生态学杂志 Chinese Journal of Ecology 2005,24(7):807～811

排序和广义线性模型与广义可加模型在植物种与环境关系研究中的应用*

朱 源 康慕谊**
(北京师范大学资源学院 中国生态资产评估研究中心 环境演变与灾害教育部重点实验室,北京 100875)

摘 要 排序和广义线性模型(Generalized Linear Model,GLM)与广义可加模型(Goneralized Additive Model,GAM)是研究植物种与环境间关系的重要方法。基于线性模型的排序方法应限定于环境梯度较短的植被数据,而基于单峰模型的排序方法更适用于梯度较长的情况。PCA、CA/RA系列和CCA系列是常用的排序方法。同时进行环境数据和植被数据分析的CCA系列,能清楚地得出植物种与环境间的关系。CCA改进后的DCCA和PCCA,是现今较理想的排序方法。GLM和GAM实质上是用环境变量的高阶多项式来拟合植物种与环境变量的关系。GLM和GAM扩展了植物种与环境变量之间的关系模型,能深入地探讨植物种与环境间的关系。GLM主要是模型决定的,而GAM主要取决于原始数据。一般来说,排序能得出研究区域的主要环境梯度,提供了物种聚集和植物群落的概略描述。GLM与GAM对于深入研究单个植物种与环境间的关系具有优势。在实际研究中,两种方法结合使用能互补不足。

关键词 植物种-环境模型,排序方法,广义线性模型与广义可加模型,方法比较
中图分类号 Q948 **文献标识码** A **文章编号** 1000-4890(2005)07-0807-05

Application of ordination and GLM/GAM in the research of the relationship between plant species and environment. ZHU Yuan, KANG Muyi (*China Ecological Assessment Research Center, Key Laboratory of Environmental Change and Natural Disaster, Minisitry of Education; College of Resources Science and Technology, Beijing Normal University, Beijing* 100875, *China*). *Chinese Journal of Ecology*, 2005, **24**(7): 807～811.

Ordination and GLM/GAM are both important methods to explore the relationship between plant species and environment. Generally, ordination methods, based on linear model, are suitable for dealing with the vegetation within a comparatively short environmental gradient, while those based on unimodal model are more suitable for tackling with the vegetation distributing along a longer gradient. PCA, CA/RA series and CCA series are the most commonly used ordination methods. Involving the environmental data in the analysis, CCA series could demonstrate the relationship between plant species and environment clearly. After modification, DCCA and PCCA are considered to be ideal ordination methods at present. Essentially, General Linear Model (GLM) and General Additive Modal (GAM) were using the high power polynomials of environmental variables to fit the response of plant species to environment. So, the models for delineating the relationship between species and environment have been enhanced through the introduction of GLM and GAM in recent years. GLM is basically driven by model, while GAM is more driven by data. In comparison, ordination can obtain the main environmental gradient and provide the general information of species assembling and plant community in the study area. GLM and GAM have advantages in probing deeply into the relationship between plant species and environment. In practice, the two kinds of methods can yield better results when used in combination.

Key words species-environment model, ordination, GLM and GAM, methods comparison.

1 引 言

植物种与环境间的关系是植被生态学研究的重要内容之一[3,15～18]。植物种的空间分布与环境之间存在密切的关系,这是比较公认的一点。但是,环境以何种方式影响植物种,在多大程度上决定植物种的空间分布格局,以及植物种对于环境做出何种反应,这些问题仍然是研究的热点。其中一个重要的原因就是,自然界植被分布格局的成因相当复杂。排序(ordination)是植被生态学研究的重要手段之一。植物种排序可用于解释植物种的空间分布及其与环境的关系。排序是基于"植被连续体"的假设,即植被的连续变化一般与环境变量的连续变化一

*国家自然科学基金资助项目(40271047和40371043)。
**通讯作者
收稿日期:2004-09-12 改回日期:2004-11-19

致。通过排序,植物种排列在多维空间中,排序轴(即空间的“维”)能够反映一定的生态梯度,每个植物种是该空间的一个点。排序的目的就是要得出影响植物种和植物群落分布的主要环境梯度,即排序轴。排序图上空间距离越近的植物种,其所处的环境也越相似。排序方法基于的模型常用的有线性和单峰模型两种[8,9]。

