阳含熙　卢泽愚 著

植物生态学的数量分类方法

科学出版社

植物生态学的数量分类方法

阳含熙　卢泽愚 著

科学出版社
1981

内 容 简 介

本书共分十三章。前三章是进行数量分类的基础知识：数据的类型、转换、标准化和衡量研究对象差异的数量指标——相似系数；四至八章介绍各种分类方法及分类后的判别方法；最后五章介绍各种排序方法。附录中的矩阵基本知识是阅读本书必需的一些数学知识；书中关于植物群落的取样问题对正确收集数据也作了简要的叙述。

书中介绍的方法相当广泛，基本上概括了二十年来这方面的主要内容。对每一方法都讲清从原始数据出发，如何一步步运算，直到得出结果，并以简单的数值例子详细说明计算过程。这样可便于不熟悉数学的读者能懂得并能初步掌握应用这些方法。对其中一些重要的方法，还介绍了在植物生态学方面应用的实例。

本书可供植物生态学、生物学、地学、气象学与农林医学等方面的科技人员，应用数学工作者，以及大专院校有关专业师生参考。

植物生态学的数量分类方法

阳含熙　卢泽愚 著

责任编辑　于拔

科学出版社出版

北京朝阳门内大街 137 号

中国科学院印刷厂印刷

新华书店北京发行所发行　各地新华书店经售

*

1981 年 9 月第 一 版　开本：787×1092　1/32

1981 年 9 月第一次印刷　印张：13 5/8

印数：0001—4,400　字数：306,000

统一书号：13031·1676

本社书号：2299·13—8

定价：2.10 元

前　言

植物生态学数量分类的研究是从五十年代开始的，由于计算工作量大，等到六十年代电子计算机普遍应用之后，它才迅速地发展起来。许多具有不同观点的传统学派，如法瑞学派、英美学派、苏联学派等，都进行数量分类的研究，并用它去验证原来传统分类的结果。目前，在国外植物生态学研究中已广泛采用数量分类的方法，每年都发表大量的文章，出了不少专著，并不断涌现新的方法。

我们参考国外文献，并根据自己对国内森林、草原资料应用的经验，编写这本书，以向读者比较全面地介绍各种数量分类方法。书中包括了二十多年来这方面的主要内容，特别着重介绍近年来创造的新方法。虽然讲的是植物生态学的数量分类方法，实际上它们已普遍应用于生物学、地学的各个分支，而且对农业、林业、畜牧业等学科都是同样适用的。

当前，国内许多生物科学工作者都接触到大量应用数学方法的国际文献，并已感到有用数学工具的必要，但有不少人对数学望而生畏，至今仍很少应用象数量分类这种多元分析的方法。其实，如果我们只限于懂得并会应用一些现成的分析方法，并不深究它的数学根源，也不求创造新的方法，那并不是很难的事情。特别是数量分类的方法中，只用到一般的代数演算，和线性代数中有关矩阵的一些基本知识，绝非一般生物工作者所难于理解的。困难主要在于他们不熟悉数学的术语，不习惯数学的符号，这是非本质的困难，是不难克服的。

因此，我们在介绍方法中力求写得浅显简明，逐步引进必要的数学术语和符号；讲述与实例结合；同时书后附录中，专门介绍一些阅读本书所必备的矩阵代数的基本知识。我们希望一般生物学、地学和农林牧等专业的科技工作者都能懂得介绍的方法，以便能顺利阅读这方面的国外文献，并在自己的科研工作中也能初步掌握应用这些方法。

