

601-610

第17卷第6期
1997年11月生态学报
ACTA ECOLOGICA SINICAVol. 17, No. 6
Nov., 1997

生物群落多样性的测度方法

V. 生物群落物种数目的估计方法*

刘灿然

马克平

Q145

(中国科学院植物研究所, 北京, 100093)

A

摘要 群落的物种数目, 即物种丰富度, 是最古老、同时也是最基本的一个多样性概念, 从对它的估计中可以得到关于物种灭绝速率方面的信息, 这对生物多样性保护是非常重要的。已经提出了很多方法来估计群落中的物种数目, 这些方法可以分为两大类, 即基于理论抽样的方法和基于数据分析的方法。前者包括经典估计方法和贝叶斯估计方法; 后者包括对数正态分布的积分方法、再抽样方法和种-面积曲线的外推方法。发现, (1) 有些方法适用于动物群落, 如大多数基于理论抽样的方法; 有些方法则适用于植物群落, 如大多数基于数据分析的方法; (2) 这些方法还没有经过全面而系统地比较; (3) 还没有一个普遍认为比较好的方法。因此, 建议采用野外调查与模拟研究相结合的方法对各种估计方法进行系统地评价。

关键词: 群落, 物种数目, 物种丰富度, 估计, 参数方法, 非参数方法, 外推法。

生物群落, 多样性, 测度方法, 估计方法;

MEASUREMENT OF BIOTIC COMMUNITY DIVERSITY

V. METHODS FOR ESTIMATING THE NUMBER OF SPECIES IN A COMMUNITY

Liu Canran

Ma Keping

(Institute of Botany, Chinese Academy of Sciences, Beijing, 100093, China)

Abstract The number of species in a community, or the species richness, is the oldest and most fundamental concept of diversity. From its estimation we can acquire the information about species extinction rate, and this is important to the diversity conservation. Numerous methods have been proposed for estimating the number of species in a community. Some of them are theoretical sampling-based, and others data-analytic. The former includes frequentist and Bayesian procedures, and the latter integration of lognormal distribution, re-sampling methods (jackknife and bootstrap) and extrapolation of species-area curves. It has been found that (1) some methods are suitable to animal communities, e. g. most of the theoretical sampling-based methods, and some to plant communities, e. g. most of the data-analytic methods; (2) they have not yet been compared systematically; (3) there is not as

* 国家自然科学基金资助项目(39570125)和国家基础重大项目(85-PD-31-03)的部分内容

收稿日期: 1997-05-15, 修改稿收到日期: 1997-08-15。

yet a globally preferable method. Therefore, It is suggested that the methods should be assessed systematically through field investigations for various communities and simulations for different conditions.

Key words: community, number of species, species richness, estimation, parametric method, nonparametric method, extrapolation.

群落中总的物种数目,即物种丰富度,是最古老、同时也是最基本的一个多样性概念;它可能是所有多样性术语中最不易引起混乱的,因而它的使用频率很高^[1]。同时,生物学家和生态学家对它的估计也很感兴趣;这除了他们本身的兴趣之外,还因为从这些估计中可以得到关于物种灭绝速率方面的信息,这对生物多样性保护是至关重要的^[2,3];并且已经提出了很多估计方法,本文拟对其中一些主要的方法做一评述,其它相关的估计问题请参看有关的文献(如[3])。

在下面的讨论中,假设群落中包含 C 个物种,并且各个种的个体数所占的比例为 $\pi = [\pi_1, \pi_2, \dots, \pi_C]'$, $\sum_{i=1}^C \pi_i = 1$, 但 C 是未知的。从中抽取一个包含 n 个个体的样本,从理论上说,其结果可以用一个随机向量 $\mathbf{n} = [n_1, n_2, \dots, n_C]'$ 来表示,其中 n_i 表示样本中第 i 种的个体数, $i = 1, 2, \dots, C$ 。但是,当且仅当 $n_i > 0$ 时,第 i 种才在样本中出现。简言之, $\mathbf{n}$ 是不可观测的。然而,可观测的随机向量是 $\mathbf{c} = [c_1, c_2, \dots, c_n]'$, 其中 c_j 是样本中由 j 个个体代表的种数, $j = 1, 2, \dots, n$; 也就是说 $c_j = \# \{n_i : n_i = j\}$, $j = 1, 2, \dots, n$ 。再用 c 来表示样本中的总种数, 则有 $c = \sum_{j=1}^n c_j$, 并有 $n = \sum_{i=1}^C n_i = \sum_{j=1}^n j c_j$ 。本文就是只用 c 来估计 C 。