2 常用排序方法概述

2.1 早期的排序方法

排序的概念最早是指用一、两个环境梯度去排列植物群落。早期如加权平均排序(weighted average)、极点排序(polar ordination)等方法的计算一般比较简单、易行,而且能得出直观且较真实的结果。但是数学基础不够严密,考虑的环境变量较少,主观的影响较大等原因限制了这些方法在更大范围内的应用[31]。

2.2 主分量分析

主分量分析(Principal Components Analysis, PCA)是一种完全基于植被数据而不考虑环境数据的排序方法,具严格的数学基础,基于线性模型,去除了主观影响。PCA 中的主分量(即排序轴)是各属性数据的线性组合。原始数据中的非线性关系会使 PCA 分析出现误差,故 PCA 的运用应限定在变异较小的数据,即较短的环境梯度。可将 PCA 的排序轴与环境变量进行相关分析,用以间接地解释植物种的空间分布及其与环境间的关系。但如果排序轴与环境变量的相关较小,解释上就比较困难[1,6,7,32~35]。

2.3 对应分析及其衍生系列

相对于 PCA 来说,对应分析/相互平均法(Correspondence Analysis/ Reciprocal Averaging, CA/ RA)基于单峰模型,更符合实际,在样地数据较大时尤为如此。CA/ RA 在多数情况下会产生“弓形效应”(Arch Effect),“弓形效应”同时产生了“边缘效应”(Edge Effect),即排序空间中植物种间的实际距离被歪曲。CA/ RA 对于极端样地和稀有种很敏感,故在 CA/ RA 分析中,极端样地和稀有种最好在分析之前去掉[12,23,36~37]。为解决 CA/ RA 的“弓形效应”问题,可在 CA/ RA 的基础上加入“除趋势”的步骤,即除趋势对应分析(Detrended Correspondence Analysis, DCA)。与 CA/ RA 相比,DCA 提高了排序精度,与高斯模型更为吻合。PCA 不适用于非线性数据,而 DCA 既适用于线性数据也适用于非线性数据。CA/ RA 和 DCA 分析完全基于植被数据,所以在分析植物种与环境间关系时,会出现与 PCA 相同的问题[4,11,15,20,21,25,28]。

2.4 典范对应分析及其衍生系列

典范对应分析(Canonical Correspondence Analysis, CCA)需要植被和环境两组数据。CCA 的排序轴与环境变量的线性组合相关,从而能直接地研究植物种、植物群落和环境变量之间的关系[5,10,22,29,30,38~40]。为解决 CCA 的“弓形效应”问题,可加入“除趋势”的步骤,即除趋势典范对应分析(Detrended Canonical Correspondence Analysis, DCCA)。当“弓形效应”不明显时,CCA 和 DCCA 的排序结果相似。当数据中的环境变量较多时,“除趋势”对于 CCA 并非最佳。Ter Braak 认为冗余变量(对植物群落或植物种分化特征影响不明显的环境变量)是产生“弓形效应”的根本原因。只要去除冗余变量,就不会产生“弓形效应”,也就不需要“除趋势”。局部典范对应分析(Partial Canonical Correspondence Analysis, PCCA)能去除冗余变量的影响,是一种真正的直接梯度分析方法。冗余变量的选择可运用经验,可通过相关分析环境变量与植被数据后选择等。前向选择(Forward Selection)是一种较常用,且能自动选出冗余变量的方法。PCCA 的排序轴与选出的环境变量的线性组合相关,且与冗余变量无相关。所以在数学上,PCCA 就比 DCCA 更严密,在植被生态学研究中,会有广泛的应用[2,13,14,19,41]。

CCA、DCCA 及 PCCA 在排序完成后,植物种、植物群落与环境变量间的关系十分清楚。DCA 侧重于描述群落间的关系(群落的物种组成差异),而 DCCA 更适于分析群落与环境的关系。如果 DCA 和 DCCA 的排序结果相差很大,可能是因为一些重要的环境变量被忽略了。

