生态学
数据库的简介

〔加拿大〕E. C. 皮洛著
石绍业
等 译
陈华豪

东北林业大学出版社

生 态 数 据 的 解 释

E.C. 皮洛 著

石绍业 陈华豪 等译

东 北 林 业 大 学 出 版 社 出 版

（哈尔滨市和兴路8号）

黑龙江省新华书店发行　东北林业大学印刷厂印刷

开本 787×1092 毫米 1/32　印张 9　字数 187 千字

1 9 8 6 年12月第1版　1 9 8 6 年12月第1次印刷

印数 1—3,000 册

统一书号 13447·003　定价 1.50 元

序　　言

本书的目的是对群落生态学家所应用的大量野外数据的方法，作一个充分而详细的引论性介绍。这些方法可以把含有多元变量的、庞大而又不易运用的野外数据处理成为综合性的和可以解释的。我深信，出这样一本书是必要的。现在已有几本论述同样材料的某些内容更为高深的书籍，如L. Orlóci的《植被研究中的多元分析》（W.Junk，1978）和A. D. Gordon的《分类——多元数据的探索性分析方法》（Chapman and Hall，1981），但是这些书对读者的数学能力的要求都比本书高。也还有一些对这一材料的更为一般性的论述，如H. G. Gauch的《群落生态学中的多元分析》（Cambridge University Press，1982），或者是R. H. Whittaker主编的二卷集《植物群落排序和植物群落分类》（W. Junk，1978）中的许多章，但是他们更关心的是一般原理和比较不同方法的优点，而不是解释每一种技术的实际的详细内容。

这样的解释是极其需要的。这个世纪以来，生态学家对数字处理已经使用过一系列辅助手段，从对数表、手摇计算机到电动计算机，然后是从台式电子计算器到可编程序的计算机。对生态学家来说，没有必要了解这些辅助手段是如何工作的。但近十年来，出现了一种新的“拐杖式”手段，即计算机程序包。这些计算机程序包使得生态学家可以不必编

出他们自己的程序，即可对他们的数据进行复杂的分析。程序包中的程序常常是长而复杂的，这是计算机专家们的作品。不要求生态学家求助于这些专家是不合理的。

然而，要求生态学家即使不懂得这些程序是如何工作的，也应该懂得这些程序将为他们做些什么，并不是不合理的。有两种人，一种人使用已编好的程序求高阶矩阵的特征值和特征向量，并且懂得这些量的含义；一种人把进行主分量分析的整个任务编成这样一个程序，而他不懂得这种分析是做什么的。在这两种人之间的差异是巨大的。程序包的优点多于缺点，它既能对大量数据作快速而又精确的分析，并且是以揭露其生态学含意的方式进行，又可使那些训练不足的人有可能在不理解的情况下通过数据分析的过程。

对那些想要完整理解分析多元数据最常用技术的人有所帮助，是本书的设想。本书不提供任何计算机程序。取代它的是用虚拟的数据解释各种技术，数据都是简单的，只要能了解从头到尾详细进行分析的所有步骤就足够了。所以必须予先具备的条件仅是相当于大学一年级学生的基础代数和解析几何知识。为使本书对自学有用，各章末都列有习题。书末有习题答案和一个综合词汇表。

我是在 Lethbridge 大学的艾伯塔油沙技术研究所担任研究教授时写这本书的。我十分感谢油沙技术研究所和大学，由于他们的支持才有可能写这本书。我还要感谢 Lethbridge 大学的 William Smienk，他绘制了全部插图。

E. C. 皮洛

加拿大 艾伯塔省 Lethbridge

1984年4月

目　　录

第一章　引　　论

大概所有生态学家都熟悉野外手册，它的各页形如图1.1。可能大多数生态学家，即使还是刚刚从事这项工作的生态学家，都曾记录过这样的表。他们的成果或者是较为整洁的，或者是较为杂乱的，取决于人和各种环境情况（风、雨、蚊子、快黑的天、正在涨的潮，或生态学家所受到的其它压力中的任何一种）。但是这类表原则上是相同的。表中列出了若干抽样单元（即样方）中每一抽样单元的若干变量（即物种的数量）的每个变量值。这样的表格是群落研究与分析的直接原始材料。

虽然生长在野外的天然的生物群落是生态学家们最根本的原始材料，但不首先用符号来表示它们那就不可能真正理解掌握它们。图1.1那样的表是天然群落的一种典型的符号表达方式，它是数据矩阵的一部分。这是最初的表达式，所有后续的分析及它们的表达式都是从这里推演出来的。因此，它是从观测到的实际群落到有关群落的理论，甚至可能到有关生态群落的一般性理论的链中的第一个环节。