1 基本理论抽样的方法

1.1 经典估计方法

1.1.1 样本来自多项式分布

假设从一个个数无限的(实际上是个体很多,很难数清有多少个个体)、包含 C 个物种的群落中随机抽取 n 个个体。并假设 $\mathbf{n}$ 服从 C 维多项式分布,具有概率分布函数:

$$P_{\pi}(\mathbf{n}) = \binom{n}{n_1 \dots n_C} \prod_{i=1}^C \pi_i^{n_i}, \text{ 且 } \sum_{i=1}^C n_i = n$$

1.1.1.1 参数模型

有两类参数模型,第一类假设 π_i 是一个包含少数几个参数的某种形式的函数,即 $\pi_i = f(i; \theta, C)$, $i = 1, 2, \dots, C$, 其中 θ 是一个参数向量, f 是 i 的递减函数。Kalinin 研究了“Zipf”模型 $f_Z(i; \theta, C) = \theta/i$ 和“Mandelbrot”模型 $f_M(i; \theta, C) = \theta_i / (\theta_i + i)^{\theta_i}$ 。他对 f_Z 和 f_M 中的参数 (θ, C) 推导出了一簇基于 c 的矩法估计量,但估计量太复杂,他没有计算它们的方差。McNeil^[4]考虑过这类模型的渐近近似,还推导出了一个估计量 $\hat{C}_{MCNEIL}$ 及其渐近分布;然而,这个过程也很复杂,并且其小样本行为还不很清楚^[5]。

第二类是用包含某些参数 θ 的一个概率密度函数(pdf)来近似 π_i ($i = 1, 2, \dots, C$) 的频率分布(histogram)。Sichel 用广义的逆高斯概率密度函数(GIG pdf)推导出了一个这样的模型。它包含3个参数,但是 Sichel 在其应用中特别地固定了其中一个参数,于是产生了逆高斯 $pdf \Psi(\pi, \theta_1, \theta_2) = \theta_1 \sqrt{\theta_2} \exp(\theta_1 - \pi/\theta_2 - \theta_1^2 \theta_2 / (4\pi)) / (2 \sqrt{\pi} \pi^{3/2})$ 。Sichel 指出,如果 π_i 的分布可以由 $(\Psi(\pi; \theta_1, \theta_2))$ (对某些 (θ_1, θ_2)) 近似给出,则 $C \approx (E_{\Psi}(\pi))^{-1} = 2(\theta_1, \theta_2)^{-1}$ 。于是,他基于 (c, c_1) 给出了 (θ_1, θ_2) 的一个估计 $(\hat{\theta}_1, \hat{\theta}_2)$, 从而得到 $\hat{C}_{MCNEIL} = 2/\hat{\theta}_1, \hat{\theta}_2$ 。他还给出了 $\hat{C}_{MCNEIL}$ 的渐近偏差和方差;并通过模拟研究了它的小样本行为,发现“在对 C 的估计中出现了很大的偏差。似乎对这个特定的目的,样本大小应大于1500”。Sichel 还研究了更灵活的 GIG 分布的一个类似的应用,用最小卡平方(χ^2)拟合方法估计了其中的参数。Burrell 和 Fenton 对与0-截断的 GIG 混和的 Poisson 分布中的参数给出了极大似然估计^[3]。

在这两类模型中,似乎第二类更适合应用;它受的约束很少,因为它没有严格地对 π 加以说明,并且它

的估计理论也更完善^[3]。

1.1.1.2 非参数模型

不作关于 π 的任何假设,就不存在关于 C 的无偏估计。Engen 指出对于一组可能的总体来说, C 的任何基于 c 的估计量的偏差都是无界的,但是,Chao 用对 $E(c_i)$ 的估计得到了关于 C 的下界的一个非参数估计:

$$\hat{C}_{\text{CHAO1}} = c + c_1^2 / (2c_2)$$

并讨论了相应的 bootstrap 置信区间。Chao 和 Lee 应用“覆盖度”(coverage)这个概念,推导出了一个非参数估计量^[3]:

$$\hat{C}_{\text{CHAO2}} = c / \hat{u}_{\text{GOOD}} + \hat{\gamma}^2 n (1 - \hat{u}_{\text{GOOD}}) / \hat{u}_{\text{GOOD}}$$