限于作者水平，应用这些方法的实践经验还很少，书中难免出现缺点和错误，希读者随时指正。

作　　者

1979.11

绪　言

这里说的"数量分类"方法是个总称，它包括两类处理大量数据的多元分析方法：分类和排序。在植物生态学研究中，往往涉及到大量有关植被和环境因素的观测数据，单凭人们头脑要直观地从中看出内部的关系，不仅需要丰富的经验，而且很不容易做到。数量分类方法，无论是分类或者排序，都是给出一系列完整的处理原始数据的计算规则，最终给出简化形式的数据结构。这就提高了人们分析数据的能力，从而揭示出一些不易发现的有意义的规律：或者对植物群落和环境因素进行了比较客观的分类，或者找出了植物种之间，或植被与环境因素之间的相互关系。

大多数的分析方法都要求非常繁多的，虽然是不难的计算，特别当原始数据量大时更是如此。因此，这些方法几乎都需要借助电子计算机才能实现，没有计算机的普遍应用就不可能如此迅速地发展这类多元分析的方法。

目前用于生态学的数学方法大致可分三类：一类是构造生态系统的数学模型，一类是统计分析，另一类就是数量分类。数量分类很接近于统计分析，但又有明显的区别。它虽然也处理大量数据，这些数据往往也是从考察的更大总体中抽取来的，但它们不是统计意义下的样本，一般并不要求从它们去推断总体的规律。因而通常不需随机取样，不必了解数据的分布性质，也不涉及显著性检验的问题。我们只是将原始数据当做总体，找寻的规律并不外推，所以数量分类的方法

是非概率的。为了区别起见，Williams(1976)称这类方法为"模式分析"(pattern analysis)。这种用法不大合适，因为同一英文名称早被 Greig-Smith(1957)用于研究种群分布格局的一种方法，称为"格局分析"，所以我们在中文译名上加以区别。

模式分析与统计分析是互相补充的。在解决植物生态学的某些具体问题中，有时既可用统计分析的方法，也可用模式分析的方法；甚至两者兼而用之。比如可以先用模式分析将数据分类，然后才便于统计分析：如先通过主分量分析找出主要相关的因子，再做单因子的回归分析；或者先通过统计分析，然后借助模式分析对其结果进行明确的解释。而且，有些分析方法本身就可用于两种分析，如数量分类方法中的主分量分析也可用于纯统计的目的，不过在本书中我们只把它当做一种排序方法。

进行数量分类的基本单位叫做实体 (entity)，描述实体数量特征的各个信息项目称为属性(attribute)。在植物生态学研究中，实体可以是样方、标地、地段(林分)或群落等等，为方便起见，我们以后将实体一般都叫做样方。样方中观测的各个种的数据项目(如种的存在不存在、种的频度、盖度或重量等等)，以及环境因素的数据项目(如样地坡度、雨量、日照、土壤深度、各种养分元素的含量等等)都是属性。数量分类的基本问题是，根据对一组实体按属性记录的原始数据，通过一系列的计算程序，将这组实体进行分类或者排序，即按属性分类实体。

从数学的角度讲，这些方法是施于原始数据集合的一套处理规则，方法本身不依赖于对实体和属性具体内容的解释，因此可用于多种学科。比如，我们的实体可以是张三、李四、王五等若干个人，而属性是记录他们对各种专业知识的掌握程度，我们就可对这些人按知识进行分类；实体也可以是收集

的一组植物标本，并记录它们的形态、化学成分、染色体数目等等数据作为属性，那就在进行植物的分类。事实上，在动植物分类学、生态学、地质学、心理学等等许多学科中，都已广泛应用这些方法。

同样的道理，对同一数据集合，实体与属性的地位也可对调。也就是说，既可按属性去分类实体，也可按实体去分类属性，前者称为正分析，后者称为逆分析。原则上讲，任何一种数量分类的方法都有正逆之分，它们的差别仅仅是将原始数据集合的排列方向改换一下，所以我们对各种方法一般都只介绍正分析。是否两种分析都有意义，应当进行正的还是逆的分析，自然要按解决的具体问题的生态意义而定。