解释这样的数据矩阵是本书的主题。本章首先一般地论述使数据解释成为可能的程序，接着一节讨论几个术语，作为所有后续内容所必须的基础知识。对于一门迅速发展中的学科来说，不可避免会出现不同作者对少数术语按不同的含义使用。因此有必要明确规定本书中使用这些术语的含义。这

July 18 Plot 6 (Rt bank, c 300 m S of mouth of Steepbank R., 40 m inland)
(P. 2 of plot 6 Quads 11–20

PLOT 6

Sp	Quad #11	#12	13	14	15	16	17	18	19	20
Equisetum sp.	4	–	1	2		7	10	13	18	17
Rubus pubescens	11	4	13	18	4	7	17		13	2
R. triflorus	1	8	1	2	19	8	3	5	2	8
Cornus stol.	6	–	–	1			1	1	/	1
C. canadensis	–	–	2	–	12			1	/	/
Rosa acic.	2	2	1	6	11	2	1		3	3
Galium 3-4"	–	–	12	3	22		2		1	/
Ribes oxyacant	–	1	–	4	15			8	/	23
R. triste	2	9	13	2		4	10	6	16	9
Mitella nuda	–	6	–	–	1	9		16	25	19
Mertensia nuda	–	11	6	10		2	10	4	1	12
Aralia nudic.	4	–	6	1	3			1	/	1
Viburnum edule	#2	#15	5	6		7	7	#5	#3	4
Lathyrobotrya	1	2	1				2	3	3	/
Calamagrostis	3	3	–	1	1	6	11	8	4	4
Pop. bal. (seedling)	2	1	–	1	1	2	2		1	/
Prunus virg. (seedling)	–	–	1	–					1	/
Pop. tremuloides	–	–	1	–				1	/	/
Actaea rubra	–	–	1	–	1				/	1
Cicuta alpina	4	–	1	1	1	3			2	11
Thal. venulosum	3	–	–	–		1	1		/	/
No. OF SPECIES	12	10	14	14	12	12	12	12	13	14

Mertensia stunted

图 1.1 是野外手册中典型的一页。这一页记录艾伯塔省 Athabasca 河的泛滥平原上的脂杨（Populus balsamifera）林地中地面植被的观察。

在第二节中完成。

第一节　数据矩阵及散点图

数据矩阵，从这个术语最普通的意义来说，是由行和列构成的任何观测值的表。群落生态学中最常遇到的数据矩阵是列出一定数量的抽样单元中每一单元内若干物种量的表。因此构成数据矩阵显然有两种途径：或者是让每一行表示不同的物种，每一列表示不同的抽样单元（本书中都是这样做的），或者相反。这里使用的方法是大多数生态学家们乐于采用的。

任何矩阵（包括数据矩阵）都可以用一个字母表示，用这个字母代表构成一张表的整个数组。通常表示一个矩阵的字母用粗体字印刷。若矩阵有 s行、n 列，把它描述为$s \times n$矩阵，或者说，它是一个 s × n 阶矩阵。为了增强记忆，本书用字母 s 与 n 表示数据矩阵的阶：s 表示物种、n 表示抽样单元的数目。在规定矩阵的阶或大小时，总是先写行数，后写列数。

现在考虑如何用符号表示 3 × 4 数据矩阵 **X**，用添加下标的 x 代表真实的数值。则：

$$\mathbf{X}=\begin{pmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ x_{21} & x_{22} & x_{23} & x_{24} \\ x_{31} & x_{32} & x_{33} & x_{34} \end{pmatrix}$$

可以看出，矩阵的每个元素，即每个单项，有二个下标，这些下标规定元素在矩阵中的位置：第一下标给出行数，第二下标给出列数。例如 x_{24} 表示 **X** 的第二行、第四列的那个元素，一般地，把 i 行、j 列的那个元素写成 x_{ij}。凡是有矩阵的数学著作都采用这个规则。由此得出：在数据

矩阵中，第 j 列是由第 j 个抽样单元中各个物种的数量构成的（如果在一个抽样单元中没有某个物种，其数量为零）。同样，第 i 行是所有抽样单元中第 i 个物种的数量。

现在我们进入本书的主题，数据的解释问题。要从野外编列的“原始”数据矩阵中看出其潜在的结构，甚至判断它是否有任何结构是很不容易的。我们用的“结构”这一词的含义是指可以显示出某些物种组有同时存在的趋势的任何系统模式，或者是抽样单元在适当排列时，在物种的组成中具有渐进的连续趋势的任何系统模式。

