

多元统计方法在节肢动物群落分析中的应用^{*}

杨 兵^{1, 2} 戈 峰^{1 **}

(1. 中国科学院动物研究所 农业虫害鼠害综合治理研究国家重点实验室 北京 100101;

2. 中国科学院成都生物研究所 恢复生态学重点实验室 成都 610041)

摘 要 近年来有关节肢动物群落的研究报道较多,而传统的、简单的多样性指数分析远远不能满足对高通量群落数据的挖掘。采用多元统计方法可以较好地分析群落数据。其中,在研究某一处理的长期效应时,主响应曲线是一种经典的多元分析方法;而非度量多维尺度分析具有能处理较大量的数据且常能以低维排序图反映结果的优点,是一种具有广泛应用前景的排序技术。本文以 Bt 棉与非 Bt 棉不同比例混作棉田的节肢动物群落为例,详尽地介绍了主响应曲线和非度量多维尺度在节肢动物群落分析中的应用。

关键词 多元分析, 主响应曲线, 非尺度多维尺度, 节肢动物群落分析

The applications of multivariate analysis in arthropod community analysis

YANG Bing^{1, 2} GE Feng^{1 **}

(1. State Key Laboratory of Integrated Management of Pest and Rodents, Institute of Zoology,

Chinese Academy of Sciences, Beijing 100101, China; 2. Key Laboratory of Ecological

Restoration, Chengdu Institute of Biology, Chinese Academy of Sciences, Chengdu 610041, China)

Abstract Increasing numbers of studies of arthropod communities have been made in recent years, but comprehensive information derived from community analyses is limited by the use of traditional and simple diversity indices. Multivariate analysis methodology is particularly well suited for meeting specialized needs for high-throughput community data mining. The principal response curve is a classic multivariate analysis method which is widely used in environmental monitoring and ecological restoration to explore the long-term effects of certain human activity. Non-metric multidimensional scaling analysis can deal with large amounts of data and show the results in a low-dimensional ordination, making it a promising method for analyzing community ecology. However, these methods have rarely been used in arthropod community analysis in China to date. Here, the uses of principal response curve and non-metric multidimensional scaling are introduced in a study of arthropod communities from Bt and non-Bt mixture cotton fields with different acreage ratios.

Key words multivariate analysis, principal response curve, non-metric multidimensional scaling analysis, arthropod community analysis

当前节肢动物群落的研究报道越来越多,而传统的、简单的多样性指数分析远远不能满足对高通量群落数据的挖掘。群落数据通常由一系列物种的丰度或生物量数据构成,包括一个或多个重复样本,它们取自:(1)同一时间的不同地点(空间分析);(2)同一地点的不同时间(时间分析);(3)不同试验处理的群落;(4)以上的组合。由于

物种×样本矩阵往往十分庞大,群落的结构不能一目了然。同时,这些数据本身具有多维性。因此,采用多元统计的方法来分析群落数据较为适合(周红和张志南,2003)。

群落数据的多元统计分析就是通过对样本之间生物关系的某种图形表达,以减少这些矩阵的复杂性;然后通过统计检验来确认群落结构的时

^{*} 资助项目:国家自然科学基金(31200321)和国家科技支撑计划项目(2012BAD19B05)。

^{**}通讯作者, E-mail: gef@ioz.ac.cn

收稿日期:2013-06-13, 接受日期:2013-06-30

空变化,并将这种变化与变化的环境或实验条件联系起来。换言之,就是将群落原始数据中的响应变量(response variable)和解释变量(explanatory variables)之间复杂的关系通过一定的形式表示出来。其中,群落中的响应变量经常是物种的组成数据,而解释变量通常是环境因子,比如作物品种、农事措施、化学防治等特征属性。

当只关注群落变化的连续性时,群落排序便是很好的方法。在排序分析中,解释变量又经常可以分为两组:一类主环境变量,即所关注的是它们与群落内物种分布的关系;另外一组叫协环境变量(在一般统计方法里面也叫协变量 covariates)。协环境变量与主环境变量同时对响应变起作用,因此在分析主环境变量效应之前,应先将协环境变量的效应剔除出来,以便更准确考察主环境变量与物种分布的关系。目前,在研究某一处理的长期效应时,主响应曲线是一种经典的多元变量分析方法,广泛运用于环境监测和恢复的群落生态学研究;而非度量多维尺度分析具有能处理较大量的数据且常能以低维排序图反映结果的优点,是一种具有广泛应用前景的排序技术。然而,上述研究方法很少应用到节肢动物群落分析中。本文以不同 Bt 棉与 Bt 棉比例棉田的节肢动物群落为例,重点介绍了主响应曲线和非度量多维尺度在节肢动物群落分析中的应用。

1 主响应曲线 (principal response curve)

通过比较试验处理(如品种、农事措施、化学防治)与对照中某种生物群落在不同取样时间的组成差异,可阐明某种试验因素或措施对群落特征的影响。然而,标准的约束排序图却很难满足。为了更好地展示处理与对照在不同时间的差异, Van den Brink 和 Ter Braak (1998, 1999)开发出了—种被称作主响应曲线(principal response curve, PRC)的新方法。主响应曲线的主要结果是一条或多条的表示不同试验处理下群落随时间的变化的曲线。在通常情况下,表示各处理群落组成的时空轨迹的主响应曲线图和显示物种在相应的冗余度分析(redundancy analysis, RDA)轴的得分的一位排序图常相互补充。因此,可利用 PRC 和物

种得分,获得处理和对照在特定时间的差异;通过计算经转换过的指数的乘积,便可以预测同一时间对照样地的该物种丰富度的相对大小。

特别注意的是,这种定量解释要求对原始的物种数据进行对数转换,而且物种数据必须是计量资料。其中,RDA 的第一轴回归系数的剩余相互作用项代表了特定处理和对照在特定时间的差异程度。回归系数必须转化,但无需标准化。回归系数的得分的具体转化公式可在 CANOCO 使用参考(Ter Braak and Šmilauer 2002)中找到,所需的计算可以通过 Canodraw Windows 运行。具体操作如下: Using the project > Import variables > Set up PRC scores command。有时可见有许多的源于 RDA 第一轴的响应曲线,因为除了对照可能还有一系列处理。

由于实验处理在不同的采样时间的效应通常无法由一条约束轴表示,因此常需要构造第二条甚至一系列的主响应曲线。其中,用一条和额外的响应曲线需要通过基于第一个典范轴的置换检验进行确定。在置换检验时,随机分配整个时间序列的试验处理,但取自同一样地的不同采样时间的样本在每次置换时需要一起被置换。为了确保置换时取自同一处理但不同时间的样本都在一起,最简单的方法是在 CANOCO 程序中选设定取自同一样地的所有样品作为整区进行置换。