如何进行野外调查取得原始观测数据，不在本书正文的讨论范围之内。我们在本书中的工作是建立在已经有了实体-属性原始数据的基础上的。

由于属性多种多样，反映它们的数据类型就有不同。比如，种的频度、雨量、日照时数等是数值；种的存在和不存在只有两种状态；土壤颜色可分为红、黑、黄等多种状态等等。所以需要考虑数据的类型，以及不同类型间的转化，最后将具有同一类型的数据排列成要求的格式。这些是第一章的内容。另外，收集的原始数据中，可能由于生态意义不大，不合某种性质的要求，或者量纲不同、数值大小悬殊等原因，为了便于分析有必要对它们预先进行适当的处理。这包括数据减缩、转换和标准化等方面的内容，本书第二章讨论它们。前两章都是必要的基础工作。

有了合乎要求的原始数据后，下一步是根据属性的数据去求出衡量各实体之间，或实体与实体组之间，或各实体组之间彼此相似或相异程度的数量指标（总称相似系数），并把它们排列成要求的格式。所有分类和排序方法都是根据这一数

量指标来进行分析的，这也是一项基础工作。同时，对不同情况已经设计了各种各样的相似系数算法，所以本书第三章专门讨论这方面的内容。

在上述工作的基础上就可进行分类或排序了，以下各章分别介绍各种不同的分析方法。

分类是根据各实体间的相似关系将实体分成若干组，使组内的实体相当相似，而组间的实体则尽量相异，从而实现对植物群落的比较客观的分类。如果对同一实体集合分别按植物种的组成和按环境因素进行分类，两者比较还可能揭示出植被与环境因素之间的关系。我们在第四章中概述分类的目的、类型、一般的分类过程，以及分类结果的图象表示与比较等内容。第五、六、七章再详细地讨论不同类型的一些比较重要的，用得较多的或者比较新的分类方法。

与分类方法密切相关的另一类型问题，是对实体集合已经进行了分类，现在又加上一个实体的数据，我们不必把它与原实体的数据合起来重新去分类，而是在原分类的基础上要判别此新实体应该属于哪一类。这是判别分析的问题，在第八章中介绍一些这方面的基本内容。

排序是将实体作为点在以属性为坐标轴的空间中按其相似关系把它们排列出来，特别是在尽量少损失原数据信息的要求下，力争在简化的空间中排列实体，从中可以发现实体间的分布格局，揭示出植物种之间或者种与环境因素之间的关系。排序方法很多，不同方法的差异很大，不象分类方法那样可以归纳出一般的规则，因而内容较多。我们大致分为较早期的、计算较容易的方法，和其它一些较新的、数学上较复杂的几类方法，分别在九至十三章讲述。

数量分类方法还在不断发展中，在解决具体生态问题时，应该进行分类还是排序，以及从许多方法中应选用哪种方法，

目前尚无客观的选择策略；哪种方法“最优”也不易回答，而且是很有争议的问题，或许永远也找不到准确的客观判别标准。这是一个需待深入探讨的问题，不是本书所要解决的。

书末的附录有两部分，一是有关矩阵的基本知识，以帮助不熟悉数学的读者阅读本书；一是介绍植物群落研究的取样问题，供进行植物生态调查的读者参考。

目　录

第一章 数据类型

第一节 数据的基本类型

数量分类方法几乎总是从对一组实体测出若干属性的数据出发的。由于属性的性质多种多样，反映属性的数据就有不同的类型。我们将属性，从而它相应的数据分为三种基本类型，分别简述如下。

一、名称属性 (nominal attributes)

有的属性只能描述为若干种不同的状态，每个实体具有其中一种状态。比如土壤的颜色(属性)可分为红、黑、黄等等；岩石可分为页岩、砂岩、玄武岩、花岗岩等等；植被可分为森林、草原、灌丛、苔原等等；都是这种情况。这种属性的基本特点是，在作为数据处理时各个状态的地位是等同的，状态之间没有一定的顺序。由于它的各种状态可用不同名称表示，所以称这种属性为名称属性。