作为一个虚拟的例子，考虑下面这个10×10数据矩阵，它是一个“原始”矩阵。

$$X = \begin{pmatrix} 2 & 2 & 0 & 0 & 3 & 0 & 1 & 1 & 4 & 3 \\ 3 & 0 & 0 & 0 & 0 & 0 & 4 & 0 & 1 & 2 \\ 0 & 0 & 4 & 3 & 0 & 2 & 0 & 1 & 0 & 0 \\ 1 & 3 & 0 & 0 & 4 & 1 & 0 & 2 & 3 & 2 \\ 0 & 1 & 3 & 4 & 0 & 3 & 0 & 2 & 0 & 0 \\ 0 & 4 & 0 & 1 & 3 & 2 & 0 & 3 & 2 & 1 \\ 0 & 2 & 2 & 3 & 1 & 4 & 0 & 3 & 0 & 0 \\ 3 & 1 & 0 & 0 & 2 & 0 & 2 & 0 & 3 & 4 \\ 4 & 0 & 0 & 0 & 1 & 0 & 3 & 0 & 2 & 3 \\ 0 & 3 & 1 & 2 & 2 & 3 & 0 & 4 & 1 & 0 \end{pmatrix}$$

象通常那样，它的行代表物种，而列代表抽样单元。

不能否认，这个矩阵缺乏任何明显结构。但是如果将抽样单元（列）与物种（行）以适当方式重新排列，其结果是一个“编排矩阵”，即：

$$X=\begin{pmatrix}
4 & 3 & 2 & 1 & 0 & 0 & 0 & 0 & 0 & 0\\
3 & 4 & 3 & 2 & 1 & 0 & 0 & 0 & 0 & 0\\
2 & 3 & 4 & 3 & 2 & 1 & 0 & 0 & 0 & 0\\
1 & 2 & 3 & 4 & 3 & 2 & 1 & 0 & 0 & 0\\
0 & 1 & 2 & 3 & 4 & 3 & 2 & 1 & 0 & 0\\
0 & 0 & 1 & 2 & 3 & 4 & 3 & 2 & 1 & 0\\
0 & 0 & 0 & 1 & 2 & 3 & 4 & 3 & 2 & 1\\
0 & 0 & 0 & 0 & 1 & 2 & 3 & 4 & 3 & 2\\
0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 4 & 3\\
0 & 0 & 0 & 0 & 0 & 0 & 1 & 2 & 3 & 4
\end{pmatrix}$$

这个矩阵所包含的信息同前面那个矩阵完全相同，只是物种的顺序与抽样单元的顺序发生了改变。例如，原始矩阵中标号为1的物种（出现在第一行中），在编排矩阵中编号为4（因此出现在第4行中）。关于列，例如，原始矩阵第一列中编号为1的抽样单元重新编号为2，因而位于编排矩阵的第二列中。

确定产生编排矩阵的编号系统的方法将在第四章中论述，而这两个行和列都标号的矩阵将在表4.11中再次给出。这里只要指出这种方法是相当精细的就足够了。要想从原始矩阵形式导出编排矩阵，不采用特定的方法（算法）则是费时的，就象一个没有学习过玩魔方的人努力去解开魔方之谜一样。

前面这个编排矩阵X是具有“结构”的矩阵的明显例子。按本书中所用术语的含义，数据解释是由能在实际数据矩阵中观察到结构的各种方法组成。尽管这些矩阵在原始形式下可能会象X的原始形式那样，好象是无结构的，这

里所举例说明的表（或矩阵）的编排是一种的方法。另外两种方法，分类与排序，要比简单的表的编排能更多地揭示数据矩阵的结构。对这些方法的论述占本书的很大篇幅。这里只需要关于这些论题的少数简短的引论性内容。

关于分类，几乎没有多少要说的。分类这个词以其最普通的意义用于群落生态学中（即把相似的事物归并为类），从而不需要特殊的定义。显然，人们可以对抽样单元进行分类，把在物种的组成上相似的那些单元归并为一类；也可以按物种分类，把有可能分布在一起的那些物种归为一类。关于分类的这些相对模式在后面将有更多的论述。

很明显，还有两种相对立的方法可以用来进行数据分类。例如，如果有一些抽样单元要进行分类，先是从一个个单元开始，将它们归为小的类，然后把这些小的类再归并（等等），逐次形成更大的类，这是聚团分类或聚类，是第二章的论题；另一种就是从整组抽样单元着手，把它视为第一个包括一切的大类，然后把它分为较小的类，这是分割分类法，是第五章的论题。