下面以 Bt 棉与非 Bt 棉不同比例(0, 75% 和 100%)混作棉田的节肢动物群落为例,简要说明主响应曲线的分析步骤。

(1) 物种丰度矩阵与环境矩阵构建

首先将要分析的 Bt 棉与非 Bt 棉不同比例混作棉田的节肢动物群落数据在 Excel 整理成如下格式(图 1:A, B):每行代表一个处理的重复,每列为一个物种或者环境因子,例如,40 个物种就有 40 列数据。

(2) 打开物种丰度数据和环境因子所在的 Excel 表格,在 Excel 里面对数据进行选择复制,然后在打开的 WCanolmp 软件里面按图 2 设置,并选择“save”。最后形成两个文件,分别为物种数据和环境因子数据。

(3) 选定物种丰度数据,环境数据和协变量数据矩阵,并指定文件名为 2009lfa. sol, 点击下一步(图 3)。

(4) 选定使用线性直接梯度分析的 RDA(图

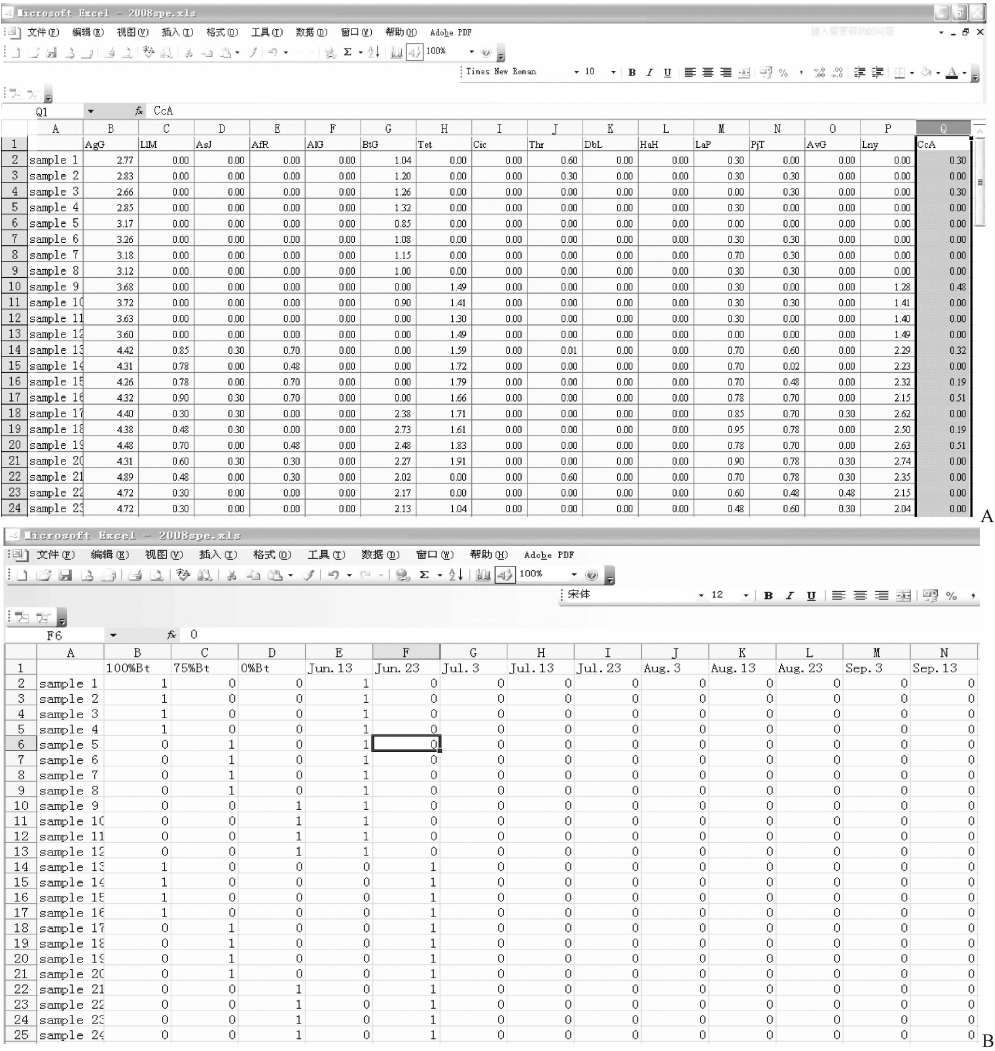


图 1 物种丰度数据矩阵(A)和环境因子数据矩阵(B)

Fig.1 Species abundance metric (A) and environmental factor metric (B)

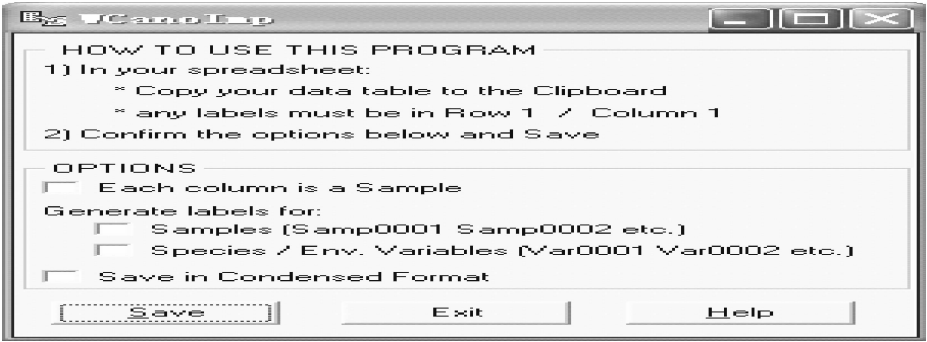


图 2 通过 WcanoImp 将物种丰度数据和环境因子数据导入 Canoco 软件

Fig.2 Input species abundance data and environmental factors into Canoco with WcanoImp procedure

4:A),点击下一步,出现如图 4(B)的对话框。在 Focus scaling on 区域选择“inter-sample distance”复选框,而在 species scores 区域选择“Do not post-

transform”。待出现“Transformation of Species Data”对话框时,选择“Log transformation $Y' = \log(A^*Y + B)$ distance”复选框(图 4:C),点击“下一步”。在