其中, $\hat{u}_{\text{GOOD}}$ 是由 Good^[6]以及 Good 和 Toulmin^[7]提出的关于“覆盖度” u 的估计量, $\hat{u}_{\text{GOOD}} = 1 - c_1/n$; $\hat{\gamma}$ 是 π_i 的变异系数 γ 的估计量,他们还给出了一个相应的方差估计量,并进行了模拟研究,发现“在非等可能的情况下(即 π_i 不全相等),由于低估了 γ 而使提出的估计量一般都偏低”。尽管如此,有某些经验的和理论的证据表明,对应用目的来说, $\hat{C}_{\text{CHAO2}}$ 比 $\hat{C}_{\text{CHAO1}}$ 优越^[3]。

1.1.2 样本来自 Poisson 分布

假设样本中第 i 种的个体数是一个具有均值 λ_i 的 Poisson 随机变量, $i=1, 2, \dots, C$, 并且这些变量是独立的。在这种情况下,总的样本大小是一个具有均值 $\lambda = \sum_{i=1}^C \lambda_i$ 的 Poisson 随机变量,并且 $P_n(n) = e^{-\lambda} \prod_{i=1}^C \lambda_i^{n_i} / n_i!$ 如果再假设 $\{\lambda_i | i=1, 2, \dots, C\}$ 本身是来自某个分布 F 的一个随机样本,则 $E(c) = C(1 - P_0(F))$, 其中, $P_0(F)$ 是与 F 混合的 Poisson 随机变量等于 0 的概率。因此,给定 $P_0(F)$ 的一个估计 $\hat{P}_0(F)$, 就可以得到 C 的估计^[3]:

$$\hat{C}_{\text{POISSON}} = c / (1 - \hat{P}_0(F)).$$

常用的分布 F 是皮尔逊 III 型分布和对数正态分布,用它们导出的混合 Poisson 分布有对数级数分布、负二项分布、几何分布(即指数分布)和离散的对数正态分布。除第一个分布不能用来估计总种数 C 以外,其它几个分布都可用来估计 C ^[3]。

Ord 和 Whitmore 还研究了这样一个模型,其中的分布 F 是逆高斯分布 $\Psi(\theta_1, \theta_2, \cdot)$, 他们用 $\hat{P}_0(F) = P_0(\hat{\theta}_1, \hat{\theta}_2)$ 推导出了 $\hat{C}_{\text{POISSON}}$, 其中, $(\hat{\theta}_1, \hat{\theta}_2)$ 是 (θ_1, θ_2) 基于 c 的极大似然估计。他们分析了几组数据,并得出结论“对 Poisson-逆高斯分布的拟合揭示出,在某些情况下,它可能是描述种-多度数据的一个适当的模型,但并不总是这样”。

Zelterman 给出了 $P_0(F)$ 的基于 c 的一簇估计量,它们对 F 来说是稳健的,即它们的方差不是很大。其中最简单的一个是 $\hat{P}_0(F) = \hat{P}_0 = \exp(-2c_2/c_1)$, 从而根据 $\hat{P}_0$ 得到 $\hat{C}_{\text{POISSON}}$ 。Zelterman 讨论了这些估计量的渐近性、稳健性和有效性,并得出结论,“尽管不主张用这些估计量去代替标准的技术,但有一点却很清楚,这就是,它们在某些情况下可能是有用的”。还没有将这些估计量与 Ord 和 Whitmore 的估计量做过比较^[3]。

1.1.3 样本来自多重贝努利分布

假设分 n 次或由 n 个人分别观察一个群落,在每一次,每一个种或者被观察到,或者未被观察到。这个模型起源于标志重捕实验,但是用它来估计 C , 样本可以用一个 $C \times n$ 矩阵 $[X_{iv}]$ 来表示,其中 $X_{iv} = 1$ 表示第 i 种在第 v 次观察到了, $i=1, 2, \dots, C$; $v=1, 2, \dots, n$, 只有至少具有一个“1”的行是可观测的,事实上, c_j 就是刚好有 j 个“1”的行数。Burnham 和 Overton^[9,10] 研究了一个模型,其中 X_{iv} 是相互独立的,并且 $P\{X_{iv}$