现在我们转到排序上来，这是一个描述群落生态学中整个很有用的一组技术的术语。就其原意，排序(ordination)这个词的含意是“排秩次”（ordering）。首先可以考虑这样的例子：植物生态学家沿着环境梯度观察一行样地（样方）收集的数据，例如沿山坡向上走或者从陆地到海洋时穿过盐滩。在这种情况下，样方的顺序是予先给定的，不再需要“排序”。然而，如果生态学家在平坦的混交林里对草本植被以随机方式布置样方，并假设尽管环境有中等程度的差异，但梯度并不明显。这时不会有直接了当的方法对这些样

方进行排序，但是假设一种自然顺序存在可以是合理的，只要人们能发现它。这种假设近于假定森林环境是由不同生境的镶嵌组成(在镶嵌块之间并不必需有清晰的边界)，而且这些不同生境本身象顺着山坡向上或横越盐滩时的连续生境具有自然顺序一样有其自然顺序。如果这种假设是正确的话，就可以大致设想出一种技术，从植被资料中发现这种自然顺序，这样一种排序就是从这种技术里产生的。

上一段论述的是最初从事生态排序的人们的目标。现在让我们从另一种出发点来探讨这个题目。为使讨论具体，设想数据是由一名森林生态学家估价一个混交林中许多样地上几个不同树种的生物量而积累起来的。需要对这些样地进行排序。完成这个工作的最好方法是按照它们所含的最丰富种的数量将样地排秩次。但是为什么只停留在一个种排序上呢？他们还可以按照样地所含第二丰富种的数量来排秩次。获得这类排秩次的方便的办法是画出一个散点图，图中第一个物种的数量取在 x 轴上，第二个物种的数量取在 y 轴上，每块样地由一个点表示。显然，如果将这些点投影到 x 轴上，则它们的顺序与第一种排序的顺序相同；同样，如果将点投影到 y 轴上，它们的顺序与第二种排序的顺序相同。但是，散点图本身将给出群落清晰的结构图。将两根轴分开考虑是得不到什么的。把散点图考虑为在二维中的排序，这是数据的图象表示，而不仅仅是列出样地标号的表。

现在应该清楚，论题正在引向何处。如果二维散点图（显示最重要的两个物种在样地中的数量）好的话，那么三维散点图肯定会更好。虽然构成这样的图是比较困难的，需要在软木板上插上不同长度的小针来做成，但这种图含有更

多的信息。可是，为什么停留在三个物种上呢？每次忽略掉一个存在的物种就会牺牲掉一定量的信息。因此最佳的散点图应该是一种展示森林中所有 s 个 物种数据的散点图——不论 s 有 多大。困难仅在于超过三维的散点图不可能想象出来，更不用谈构成了。给人留下的是一个“概念性的散点图”这是一种不令人满意的研究对象。

然而，现在所用的生态排序确实是从概念性散点图出发的。各种不同的排序方法相当于把这些多维模式绘制在二维（有时是三维）空间中，用这种方法可以把想象的概念模式带回到真正可以看到的世界中，对它进行观察和研究。

这个过程类似于以三维地球为模式绘制二维地图的过程。差别在于当地理学家在纸上画出世界地图时，他可以容易地参考表示陆地、海洋的真实模式的三维地球仪；维度的减少只是从 3 到 2 。相反，生态学家们不能看到，甚至也不能想象出他所要作图的多维模式，而且所要求在维度上的减少量通常要大好多倍。但画地图和生态排序间的相似仍然是紧密的，也是有启发性的。这种相似性表明，在二维平面上画多维散点图可以有多种可能的技术，而且没有一种会自动地比所有其它的种类更好些。正如不同的作图投影法适用不同的目的一样，不同的排序方法可以将所研究的生态学数据的不同侧面强调出来。遗憾的是，技术与目的的配合，在生态学中远不如在地理学中那样清楚；目的本身并不那么十分明确，而达到目的的方法也不完全清楚。

现在我们转向考虑实践问题。为了实现对多维散点图的二维排序，人们必须对给定的数据，即 s × n 数据矩阵 （要记住 s 是物种数，n 是抽样单元数）进行工作。可以把数据

视为给出在由 s 个相互垂直的轴（每个物种一根轴）构成的坐标系中和由 n 个数据点（每一抽样单元一个点）组成的散点图中的各个点的坐标。第 j 个点的坐标由数据矩阵中第 j 列中的各个元素给出。形成一个“群”的数据点的整个集合称为数据群，生态学家使用的各种各样的排序方法相当于将 s 维数据群绘制在一张二维的纸上的各种不同方法。所得到的是抽样单元的排序。其中人们最熟悉而又广泛应用的方法将在第四章中论述，一些必要的数学初步知识放在第三章里。