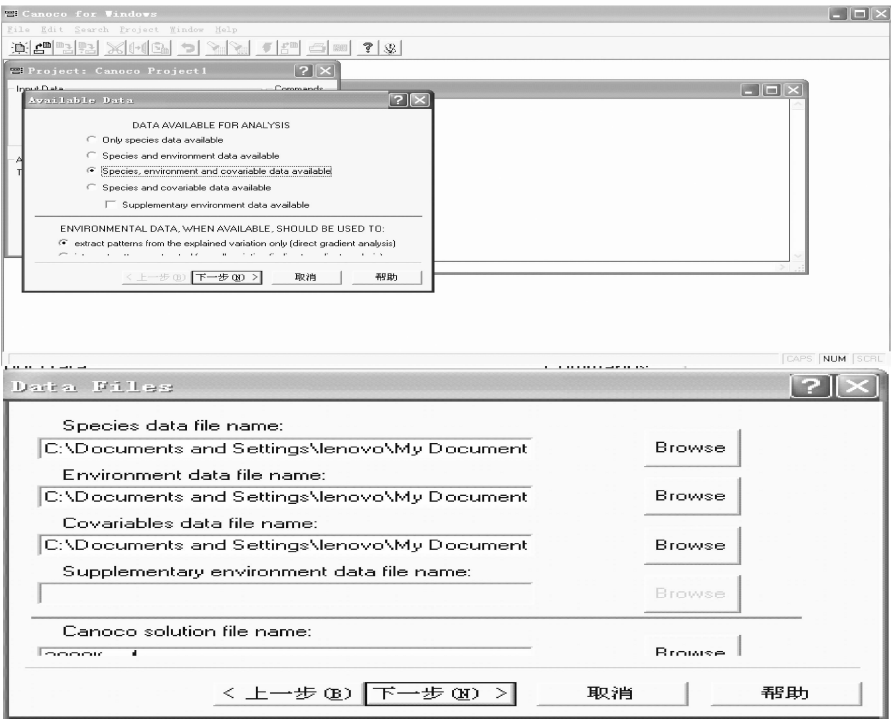


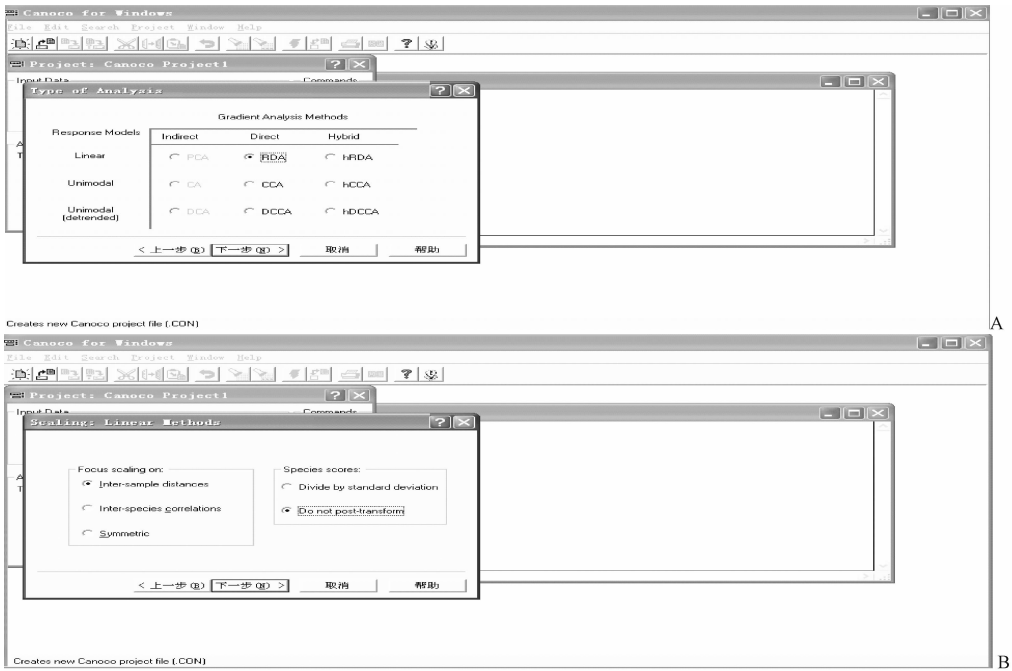
图 3 指定待分析的数据类型 (A) 并将物种丰度数据,环境数据和协变量数据矩阵导入 Canoco 软件 (B)

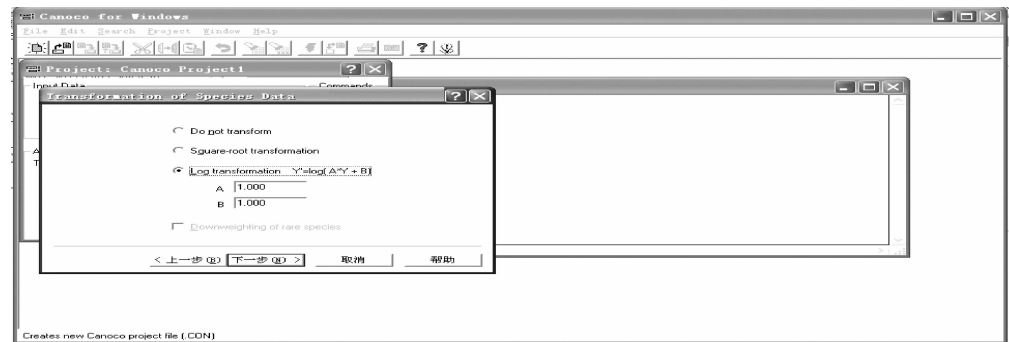
Fig.3 Assign data type for analysis (A) and input the file of species data, environmental data and covariables data to Canoco (B)

此之后按图 4(D ~ O) 进行。

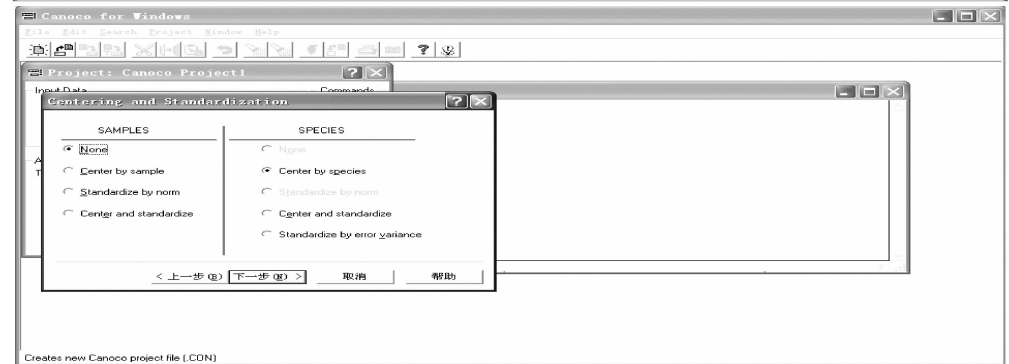
(5) 点击 Analyze, 出现如下对话框, 查看第一轴的特征值及差异显著检验结果, 点击“Save