$= 1\} = \pi_i, v=1, 2, \dots, n$ (即在每一次观察时概率都是一样的)。每一个种对样本的贡献 $n_i = \sum_{v=1}^n X_{iv}$ 是一个服

从二项分布 $B(n, \pi_i)$ 的随机变量, $p_n(n) = \prod_{i=1}^C \binom{n}{n_i} \pi_i^{n_i} (1 - \pi_i)^{n - n_i}$ 。假设 $\{\pi_i | i=1, 2, \dots, C\}$ 是来自某个分布 F 的一个随机样本,他们推导出了 C 的一个 k 阶 jackknife 估计量 $\hat{C}_{\text{BOK}}$, 例如:

$$\hat{C}_{\text{BOI}} = c + c_1(n-1)/n$$

Burnham 和 Overton^[10]根据 F 计算了 $\hat{C}_{\text{BOI}}$ 的均值和方差,并给出了选择最优 k 值的方法。对这个模型,Chao^[11]建议用 $\hat{C}_{\text{CHAO1}}$,而 Yip^[12]取 F 为 β 分布,用鞅论(Martingale theory)推导出了一个关于 C 的估计方程,并得到了相关的方差估计量。

Doss 和 Sethuraman^[14]指出,当不存在无偏估计量时,试图(例如用 jackknife 方法)来减小有偏估计量的偏差,势必会导致方差的增大;事实上, $\hat{C}_{\text{BOI}}$ 具有交错级数的形式,可能是不稳定的。Esty^[14]指出 $\hat{C}_{\text{BOI}}$ “强烈地依赖于对 $1/X$ 的一个很不准确的多项式逼近”,并且“不应该应用它”。另一方面,Chao^[11]的模拟研究显示,在某些情况下 $\hat{C}_{\text{CHAO1}}$ 比 $\hat{C}_{\text{BOI}}$ 优越,在另一些情况下则不然。Yip 通过模拟,将他的方法与 $\hat{C}_{\text{BOI}}$ 和 $\hat{C}_{\text{CHAO1}}$ 做了比较,得出结论,“鞅方法的主要优势在于计算简便,概念简单且灵活”。但是,Yip 的方法还很不完善,有很多问题未解决,例如, β 分布的参数估计。Bickel 和 Yahav 以及 Chao 和 Lee^[15]对这个模型的替代形式及其扩展做过研究,但还不能对其结果做出定论。

1.2 贝叶斯估计方法

1.2.1 基于个体抽样的方法

Efron 和 Thisted^[16]采用了一种“非参数经验贝叶斯”途径。他们基于负二项分布模型和欧拉变换方法得到了未观察到的种数的几个估计量。他还建立了一个线性规划模型,在一些来自 c 的约束下,通过使 $CP_c(F)$ (这是未观察到的种数的期望值)达到最小而得到一个值 C^* 和一个分布 F^* ,从而得到 C 的下界的一个估计:

$$\hat{C}_{\text{ET}} = c_{\text{adj}} + C^* P_c(F^*)$$

其中, c_{adj} 是 c 的一个修正值。Efron 和 Thisted 发现,尽管这个下界是“合理地保守,但没有参数模型,数据也只能给出关于 C 的很少的信息”。

Hill^[17]考虑了有限总体(N 为总的个体数, N_i 为第 i 种的个体数)的情况,采用了关于 $(C, N_1, \dots, N_c)$ 的先验分布:

$$\prod (C, N_1, \dots, N_c) = \prod_{N_1, \dots, N_c | C, N} (N_1, \dots, N_c | C, N) \prod_{C, N} (C, N) = \left(\frac{N-1}{C-1} \right)^{-1} \prod_{C, N} (C, N)$$

其中, $\prod_{C, N} (C, N)$ 是定义在 $N \times N$ 上的任意分布。Hill 计算了二元后验分布 $P_{C, N | c} (C, N | c)$, 即 $P_{C, N | c} (C, N | c) = P_{C, N | c} (C, N | c)$ 。他只对 $P_{C, N | c} (C, N | c)$ 感兴趣,并认为可以通过后验众数得到 C 的估计。