我们依其状态的数目，把它分成两类。

1. 二元属性 (binary attributes)

名称属性的一种重要特别情况，是只具有两个状态。如某植物种的存在不存在，某昆虫的有翼无翼，某植物的有刺无刺，动物是雌是雄，等等。这种属性叫做二元属性，因为它往往是确定某种性质的有无，所以也称为定性属性 (qualitative

attributes)。

对二元属性的两个状态常用两个数字 0 和 1 来表示。当属性是指某种性质的有无时，一般用 0 表示不具有该性质，1 表示具有该性质，比如用 0 表示种不存在，1 表示存在。当属性是指两个对立的状态时，0 和 1 各表哪个状态可以随便指定，比如可用 0 表雄、1 表雌，或者反之。

在植被考察中经常采用这种定性的属性，它比调查植物种的量(如频度、盖度、重量等)，无疑要简便得多。但是用 0、1 数据显然要比定量数据少获得信息，而且加强了罕见种的作用。如果费事不大，自然以用定量数据为好。也有人认为，当定量数据中有相当多的 0 时，定性数据带来的信息损失比直观预料的要小。

二元数据还有对称与不对称之分。如果我们在其后的分析中，把 0 和 1 的地位等同对待，它就是对称的；如果偏重一方就是非对称的。在植物分类学中，有毛或无毛是一个很重要的特性，我们可将二者等同对待，是对称的二元数据。而在植被考察中，一个种的存在或不存在，很多情况下不能等同对待，因而有时看成是对称的数据，有时又认为是不对称的数据。

二元数据用得很多，绝大多数数量分类方法都适用于这种数据，甚至有的方法还是专为它设计的。

2. 无序多状态属性 (disordered multistate attributes)

具有三个以上状态的名称属性，又称为无序多状态属性，以强调它的状态间无一定顺序。

假设某属性有 n 个状态，我们可分别用数字 1、2、3、···、n 代表各个状态。比如上述岩石类型，可用 1 代表页岩，2 代表砂岩，3 代表玄武岩，4 代表花岗岩。但是，这些数字不能反

映各状态间的差异，1 与 4 之差在数量上三倍于 1 与 2 之差，这样来反映岩石的差别显然是荒唐的。所以这些数字不能用来进行一般的数值计算，只能看作记号而已，它无异于用 A、B、C、D、···或甲、乙、丙、丁……来代表不同的状态。

这种属性还没有理想的数据表示方法，也很少数量分类方法适用于这种数据。但已有一种信息的相似系数及相应的信息分类方法可适用于这种数据(参看第三章四节)。

二、顺序属性 (ordinal attributes)

它与无序多状态属性一样，也只能分成多个状态。所不同的是，现在的状态有确定的顺序，所以也称为有序多状态属性。例如某植物种的多度分为大量、常见、普遍、罕见和不出现五种状态；土壤酸碱度分为强酸性、弱酸性、中性、弱碱性、强碱性等状态；植物种子分成大、中、小三级，等等，都是顺序属性的例子。显然各状态之间的顺序是有意义的。

与无序多状态属性一样，对顺序属性也没有理想的数据表示方法。用 1、2、3、···、n 依次表示各个状态，虽然数字间的大小差别反映了属性状态间的顺序关系，但是不能恰当地表示各状态间的差距。因为顺序属性相邻状态间的差距没有明确的规定，并不如 1、2、3、4……那样，相邻数字间的差距总是等间隔的，所以这种数据表示也不很适于做为数值运算。同样，信息的相似系数及信息分类方法可用于这类数据。

可以看出，顺序属性往往是根据某种定量的属性在所有实体的可能取值范围内划分等级的。比如上述种的多度可以测出数值，我们综观所有样方的取值范围而粗略地分成五等；土壤的酸碱度也有 pH 值的数量指标，也是综观所有实体的数值而划分等级的。这说明顺序属性与数量的属性密切相关，如