log”, 出现如下对话框, 指定文件名, 并点击保存, 如图 5。

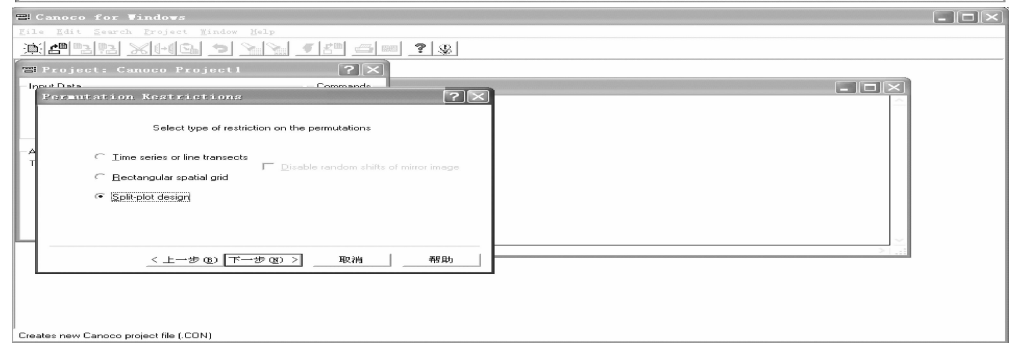




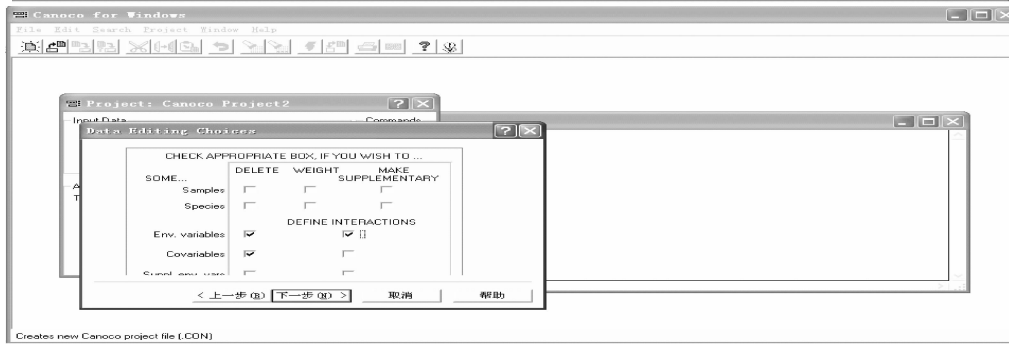
C



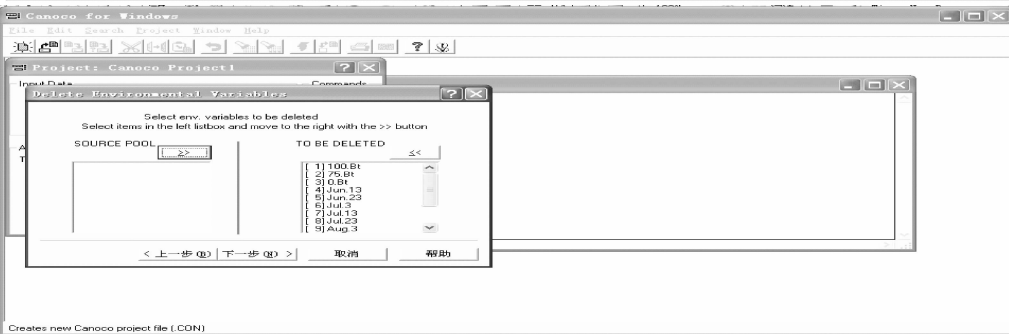
D



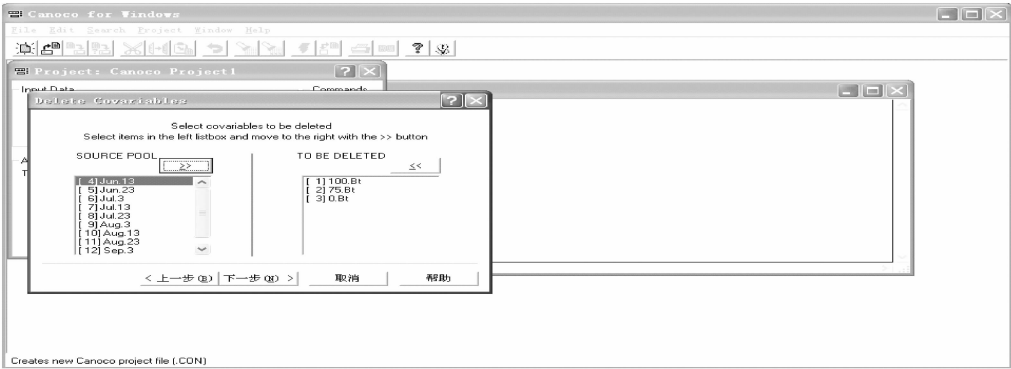
E



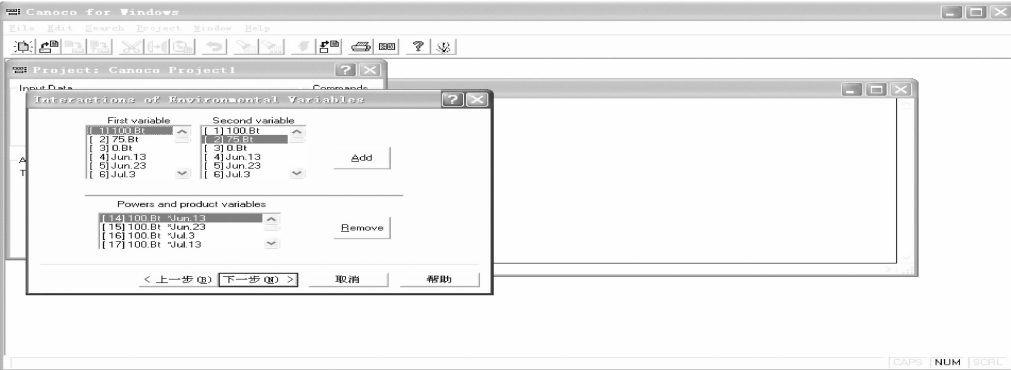
F



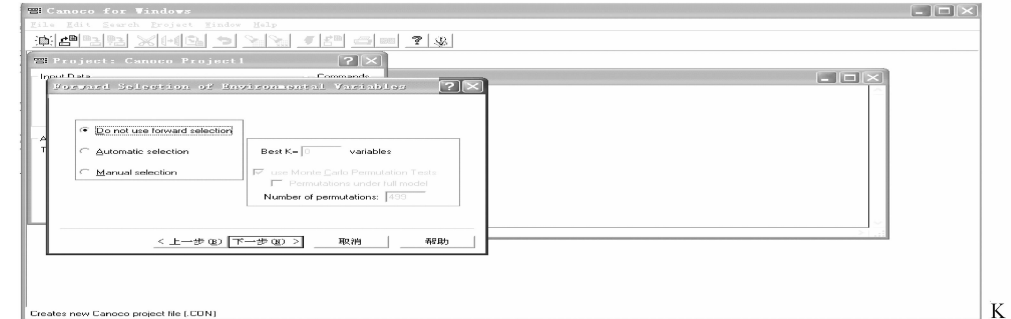
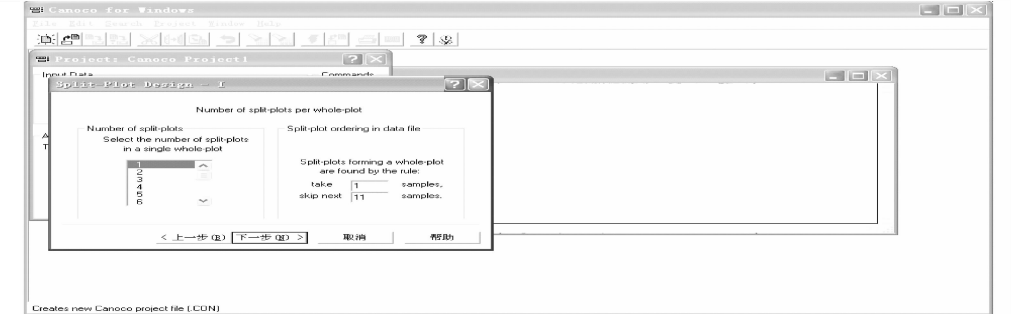
G