3 GLM 和 GAM

3.1 GLM 和 GAM 的原理

回归分析是研究植物种与环境间关系的利器,它认为植物种数据是环境变量的响应。多元线性回归将环境变量和植物种数据分别定义为预测变量和响应变量,环境变量可以有多个,而响应变量(植物种数据)只有一个。多元线性回归的数学方程如下:

$$Y = a + b_1 X_1 + b_2 X_2 + \cdots b_p X_p$$

式中,Y 为植物种的响应变量,可以是植物种的重要值、盖度等表征植物种在群落中优势程度的数值,$X_1 \cdots X_p$ 为 p 个环境变量,$b_1 \cdots b_p$ 为 p 个回归系数,a 是回归方程常数项。线性模型一般要求响应变量(Y)服从正态(高斯)分布。同 PCA 一样,线性模型是指植物种随着环境变量的变化呈线性变化[24,33]。

GLM 是多元线性回归的推广,是在线性模型的基础上加入一个单调且可二次微分的联系函数,其方程形式为:

$$g\{E(Y)\} = LP = a + b_1X_1 + b_2X_2 + \cdots b_pX_p$$

相同字母的含义同上,LP 是线性预测值,$E(Y)$ 是的期望值,$g\{\}$ 是一个单调且可二次微分的联系函数。Y 的分布属于指数分布族。

GLM 相对于线性模型有了相当的改进。GLM 得出的植物种与环境变量关系的曲线属于指数型分布族,即 GLM 容纳的模型不仅有高斯模型,还有如泊松分布、双峰模型等其它模型。响应变量 Y 与线性预测值 LP 是通过联系函数 $g\{\}$ 连接的,使得方程的形式更为多样。数据"溢出"(overdispersion)(即数据的变异大于模型所能提供的变异范围)是生态数据常见且重要的问题,GLM 本身具有解决数据"溢出"这个问题的性质。GLM 中环境变量的高次项可用于解释原始数据中的非线性关系,一般来说,阶数越高,如三次项,则该环境变量的重要性越大。

为符合 GLM 的模型,需选择环境变量合适的多项式,且要估计回归参数,这往往是麻烦且不精确的。GAM 是 GLM 的进一步推广,能自动地选择合适的多项式,且不需要估计回归参数,方程的一般形式为:

$$g\{E(Y)\} = LP = a + f_1(X_1) + f_2(X_2) + \cdots + f_p(X_p)$$

相同字母的含义同上,$f_1 \cdots f_p$ 是 p 个环境变量 X 的平滑函数,平滑函数一般可用三次样条函数来估计。

GAM 相对于 GLM 同样有所改进。GLM 主要是模型决定的,而 GAM 主要取决于原始数据,故 GAM 能更加深入地探讨植物种与环境变量的关系,预测植物种的空间分布。GAM 中容纳的模型(即植物种与环境变量的关系)要多于 GLM,但仍然服从一定的范围。GAM 的关键是选择合适的平滑函数。GLM 和 GAM 实质上是用高阶的多项式来模拟原始数据中植物种与环境变量之间的关系。一般来说,函数越复杂,结果就越精确,但计算也越复杂。所以选择函数时,要兼顾精确性与计算量[26,27,33]。

3.2　GLM 和 GAM 的难点和进展

GLM 和 GAM 的回归方程中存在交互项(即多个环境变量共同作用,影响植物种的空间分布),关于交互项的数学形式以及解释还是一个难点,已经有这方面的研究,但还不是很清楚。原数据中物种出现的次数(即样本数)越多,GLM 和 GAM 的结果就越精确,但当物种为稀有种或观测较少时,GLM 和 GAM 结果的可信度较差。GLM 和 GAM 希望能将所有对植物种空间分布产生较大影响的环境变量纳入模型中,但在实际研究中,往往是难以实现的。所以对于不能定量为环境变量的影响因素(如干扰),GLM 和 GAM 很敏感,而且这些因素的影响越大,其结果就越不准确[26,27,33]。

