均可以进一步嵌套在另一个整区(blocks)限制。在排序分析里面，整区(blocks)的层次可以当作一种属性变量进入分析。

5.8 裂区约束（Split-plot constraints）

在裂区设计置换限制中，在Canoco 4.0和4.5使我们能够描述变量两个层次结构（两个误差水平），见第2.5节。

裂区里面比较高层次的叫所谓的“整区”（whole-plots）。每个整区都包含裂区，裂区代表实验设计里面低水平层次（见图2-5）。Canoco 为整区和裂区两个水平提高很多灵活的置换，从无置换开始，到通过在两个水平之内的空间和时间的限制置换均可实现（见图5-5）。CANOCO裂区设计置换限制是比较重要的，因为经常会用于评估多次观测群落组成的变化

的数据，详见第15章的例子。

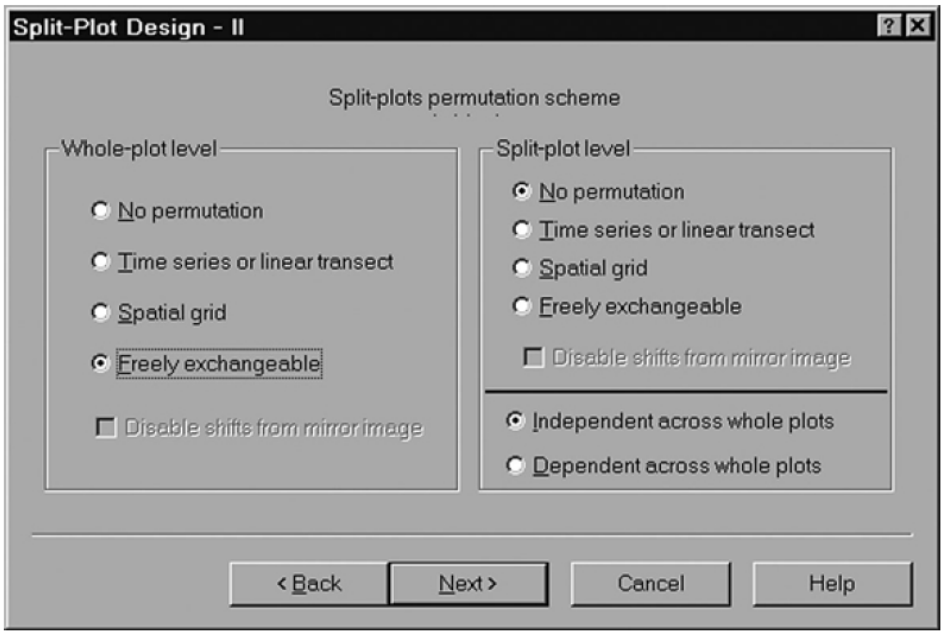

Figure 5-5. Project setup wizard page for specifying permutation restrictions for a split-plot design.

5.9 环境因子变量的预选 (Stepwise selection of the model)

在5.1节的最后部分，关于回归模型的环境变量预选已经详细说明。在CANOCO程序里面，也有相同的目标和方法来进行环境因子的预选。可CANOCO里面，可以用偏蒙特卡罗置换检验（partial Monte Carlo permutation test）来评估每个备选的环境变量的对于解释物种变量的贡献。

在CANOCO分析过程，如果我们选择手工预选变量（manual forward selection），程序会自动弹出如图5-6 那样的可选活动对话框。对话框可以展示了输入的环境变量的解释能力强弱的依次排列，图5-6展示现在已经有两个最强解释能力变量（moisture and manure）被选中（进入界面下面的框框里）。窗口上方的数字表示已经入选的两个变量的解释量占所有环境因子解释量的57%（0.624 of 1.098）。