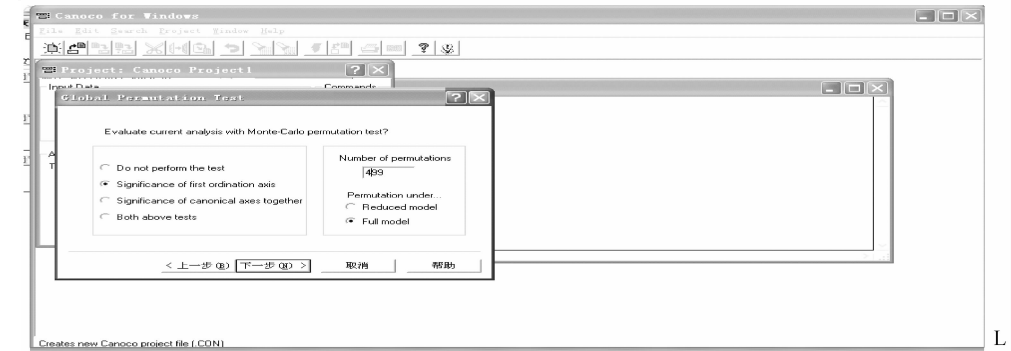
H



I



K



L

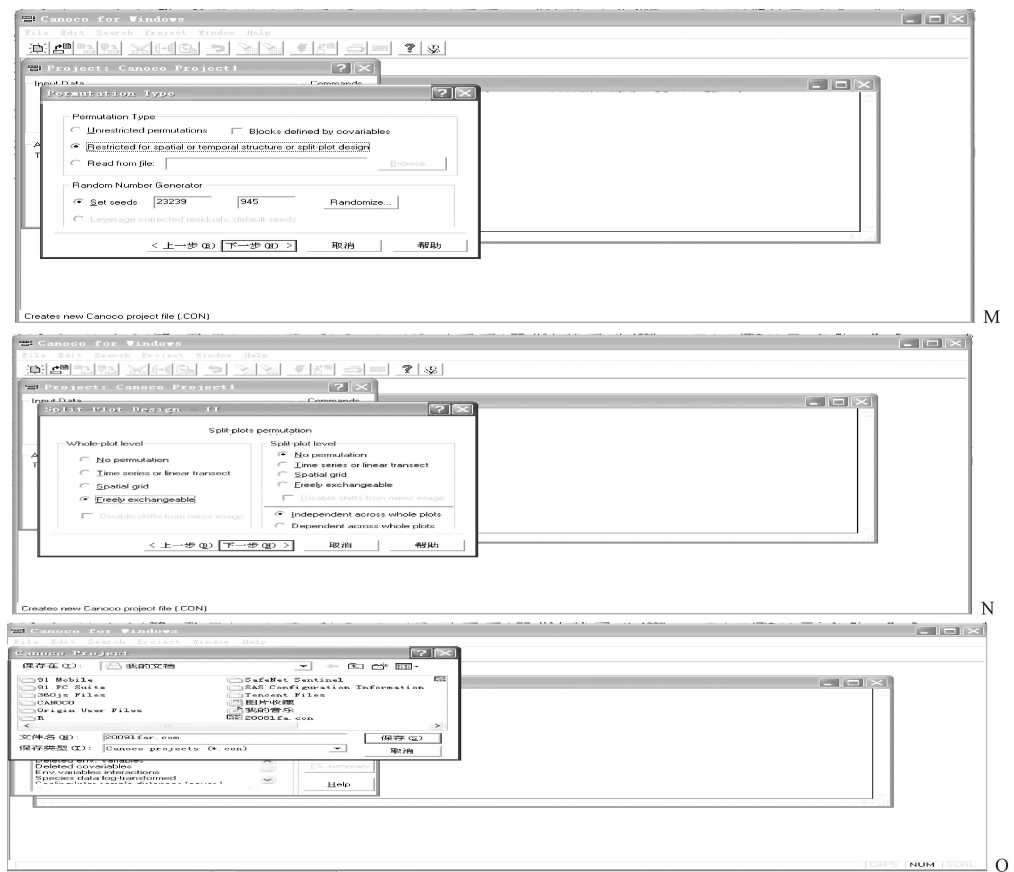


图 4 指定数据分析方法 (A), 数据转化方式 (B, C)、样本和物种数据中心化和标准化处理方式、样本数、显著性和置换检验对象后计算 PRC 并将分析结果保存在指定的文件夹中 (D ~ O)

Fig. 4 Compute the PRC scores after assigning analysis method (A) data transformation pattern (B, C), centering and standardization for sample and species, sample size, significant test for the first PRC axis and permeation test and save results in an assigned file(D-O)

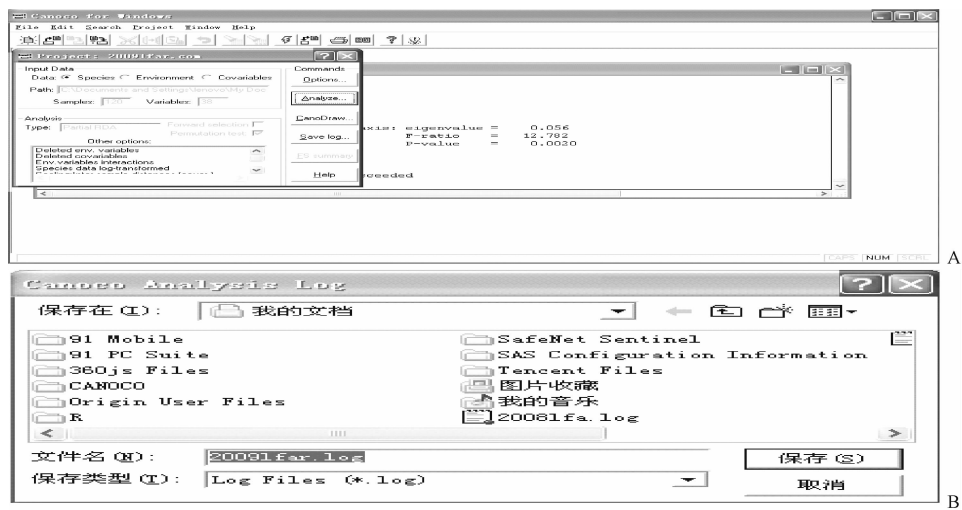


图 5 运行 Analyze 并将第一轴的特征值 (A) 及差异显著检验结果保持到指定文件中 (B)