当植物种有多个响应变量时(如对于每个物种,有高度、盖度等多个数据),矢量广义线性模型和矢量广义可加模型(Vector GLM/GAM,VGLM/VGAM)可用于解此类问题,VGLM 与 VGAM 是 GLM 与 GAM 的进一步推广。与 GLM 与 GAM 类似,VGLM 主要是模型决定的,而 VGAM 主要取决于数据。VGLM 和 VGAM 的结果一般比较复杂,呈复杂的方程形式。为便于解释,得出数据的主要变异,提供低维的视图(类似于排序图),可在 VGLM 和 VGAM 中加入一个"降维"的步骤,即降维矢量广义线性模型和降维矢量广义可加模型(reduced rank VGLM/VGAM,rrVGLM/rrVGAM)。CCA 可以认为是 rrVGLM 的一个特例。所以,Yee 等[42]认为 VGAM 容纳了更多的植物种与环境关系模型,是植物生态学研究方法的一个重要发展。VGAM 能精确地研究植物种与环境间的关系,对于植物种空间分布格局、生态位等生态理论的研究,也具有重要的意义。

4　排序和 GLM 与 GAM 的比较

排序和 GLM 与 GAM 在植物种与环境间的关系研究中,应用都比较广泛。两种方法各有侧重,也有着相同之处。

排序和 GLM 与 GAM 研究的侧重点不同。排序侧重于对研究区域一般生态梯度的探讨。由于较好地处理了植物种聚集的问题,所以能用于植物群落的划分与解释。GLM 与 GAM 对于预测单个物种的空间分布较准确,而且样本越多,结果越精确,

故能用于植被制图。GLM 与 GAM 模型本身不能直接处理种间关系的问题,故难以区分植物群落,不过对于物种聚集分布或相互排斥的问题,GLM 也有这方面研究的进展[33]。

排序和 GLM 与 GAM 的结果在表现形式上不同。在排序的结果中,植物种在排序图上是一个点,代表该种空间分布的平均位置(即最大值分布的位置),是定性的描述。当然,也可以通过空间插值的方法,得出排序图上植物种优势度的等值线图[36]。GLM 与 GAM 对每个物种以数学方程的形式,预测该种对于多个环境变量的综合响应,是定量的预测。

排序和 GLM 与 GAM 的结果存在差异。总的来说,GLM 与 GAM 解释数据总变异的比例要大于排序。当调查的植物种样本充足时,GLM 与 GAM 对于每个物种可选择能解释较大比例变异的环境变量,而排序对所有物种选用同样的排序轴。所以,对于单个物种来说,GLM 与 GAM 的结果与实际情况更吻合。当拟合较少出现的植物种时,排序可能要优于 GLM 和 GAM。因为排序轴与环境变量相关,故具有相同或相近生态需求(即对环境条件的要求)的植物种会分布在类似的环境中,即在排序空间中相邻。少见种大多数情况下会与常见种关联,由于排序分析中常见种的结果较精确,少见种的拟合也会相对准确一些[33]。

排序和 GLM 与 GAM 也存在联系。一般来说,GLM 和 GAM 方法能较好拟合的物种(即与实际情况符合),排序的结果也较好。反之也是一样。而拟合不好的植物种(两类方法同样),往往是由于样本数太少或干扰等难以定量化的因素。排序和 GLM 与 GAM 在方法原理上也是有联系的。如上文所述,CCA 是 rrVGLM 的一个特例。rrVGLM 与 rrVGAM 的“降维”与排序在本质上可以说是一致的,那么这两种方法就必然存在联系。排序基于的模型主要是线性模型和均衡的单峰模型,这在自然界中往往是难以符合的。GLM 与 GAM 容纳的模型更多,适用于处理如“偏峰”、“双峰”等更复杂的植物种与环境间的关系[42]。

5 结 语

植物种排序概略性地描述了植物种的空间分布,并能给出相应的解释。线性模型适用于环境梯度较短的植被数据,而单峰模型更适合较长梯度的情况。PCA、CA/RA 系列和 CCA 系列是较为常用的排序方法。对于植物种排序来说,同时分析植被数据和环境数据的 CCA 系列,可以清楚地得出植物种与环境间的关系。经过改进的 DCCA 和 PCCA 不仅保留了 CCA 的优点,而且减少了误差,更加精确,是现今较理想的排序方法。