让 N 固定,并取 $\prod_{C, N} (C, N) = \prod_c (C)$ 为一个截断的负二项分布,使得 $\prod_c (C) \propto \left(\frac{C + \theta_1 - 1}{C} \right)^{\theta_2} \theta_2, C = 1, \dots, N; \theta_1 \in (0, \infty), \theta_2 \in (0, 1]$ 。Hill (1979)^[17]令 $N \rightarrow \infty$, 而得到了无限总体情况下的后验分布。然而, Hill 最后的分析仅限在物种数目的一个均匀的先验分布上;并声称该方法给出了相当好的近似^[17]。Lewins 和 Jones^[18]直接从无限总体的假设出发,扩展了 Hill 的结果。他采用了先验分布 $\prod_{C, \pi} (C, \pi) = D(\pi; C, \theta_3) \prod_c (C)$, 其中, $D(\pi, C, \theta_3)$ 为具有参数 $\theta_3 \in (0, \infty)$ 的 C 维对称 Dirichlet 分布密度, $\prod_c (C)$ 是定义于 N 上的一个任意的先验分布。据此,他们推导出了后验分布 $P_{C, c} (C | c) = P_{C, c} (C | c)$;特别地,如果 $\prod_c (C)$ 是 0-截断的负二项分布,则

$$P_{C | c} (C | c) = P_{C | c} (C | c) \propto \theta_2 \left(\frac{C + \theta_1 - 1}{C} \right)^{\theta_2} \left(\frac{C}{c} \right)^{\theta_3 C + n - 1}$$

Lewins 和 Joanes 将 $P_{C | c} (C | c)$ 的众数作为对 C 的估计。他们发现“该模型对 θ_1 和 θ_2 很稳健”,并且“很明显地, θ_3 的变化对该模型具有最大的影响”,他们还讨论了各种处理未知参数的方法。

Keener, Rothman 和 Starr^[19]采用一种“经验贝叶斯”途径。假设 $\pi \sim D(\pi; C, \theta_3)$, 其中, θ_3 已知,他们计算了 c 的无条件分布 $P(c; C, \theta_3)$ 。据此,他们推导出了 C 的极大似然估计量和其它的估计量。他们也考虑了 θ_3 和 C 的联合估计,发现在某些条件下,这些估计对 θ_3 的选择(当它给定时)是敏感的,并且对 θ_3 的估计可能给出不满意的结果,尽管在其它条件下,似乎这些估计量表现得很好。

Solow^[20]提出了对物种数目 C 的贝叶斯估计方法,他假设物种相对多度向量的概率分布服从“分割线段”模型(broken stick model),通过相对多度向量的模拟而得到 C 的贝叶斯估计。由于理论上和经验上的支持,再加上“分割线段”模型不包含任何参数,因此,该方法避免了利用 Dirichlet 分布所带来的问题。

1.2.2 基于样方抽样的方法

对个体的随机抽样实际上是很难实施的,有时甚至是不可能的。而在很多实际的情况下,用样方抽样往往更可行。假设将一个区域分成 N 个等面积的互不相交的子区或样方,从中抽取一个包含 n 个样方的样本, $n < N$ 。Mingoti 和 Meeden^[21]采用经验贝叶斯途径,用存在-不存在的样方数据来估计该区域中的总种数。用于推导这些估计量的概率模型实际上是 Efron 和 Thisted^[22]的模型在样方抽样下的一种变形。这些估计量在实际中很容易应用,并可得到很好的点估计。由于它们考虑了样本的相对大小(与整个区域相比),因而比 jackknife 和 bootstrap 估计量的行为要好。他们还给出了这个经验贝叶斯估计量的标准差的一个近似估计,并研究了基于这个估计的置信区间的行为。

2 基于数据分析的方法

2.1 基于对数正态分布拟合的方法

Preston^[22]发现 $(\log_2 j, c_j)$ 的图形常常与正态曲线很相似;这条曲线是左端截断的,因为最左端可以得到的点是 $(0, c_1)$ 。通过对观察到的图形拟合正态曲线,并将该曲线外推到左边,然后在 $(-\infty, +\infty)$ 上积分可以得到 C 的估计。Slocomb 等^[23]提出通过极大似然估计用种多-度数据拟合截断对数正态分布的方法,发现该方法不是很满意,Palmer^[24,25]通过已知的 C 研究了这个方法在植物生态学研究中的适用性,发现其结果并不比 c 本身好多少。

Slocomb 和 Dickson 认为除非样本很大(>1000 个个体)并且观察到的物种数目不低于群落中总物种数的80%,否则,用这种方法得到的物种数目的不可靠的估计将成为一个严重的问题^[26]。而 Baltanás^[26]根据模拟研究的结果,认为该方法并不比 jackknife 方法差,实际上该方法的精度最高。