Fig. 5 Analyze PRC and save the results of eigenvalue (A) and significant test for the first PRC axis into an assigned file(B)

(6) 点击“CanoDraw”, 出现如下对话框, 指定文件名, 并点击保存。

(7) 运行“Project > Import variables > Setup PRC scores”指令。待出现如下对话框时(图 6), 选定列表框中的所有项目, 如图蓝色区域所有项目。通过“Browse”按钮打开 RDA 日志文件“2009lfar. log”。点击“OK”, 出现对话框, 单击“确定”。

(8) 运行“View > Project Details”指令, 双击“Project Details > Labels”文件的 EnvV. Lables 文件夹, 单击鼠标右键弹出如下对话框(图 7)。

(9) 在“Variable summary”对话框点击“Copy”将环境因子变量的指数和标签全部复制到剪贴板。新建一个 Excel 表格, 将剪贴板上的数据粘贴到表格中。然后按具体的日期替换第二列中的内容(图 8)。

(10) 将转换后的数据复制到剪贴板后关闭 Excel 表格, 返回 CanoDraw 程序。

运行“Project > Import variables > From Clipboard”指令, 出现如下对话框。选择 Environm. variable, 然后点击“Import”按钮(图 9)。

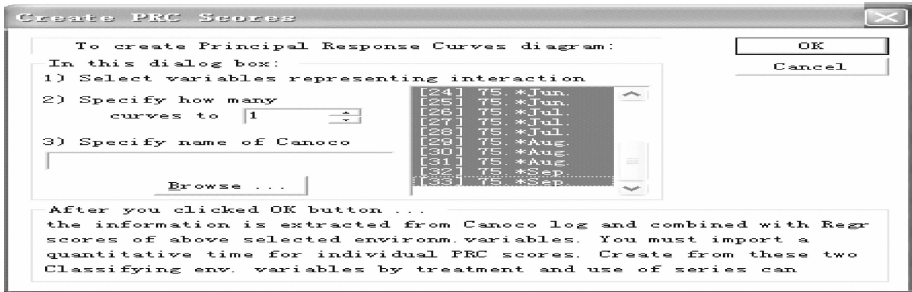


图 6 运行 Create PRC Scores 指令
Fig.6 Create PRC Scores in Canoco

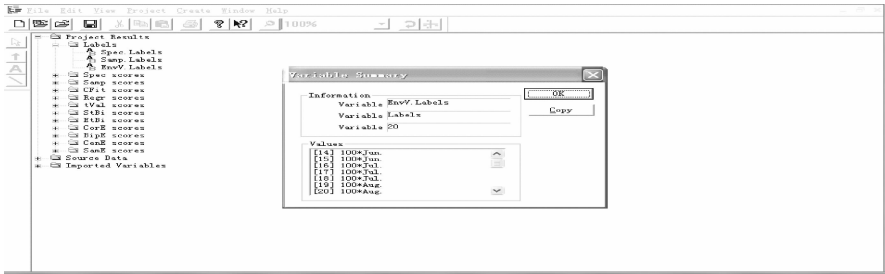


图 7 复制环境变量标签到剪贴板
Fig.7 Copy labels of environmental factors to clipboard

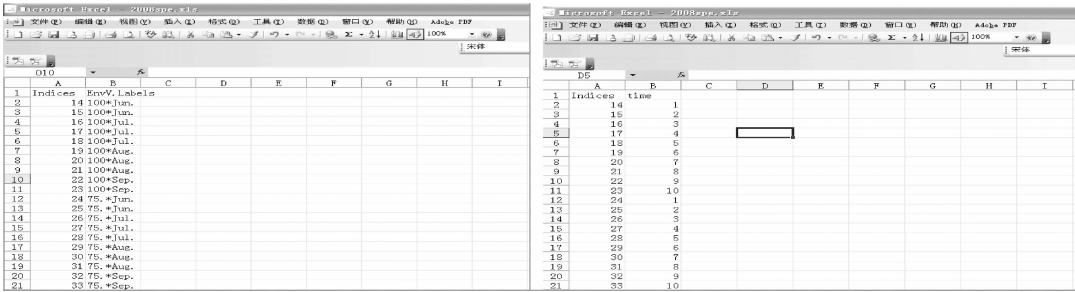


图 8 在 Excel 表格中将环境因子重新赋值
Fig.8 Reassign environmental factors in excel

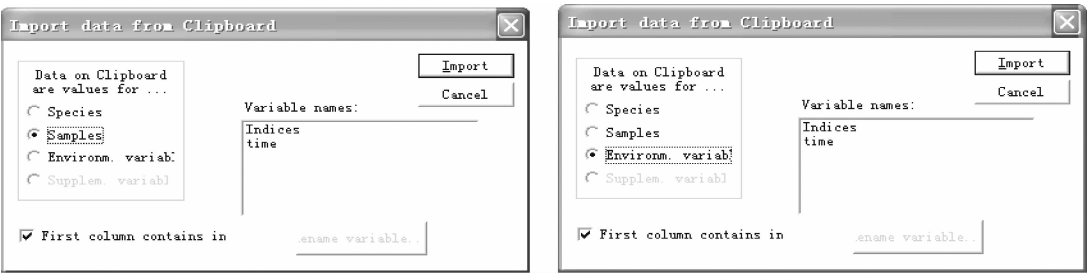


图 9 输入并调用环境因子指数

Fig.9 Input and assign the indices of environmental factors

(11) 运行“Project > Classify > Environm. variables”指令,然后点击“New select”按钮。在“Manual classification”对话框,根据取样日期重新定义 10 个分类变量,注意同一日期的处理和对照同时放入同一分类变量下。定义完分类变量后,点击 OK 关闭对话框。在弹出的对话框中勾选“Use this classification in diagrams”选项。

(12) 运行“Project > Define Series of > Environ. variables”指令,弹出如下对话框,选择“From Class”选项并点击“Close”结束。然后在弹出的对话框中直接点击“OK”。

(13) 运行“Create > Attribute Plots > XY (Z) Plot”指令,在“X VARIABLE”区域选择“Import data > For env. Variables > Time”,同时按相同方法在

“Y VARIABLE”区域选定 PRC1 变量,然后将“PROPERTIES > Labeling”和“VISUALIZATION MODEL”设定为“None”。完成上述设置后单击 OK,创建排序图。

(14) 选择“create > Attribute Plots > XY (Z) Plots”指令,在“X VARIABLE”区域选择“Analysis Results > Species Scores > constant”,同时按相同方法在“Y VARIABLE”区域选定 Analysis Results > Species Scores > Spec. 1,其他选项按自定义。完成上述设置后单击 OK,创建物种得分排序图。