GLM 和 GAM 实质上是用高阶的多项式来模拟原始数据中植物种与环境变量之间的关系。GLM 主要是由模型驱动的,而 GAM 主要取决于数据。一般来说,多项式越复杂,结果越精确。但是,越复杂的模型,运用时就越麻烦,解释上也越困难,而且可能会将原始数据中的“噪声”(如测量误差)也模拟成植物种对环境的响应。当调查样本较少或外部影响较大时,GLM 和 GAM 的结果偏差就越大。GLM 和 GAM 在运用中出现了不少难题,需要解决。在 GLM 和 GAM 的理论上也有一些发展,VGAM 就被认为是近年来植被生态学研究方法的重要进展。

排序能用于划分与解释植物群落,描述植物种空间分布的一般格局,能较好处理植物种聚集的问题。GLM 和 GAM 对于深入地研究单个植物种与环境变量的关系十分有益。所以,两种方法同时使用,相互参照,互补长短,对于实际的研究具有重要的意义。

参考文献

[1] 王孝安,冯 杰,张 怀. 1994. 甘肃马衔山林区植被的数量分类与排序[J]. 植物生态学报,**18**(3):271～282.

[2] 王国宏. 2002. 祁连山北坡中段植物群落多样性的垂直分布格局[J]. 生物多样性,**10**(1):7～14.

[3] 朱 源,邱 扬,傅伯杰,等. 2004. 河北坝上草原东沟植物群落生态梯度的数量分析[J]. 应用生态学报,**15**(5):799～802.

[4] 江 洪. 1994. 川西北甘南云冷杉林的 DCA 排序、环境解释和地理分布模型的研究[J]. 植物生态学报,**18**(3):209～218.

[5] 米湘成,张金屯,张 峰,等. 1996. 山西高原植被与气候的关系分析及植被数量区划的研究[J]. 植物生态学报,**20**(6):549～560.

[6] 阳含熙,卢泽愚. 1983. 植物生态学的数量分类方法[M]. 北京:科学出版社.

[7] 张利权. 1987. 瑞典河漫滩草甸植被的数量分类和排序[J]. 植物生态学与地植物学学报,**11**(3):171～182.

[8] 张金屯. 1991. 植被数量分类和排序的发展[J]. 山西大学学报(自然科学版),**14**(2):215～224.

[9] 张金屯. 1995. 植被数量生态学方法[M]. 北京:中国科学技术出版社.

[10] 张金屯. 1998. 典范主分量分析及其在山西植被与气候关系分析中的应用[J]. 地理学报,**53**(3):256～263.

[11] 张新时. 1991. 西藏阿里植物群落的间接梯度分析:数量分类与环境解释[J]. 植物生态学与地植物学学报,**15**(2):101～113.

[12] 李绍忠. 1987. 辽东三块石天然次生林的排序[J]. 植物生态学与地植物学学报,**11**(4):264～275.