2.2 再抽样方法

Heltshie 和 Forresdter^[27]研究在样方取样的情况下,用刀切法估计群落中物种数目的问题。刀切估计是“单一一种”(只在一个样方中出现的种)的数目的一个函数,即:

$$\hat{C}_{HF} = c + [(L-1)/L]k$$

其中, L 是抽取的样方数, k 是样本中出现的单一一种的总数。 $\hat{C}_{HF}$ 的方差为:

$$\text{Var}(\hat{C}_{HF}) = \frac{L-1}{L} \left(\sum_{j=1}^c j^2 f_j - k^2/L \right)$$

其中, f_j 为包含 j 个单一一种的样方数。近似的95%置信区间为:

$$\hat{C}_{HF} \pm t_{L-1, 0.025} [\text{Var}(\hat{C}_{HF})]^{1/2}$$

他还通过模拟研究了样方大小、样方数量和抽样面积对刀切估计行为的影响。

陈华豪^[28]将这一方法用于估计云冷杉红松林林下植物的物种数目的研究。当 jackknife 置信区间的左端小于样本中观察得到的种数时,他建议用截尾分布来求取一个不对称的估计区间。

Smith 和 van Belle^[29]讨论了在样方抽样下,用 jackknife 和 bootstrap 估计物种数目的问题。他们给出了估计公式,并用个体随机分布的模型对它们进行了评价。尽管在有大量的稀疏种和抽取少量的样方时, jackknife 和 bootstrap 低估了真实的物种数目,然而,它们还是降低了偏差。当抽取少量的样方时, jackknife 给出较好的估计;当抽取大量的样方时, jackknife 趋于过高地估计物种数目,而 bootstrap 行为较好。

Palmer^[24,25]在利用外推来估计物种数目时,比较了4种方法。通过1200个样方资料的计算,认为非参数估计量(一、二阶 jackknife 估计量和 bootstrap 估计量)比其它几个估计量都好。Baltanás^[26]认为 jackknife 方法与2.1节中的方法结果很相似。

2.3 种-面积曲线的外推

对于前述的很多模型原则上都可以导出一个“种-面积”曲线;也就是,观察的种数的期望值作为样本大小 n 的一个函数(记为 $E(c^{(n)})$)的图形(一般认为样本大小 n 与面积成正比)。例如,对于无约束的多项式

模型,有:

$$E(c^{(n)}) = C - \sum_{i=1}^C (1 - \pi_i)^n$$

注意,随着 n 的增大, $E(c^{(n)})$ 趋近于 C , 这意味着, 给定这样的曲线形式, 以及几个样本大小时 c 的观察值, 就能够通过外推来估计 C , 而不必考虑理论抽样模型。换句话说, 假定只假设: (a) $E(c^{(x)}) = f(x; \theta)$, 其中, x 是样本大小的一个度量 (不必是个体数), θ 是一个参数向量, f 是 x 的一个给定的递增有界函数; (b) $\lim_{x \rightarrow \infty} f(x; \theta) = C (< \infty)$ 。如果可以通过数据 $c^{(x_1)}, c^{(x_2)}, \dots$ (其中 $c^{(x_i)}$ 是具有样本大小 X_i 的 c 的一个观察值) 得到 θ 的一个估计 $\hat{\theta}$, 那么 $\lim_{x \rightarrow \infty} f(x; \hat{\theta})$ 就是 C 的估计^[3]。

Brainerd 提出了几个这样的 f , 其中最简单的一个就是:

$$E(c^{(x)}) = f_{BR}(X; \theta) = (1 - \theta^x) / (1 - \theta)$$

其中, $\theta \in (0, 1)$, x 是个体数。于是, $\lim_{x \rightarrow \infty} f_{BR}(x; \theta) = 1 / (1 - \theta)$ 。通过估计 θ , 而得 $\hat{C}_{BR} = 1 / (1 - \theta)$ 。