通过对图 10 分析的结果可知,在占 75% 和 100% Bt 棉田,蚂蚁和叶蝉的数量显著高于常规棉田,而叶螨,寄生蜂以及食蚜蝇的数量显著低于常规棉田。

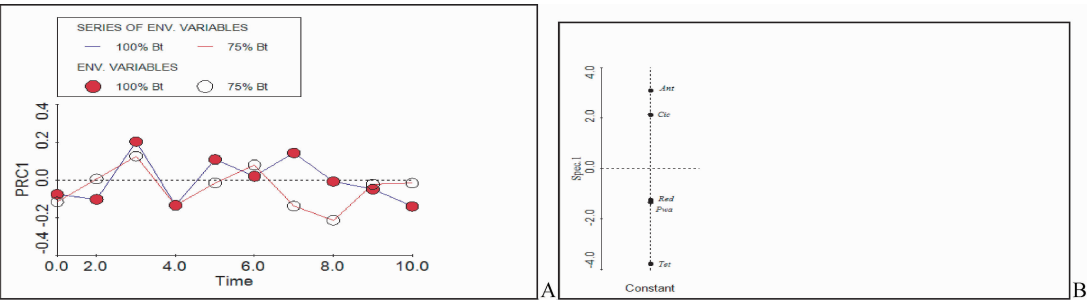


图 10 Bt 棉与非 Bt 棉不同面积比例混作棉田节肢动物群落的主响应曲线排序图 (A) 和物种得分排序图 (B)

Fig.10 Principal response curve (A) and species scores ordination (B) of arthropod communities from cotton fields of different Bt to non-Bt acreage ratios

2 非度量多维测度法

非度量多维测度法 (non-metric multidimensional scaling, NMDS) 是由 Shepard (1962a, 1962b) 提出,经 Kruskal (1964a, 1964b)

修改,最终由 Shepard (1962a, 1962b) 和 Kruskal (1964a, 1964b) 共同改进并运用于尺度和因子分析,适用于非线性数据结构分析的一种复杂的迭代排序方法。它的基本思想是通过排序,将 n 个实体在尽可能低维 ($t < n$) 排序空间上之间的距离

与实体之间的实际相异性一致(余世孝,1995)。因此,其排序结果明显优于其它分析方法。同时,非度量多维测度具有能处理较大量的数据且常能以低维排序图反映结果的优点,是一种具有广泛应用前景的排序技术。但其缺点是要求事先确定排序维数,即使应力系数值最小;因为同时获得较低的维数和较低的应力系数常相互矛盾。

NMDS 分析结果的优劣用应力系数(stress index)来衡量。通常认为当应力系数<0.2 时,可用 NMDS 的二维点图表示,其图形有一定的解释意义;当应力系数<0.1 时,可认为是一个好的排序;当应力系数<0.05 时,具有很好的代表性(马藏允等,1997;李圣法等,2005;王雪辉等,2013)。

下面以不同 Bt 棉与非 Bt 棉不同面积比例(0,

75% 和 100%)混作棉田的节肢动物群落为例,简要说明 NMDS 的分析步骤。

① 构建物种数据和环境数据的原始矩阵

将野外调查所获得的不同 Bt 棉比例棉田的节肢动物群落数据直接输入 Excel 工作表,形成一个*i* 行(物种)和*j* 列(样本)的原始矩阵,输入值为每个样本中各物种的丰度,称为物种丰度矩阵。在实际运用中,矩阵样本可以包括一个或多个取自不同时间、地点或者处理的重复。本例中为同一时间点 Bt 棉与非 Bt 棉不同面积比例混作棉田中节肢动物群落各组成成分的丰度数据。Excel 中的数据复制到剪贴板上,并将剪贴板上的内容粘贴到先建“lf.txt”文件中(图 11)。

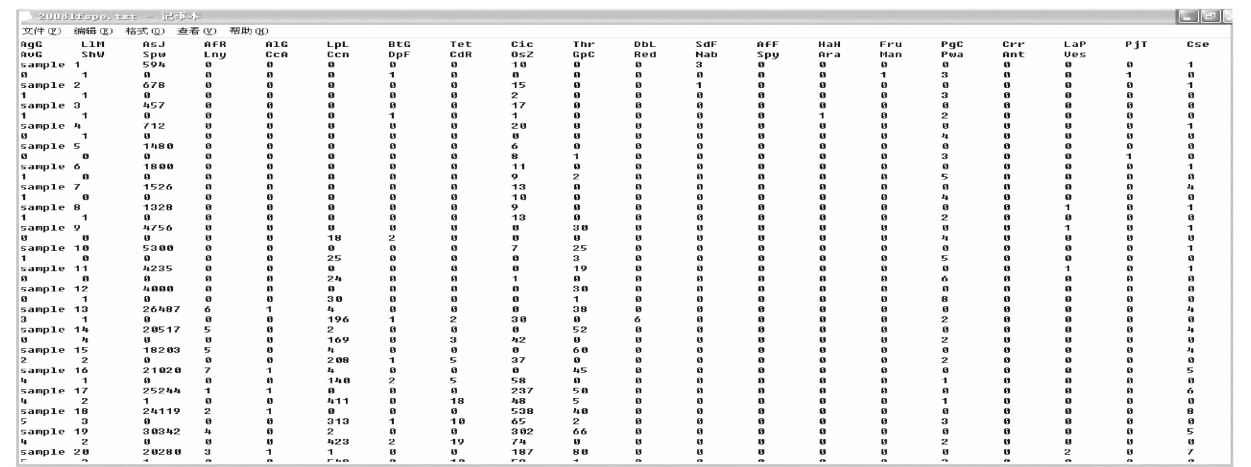


图 11 不同 Bt 棉面积比棉田节肢动物群落组成数据矩阵

Fig.11 Species composition metrics for arthropod communities from Bt and non-Bt mixture cotton fields of different acreage ratios

②在 R 2.14.1 中运行如下程序,进行排序分析

显示物种名的 NMDS 程序如下:
library(vegan) # 调用程辑包 Vegan
data = read.table("lf.txt", header = T) # 通过 read.table() 语句直接从“lf.txt”调用数据集,并将数据集命名为“mydata”
Global NMDS using monoMDS #指定非尺度多维测度方法的类型为正态的单调回归 NMDS
sol <- metaMDS(mydata) #调用“mydata”数据集,运算样本之间的 Bray-Curtis 相似性系数
sol # 显示排序结果

plot(sol, type = “t”) # 以物种和样本为点作图

```
## Start from previous best solution  
sol <- metaMDS(mydata, previous.best = sol)  
## Local NMDS and stress 2 of monoMDS #  
“local” is non-metric MDS with separate regressions for each point  
sol2 <- metaMDS(mydata, model = “local”, stress = 2) # Use stress type 1 or 2  
sol2  
## Use Arrhenius exponent ‘z’ as a binary
```

dissimilarity measure

```
sol <- metaMDS ( mydata, distfun =  
betadiver, distance = "z" )
```

sol

```
plot(sol, type = "t") # The type of the plot: "  
t" for text, "p" for points, and "n" for none.
```