[13] 李　斌,张金屯. 2003. 黄土高原植物群落生态关系研究[J]. 农业环境科学学报,**22**(4):471～473.
[14] 沈泽昊,张新时. 2000. 三峡大老岭地区森林植被的空间格局分析及其地形解释[J]. 植物学报,**42**(10):1089～1095.
[15] 沈禹颖,阎顺国,朱兴运. 1994. 河西走廊盐化草甸主要植物群落分布特点及其土壤环境特征[J]. 植物生态学报,**18**(1):95～102.
[16] 郑慧莹,李建东,祝廷成. 1986. 松嫩平原南部植物群落的分类和排序[J]. 植物生态学与地植物学学报,**10**(3):171～179.
[17] 施维德. 1983. 四川省缙云山森林群落的分类和排序[J]. 植物生态学与地植物学丛刊,**7**(4):299～312.
[18] 赵志模,郭依泉. 1990. 群落生态学原理与方法[M]. 重庆:科学技术文献出版社重庆分社.
[19] 郭水良,曹　同. 2000. 长白山森林生态系统树附生苔藓植物分布与环境关系研究[J]. 生态学报,**20**(6):922～931.
[20] 钱　宏. 1990. 长白山高山冻原植物群落的数量分类和排序[J]. 应用生态学报,**1**(3):254～263.
[21] 黄　净,韩进轩,阳含熙. 1993. 长白山北坡阔叶红松林的DCA排序分析[J]. 植物生态学与地植物学学报,**17**(3):193～206.
[22] 潘代远,孔令韶,金启宏. 1995. 新疆呼图壁盐化草甸群落的DCA、CCA及DCCA分析[J]. 植物生态学报,**19**(2):115～127.
[23] Austin MP. 1985. Continuum concept, ordination methods, and niche theory [J]. *Ann. Rev. Ecol. Syst.*, **16**:39～61.
[24] Austin MP. 2002. Spatial prediction of species distribution: An interface between ecological theory and statistical Modeling [J]. *Ecol. Model.*, **157**:101～118.
[25] Goodall DW. 1954. Objective methods for the classification of vegetation Ⅲ An assay in the use of factor analysis [J]. *Aust. J. Bot.*, **3**:304～324.
[26] Guisan A, Weisis SB, Weisis AD. 1999. GLM versus CCA spatial modeling of plant species distribution [J]. *Ecology*, **143**:107～122.
[27] Guisan A, Edwards TCJ, Hastie T. 2002. Generalized linear and generalized additive models in studies of species distributions: Setting the scene [J]. *Ecol. Model.*, **157**:89～100.
[28] Hill MO, Gauch HG. 1980. Detrended correspondence analysis: An improved ordination [J]. *Vegetation*, **42**:47～58.
[29] Hill MO. 1974. Correspondence analysis: A neglected multivariate method [J]. *J. Roy. Statis. Soc. Series C*, **23**:340～354.
[30] Hill MO. 1973. Reciprocal averaging, an eigenvector method of ordination [J]. *J. Ecol.*, **61**:237～249.
[31] Jongman RHG, Ter Braak CJF, Van Tongeren OFR. 2002. Data Analysis in Community and Landscape Ecology [M]. Wageningen: Pudoc.
[32] Kershaw KA, Looney JHH. 1985. Quantitative and Dynamic Plant Ecology (3rd edition) [M]. London: Edward Arnold.
[33] Oksanen J, Minchin PR. 2002. Continumm theory revisited: What shape are species responses along ecological gradients [J]. *Ecol. Model.*, **157**:119～129.
[34] Swaine MD, Greig-Smith P. 1980. An application of principal components analysis to vegetation change in permanent plots [J]. *J. Ecol.*, **68**:33～41.
[35] Ter Braak CJF, Prentice IC. 1988. A Theory of Gradient Analysis [J]. *Adv. Ecol. Res.*, **18**:271～317.
[36] Ter Braak CJF, Smilauer P. 2002. CANOCO Reference Manual and CanoDraw for Windows User's Guide: Software for Canonical Community Ordination (version 4.5) [M]. Ithaca, NY USA: Microcomputer Power, 500.
[37] Ter Braak CJF. 1985. Correspondence Analysis of Incidence and Abundance Data: Properties in Terms of a Unimodal Response Modal [J]. *Biometrics*, **41**:859～873.
[38] Ter Braak CJF. 1986. Canonical correspondence analysis: A new eigenvector method for multivariate direct gradient analysis [J]. *Ecology*, **67**:1167～1179.
[39] Ter Braak CJF. 1987. The analysis of vegetation-environment relationships by canonical correspondence analysis [J]. *Vegetation*, **69**:69～77.
[40] Ter Braak CJF. 1994. Canonical community ordination. Part Ⅰ: Basic theory and linear methods [J]. *Ecoscience*, **1**:127～140.
[41] Ter Braak CJF. 1995. Canonical correspondence analysis and related multivariate methods in aquatic ecology [J]. *Aqua. Sci.*, **55**:255～289.
[42] Yee TM, Mackenzie M. 2002. Vector generalized additive models in plant ecology [J]. *Ecol. Model.*, **157**:141～156.

作者简介　朱　源,男,1983年生,博士生。研究方向为植被生态学。E-mail:zhuyuan@ires.cn

责任编辑　王　伟