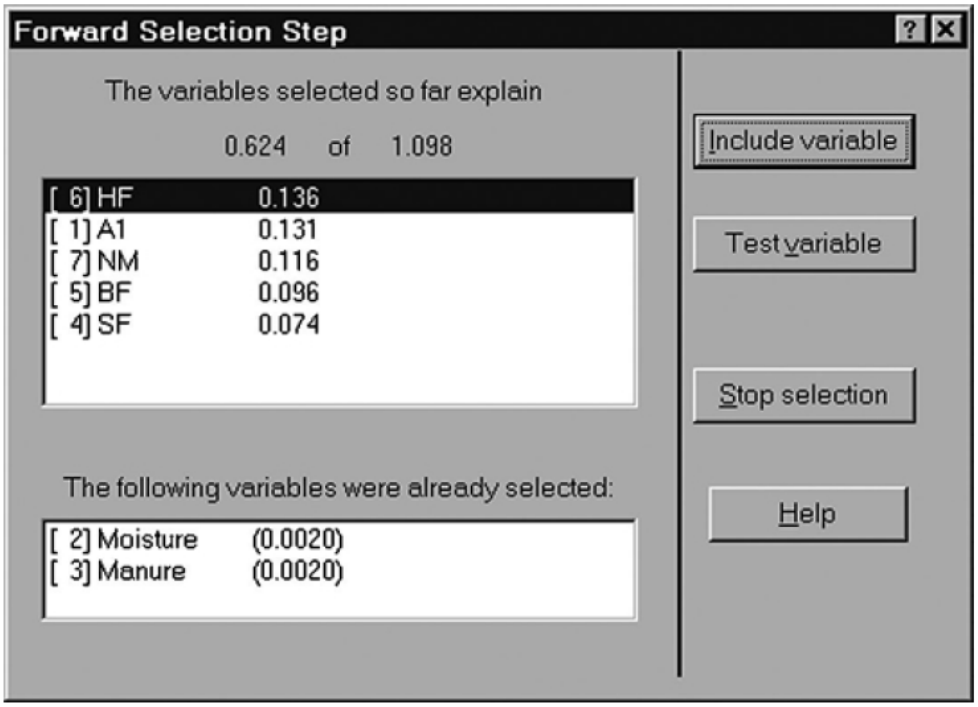

Figure 5-6. Dialog box for the forward selection of environmental variables.

窗口上面的框框展示的是备选的环境因子，环境因子后面的数字表示，如果将所对应的环境因子选人分析，可以增加的解释量。比如，HF(‘hobby farming’ type of management)就是下一个优先入选的变量，此变量如果被选，那么这三个变量所能解释的变量变成了0.760(0.624+0.136)。

如何来判断备选环境因子的解释能力呢？我们可以使用偏蒙特卡罗置换检验（partial Monte Carlo permutation test）来评估。在偏置换检验中，我们可以将每个候选的变量作为唯一的变量进入分析（此时排序模型只有一个轴），同时，将已经被选定的环境变量作为协变量，排序模型进行蒙特卡罗置换检验。如果我们能够拒绝零和假设，表示我们可以将该环境因子选入分析中。

上面这种用偏蒙特卡罗置换检验来评估备选环境因子的解释能力是一种有限制条件的解释量，是去除已经入选的环境因子的解释量后的所能解释的变化量。但在开始选择第一个变量的时候，那时并没有入选的变量，我们先可以分别将每个因子作为独立的变量进行分析，将单独解释能力（marginal effect）最强的作为第一个入选的变量。

5.10 变量分解方法（Variance partitioning procedure）

在前面一节，我们已经解释了关于单个环境变量的条件（conditional）解释量和边缘（marginal）解释量。这两种解释量的差异可以判断环境变量之间的交互作用（即解释量重叠的部分）。比如一对变量A和B重叠解释部分可以表示为A的marginal effect 和它的conditional effect差别。

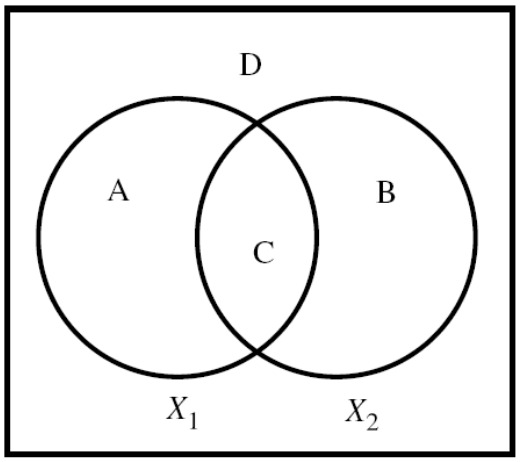

Figure 5-7. Partitioning of the variance in the species data into the contributions of two subsets of environmental variables (A, B, and the shared portion C) and the residual variance (D).

这个概念的基础来源于方差分解（variance partitioning）。实际上，我们通常不是量化两个或多个环境变量的单独或重叠解释量，而是两组或是多组的解释量。最典型的变量是时间和空间两组变量之间的单独和重叠的解释量。

以下，我们将用两组环境因子变量（X1和X2）的最简单的例子来描述解释量分解的过程。每组变量包含1个或是多个单变量。图5-7展示的是物种变化量可以分解成为很多部分。

图中D表示不能被排序模型（两组环境因子）所解释的部分。A表示单独被X1解释部分。B表示单独被X2解释的部分，C表示能被X1和X2共同解释部分）。如果忽略X2,那么能被X1解释的变量应该A+C。我们可以通过偏约束分析来计算出A、B、C、D各部分的数量。

将X1作为主环境变量，X2作为协变量，得出的解释量是A。同样，如果X2做主变量，X1做写变量，就可以得出B。C可以通过X1和X2一起作为环境变量时的解释量减去A和B得到。C偶尔会是负值，表示两组变量的交互效应会大于它们各自能解释量之和。具体更详细的内容，可以参见Legendre编写的《Numerical Ecology》(1998)第533页。