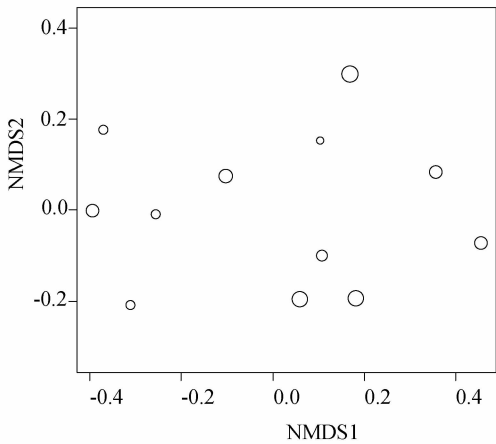
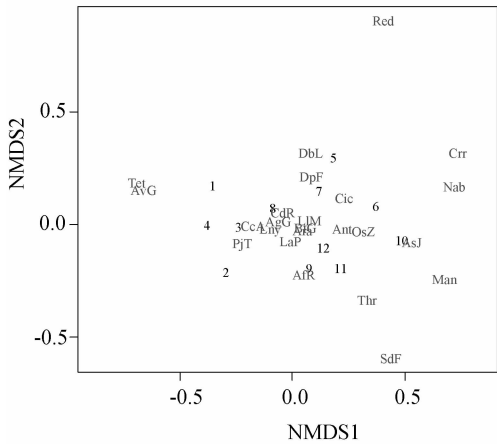


图 12 Bt 棉与非 Bt 棉不同面积比例混作棉田节肢动物群落非尺度多维测度排序图
Fig. 12 Non-metric multidimensional scaling ordination for arthropod communities from Bt
and non-Bt mixture cotton fields of different acreage ratios

上述图 12 中,1,2,3,4 表示 100% Bt 棉田样地;5,6,7,8 表示 75% Bt 棉田样地,而 9,10,11,12 表示常规棉田样地。Stress index = 0.1267,即不同 Bt 棉比例棉田节肢动物群落结构组成有显著差异。

```
只显示样方位置的 NMDS 程序  
library(vegan) # 调用程辑包 Vegan  
data = read.table("lf.txt", header = T) # 通过 read.table() 语句直接从 "lf.txt" 调用数据集,并将数据命名为 "mydata"  
data(mydata) #调用数据集 "mydata"  
mod <- metaMDS(mydata)  
stressplot(mod)  
gof <- goodness(mod)  
gof  
plot(mod, display = "sites", type = "n") #  
The type of the plot: "t" for text, "p" for points, and "n" for none.  
points(mod, display = "sites", cex = 2 * gof/mean(gof))
```

3 讨论

在研究某一处理的长期效应时,主响应曲线

是一种经典的多元分析方法。它不仅清晰地展示出处理与对照之间群落组成的差异,还能显示导致这种差异的物种及各个物种贡献大小。因此,主响应曲线可用于指示生物的筛选,甚至其意义远远高于生物指示值方法。

与聚类分析一样,NMDS 由样本间的相似性三角矩阵开始。在进行相似性分析之前,须对原始数据进行转换 (transformation) 以便对稀有种给予不同程度的加权。具体做法是先将原始数据标准化 (standardization),并去掉部分稀有种,只保留那些在任一样品中的丰度 (或生物量) 占该样品总丰度至少为 P% 的种 (P 一般取 3 ~ 10) (周红和张志南,2003)。但聚类分析和 NMDS 的目的和适用范围也不相同。聚类分析的目的在于找出样本的“自然分组”以使组内样本间较组间的样本更相似,它强调的是组的划分,而不太关注样本间在连续尺度上的关系;而非度量多维测度按照样本间的相似性或者非相似性秩次将样本排在二维的排序图中,并通过应力系数反映相似性 (或者不相似) 等级与排序图中相应的距离等级的不一致程度。此外,同时显示样本和物种的 NMDS 能清晰地展示样本间的相似或不相似程度及导致样本之间相似或不相似性的特征种。因此,NMDS 能更好

地表达群落对于比较连续的非生物环境梯度的响应。但在实际运用中,聚类图和排序图的组合(即将在某种相似性水平上得到的聚类分组叠加在以同一种相似性系数为基础的排序图上)能更充分地展示群落结构格局。虽然目前 CANOCO 软件, SAS、SPSS 及 R 均能分别进行聚类分析和 NMDS, 但聚类图和 NMDS 排序图的组合目前只能在大型多元统计软件 PRIMER (Plymouth Routines In Multivariate Ecological Research) 进行。

参考文献 (References)

Kruskal JB, 1964a. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. *Psychometrika*, 29(1):1 – 27.

Kruskal JB, 1964b. Nonmetric multidimensional scaling: a numerical method. *Psychometrika*, 29(2):115 – 129.

Lepš J, Šmilauer P, 2003. Multivariate Analysis of Ecological Data Using CANOCO. Cambridge University Press. 144 – 146.

Shepard RN, 1962a. The analysis of proximities: Multidimensional scaling with an unknown distance function. I. *Psychometrika*, 27(2):125 – 140.

Shepard RN, 1962b. The analysis of proximities: Multidimensional scaling with an unknown distance function. II. *Psychometrika*, 27(3):219 – 246.

Shepard RN, Kruskal JB, 1964. Nonmetric methods for

scaling and for factor analysis. *Am. Psychol.*, 19:557 – 558.

Ter Braak CJF, Šmilauer P, 2002. CANOCO reference manual and user’s guide to Canoco for Windows: software for canonical community ordination (version 4.5). Microcomputer Power, Ithaca, New York, USA.

Van den Brink PJ, Ter Braak CJF, 1998. Multivariate analysis of stress in experimental ecosystems by principal response curves and similarity analysis. *Aquat. Ecol.*, 32(2):163 – 178.

Van den Brink PJ, Ter Braak CJF, 1999. Principal response curves: analysis of time-dependent multivariate responses of biological community to stress. *Environ. Toxicol. Chem.*, 18(2):138 – 148.

李圣法, 程家骅, 严利平, 2005. 东海中南部鱼类群聚结构的 空间特征. *海洋学报*, 27(3):110 – 118.

马藏允, 刘海, 王惠卿, 王世权, 1997. 底栖生物群落结构 变化多元变量统计分析. *中国环境科学*, 17(4):297 – 300.

王雪辉, 林昭进, 杜飞雁, 邱永松, 孙典荣, 王跃中, 梁 新, 黄硕琳, 2013. 南海西北部陆架区鱼类的种类组成 与群落格局. *生态学报*, 33(7):2225 – 2235.

余世孝, 1995. 非度量多维测度及其在群落分类中的应 用. *植物生态学报*, 19(2):128 – 136.

周红, 张志南, 2003. 大型多元统计软件 PRIMER 的方法 原理及其在底栖群落生态学中的应用. *青岛海洋大学 学报 (自然科学版)*, 33(1):58 – 64.