读者无疑会注意到，如果象刚才所论述的那样，把一大群数据看作等同于 s 维空间中的 n 个点是合理的话，那未把这样数据看作 n 维空间中的 s 个点同样也是合理的。这样做时，数据矩阵中每一行（代替每一列）给出一个数据点的坐标。总共会有 s 个点，把这些点绘在 n 根相互垂直的轴所构成的坐标系里，每个抽样单元有一根轴，因而数据群的排序给出物种的排序。

抽样单元的排序为 R 型排序，而物种的排序为 Q 型排序。同样，有 R 型与 Q 型分类，但这些术语很少用。

完成 R 型和 Q 型排序的技术是相同的，起初觉得，这两种分析似乎都是合理的。但是 Q 型分析有一个大缺点。如果计划要对数据进行任何统计检验，哪未要求抽样对象间相互独立是主要的。群落抽样几乎总是以保证抽样单元之间的相互独立的方式进行的，而且抽样单元是 R 型排序的对象。但是，抽样单元中各个物种间肯定不是独立的，这些物种正是 Q 型排序的对象。统计假设的检验已超出本书的范围，我们不会再有机会去考虑抽样单元的随机性和独立性。然而，从统计

学的观点出发的R型与Q型分析的明显差异是应该记住的。

第二节　一些定义与其它初步知识

在继续论述之前，最重要的事情是给出群落生态学中应用很广的两个术语在本书中的定义：Sample 与 Clustering*。

Sample 这个词是生态学中许许多多混乱的一个根源，这是最不幸的。为了说明，设想一位植物生态学家在许多样方上进行了观测，对统计学家及许多生态学家来说，每一个样方是一个抽样单元 (Sample unit)，而全体样方的集合是一个样本 (Sample)。对另一些生态学家 (如 Gauch, 1982) 来说，每一个样方是一个样本 (Sample)，而全体样方的集合是一个样本组 (Sample set)。这种混乱可以清楚地表示在一个 2 × 2 表中，其中四个格内的词是由“列”内所规定的人给“行”内所规定的对象的名词。那么：

	统计学家和某些生态学家	另一些生态学家
单个单元(如样方)	抽样单元 (Sample unit)	样　本 (Sample)
单元的集合	样　本 (Sample)	样本组 (Sample set)

本书使用表中左列的术语。但是这两种术语都不是完全令人满意的，因为需要使用二个字构成的术语 (Sample

* 译者注：这里要说明使用这两个英文名词中的混乱现象，而不一定是指其中译名在使用中有混乱。

unit 或 Sample set）来表示一个单元实体是一种令人讨厌的事情。因此本书中，我使用 Quadrat* 这个词表示任何类型的抽样单元，而常添加几个字或用 Sample unit 作为一种提示。这是上述问题的一种方便的解决办法。但是，这样是否使那些视 Sample unit 并非 Quadrat（样方）的生态学家满意，还有待以后分晓。例如，这样的生态学家有：用 Suber 抽样器收集材料研究水生动物区系的学生、以沉积岩心为抽样单元的孢粉学者、以抽样网中的捕获物为抽样单元的浮游生物学家、以一个大捕虫网里的捕获物或者是诱虫灯下的捕获物为抽样单元的昆虫学家、以一片显微载玻片为抽样单元的硅藻专家、以一个小区或比传统的样方要大得多的一个标准地（虽然标准地象样方一样是一块限定面积的地面）为抽样单元的林学家。

我将尽可能避免使用 Sample 一词，因为它是含糊不清的，但是在它确实出现的地方，它是从统计意义上使用的，Sample 的含意是所有样方的集合。

Clustering 一词也是含糊的，某些生态学家把它视为与聚团分类（agglomerative classification）是同义的，另一些生态学家则视其与分类（classification）的通常意义同义，既包括聚团方法，也包括分割方法。这两种可能性

---

* 注：术语样方（Quadrat）肯定是被所有生态学家们所熟悉的。A. G. Tansley 和 T. F. Chipp 在他们的经典的《植被研究中的目标与方法》（不列颠帝国植被委员会，1926）一书中把样方定义为："仅仅是临时或永久性地划出的一块方形面积，作为所欲精确地研究的任何植被的一个样本"。较现代的定义则省略"方"这个字样，样方可以是任何形状的。还应该注意到，虽然定义中关于样方的大小没有说，但按通常习惯，都认为样方是小于"小区"（Plot）的。然而，既没有大家都同意的样方大小的上限，也没有小区大小的下限。某些"划出的面积"可以有理由用两个名字中的任何一个来称呼。