Stout 和 Vandermeer^[30] 基于 Jenzen^[31] 使用的一般技术利用种-面积曲线的概念导出了一个模型。它与 Pielou^[32] 的收集者曲线很类似, 并且相信物种数目随抽样面积的增加而增加, 最后趋向于一个极限值, 这个值可以认为是与取样面积无关的群落总的物种数目^[26]。他们给出的模型为:

$$E(c^{(x)}) = \theta_1 / (x^{-\theta_2} + \theta_1 / C) \quad (1)$$

其中, $\theta_1 (> 0)$ 是最大迁入速率, $\theta_2 (> 0)$ 是与岛屿面积和灭绝速率相关的一个比例因子, C 为种源库中的物种总数。(1) 式是根据 MacArthur 和 Wilson 1967 年的岛屿生物地理学的公式导出的。然而, Baltanás^[28] 认为没有必要假设岛屿生物地理学理论的所有前提条件都成立, 而从最流行的种-面积关系模型 (即幂函数模型)^[33~34] 也可以导出来。

$$E(c^{(x)}) = \theta_1 x^{\theta_2} \quad (2)$$

这只需在 (2) 中引入一个校正因子 $(1 - c^{(x)} / C)$ 而使曲线在较大面积时变得更平缓, 即 (2) 式变为

$$E(c^{(x)}) = \theta_1 x^{\theta_2} (1 - c^{(x)} / C) \quad (3)$$

将 (3) 变形即得到 (1)。

从 (1) 可以看出, $\lim_{x \rightarrow \infty} E(c^{(x)}) = C$, 即 C 就是要估计的群落的物种数目, 亦即 $\hat{C}_{SV} = C$ 。模型 (1) 中的参数 (包括 C) 可以通过最小二乘法用观察数据估计出来, 期望得到特定群落物种数目一个合理的估计。这需要面积数据要有一个相当大的跨度。然而, 在野外研究中, 常常只有几个样本单元 (即样方), 并且还都是同样大小。为了解决这个问题, 可以从原始数据中通过有放回的再抽样方法得到物种数随面积逐渐增加的数据。这个过程必须重复很多次 (上百次甚至更多), 再将结果平均^[28]。Stout 和 Vandermeer^[30] Minshall 等^[37] 等都采用了这个方法, 尽管他们采用的再抽样方法稍有不同。然而, Baltanás^[26] 的模拟研究认为, 与 2.2 和 2.3 节的方法相比, 该方法的效果最差。

de Capraais 等^[38] 提出了一个“双曲线”模型:

$$E(c^{(x)}) = f_{DCLH}(x; \theta) = x\theta_1 / (1 + x\theta_2) \quad (4)$$

其中, $\theta_1, \theta_2 > 0$ 。在他们的应用中, x 是砂子样本的重量。这里 $\lim_{x \rightarrow \infty} f_{DCLH}(x; \theta) = \theta_1 / \theta_2$ 通过将上述方程 (1) 两端取倒数, 而使其变成关于 $1/\theta_1$ 和 θ_2/θ_1 的线性方程; 再通过线性回归方法而得到 (θ_1, θ_2) 的估计 $(\hat{\theta}_1, \hat{\theta}_2)$, 进而得到 $\hat{C}_{DCLH} = x\hat{\theta}_1/\hat{\theta}_2$ 。这个方法的一个基本假设是, 个体在样本中是均匀分布的。显著地偏离这个假设将会导致对物种数目的过高估计。de Capraais 等^[39] 讨论了样本大小与物种数目估计精度之间的关系。

Kalantar^[40] 进一步分析了方程 (1) 中的参数的估计问题。可以通过对方程 (4) 的 4 种变形来估计参数和物种数目:

$$1/f_{DCLH} = \theta_2/\theta_1 + (1/\theta_1)(1/x) \quad (5)$$

$$f_{DCLH} = \theta_1/\theta_2 - (1/\theta_2)(f_{DCLH}/x) \quad (6)$$

$$f_{DCLH}/x = \theta_1 - \theta_2 f_{DCLH} \quad (7)$$

$$x/f_{DCLH} = 1/\theta_2 + (\theta_2/\theta_1)x \quad (8)$$

他的实例计算及 Dowd 和 Riggs^[41] 的结果都认为, 应该应用方程 (6)、(7)、(8), 而不是 (5)。Dowd 和 Riggs

用非加权最小二乘法,通过对几百套模拟数据的计算,认为用方程(5)估计的参数的偏差最大,而用方程(6)、(7)、(8)估计的参数的偏差最小。Kalantar 还提出了参数估计的非参数方法,这个方法不要求对误差分布的任何假设,也不要求数据的独立性。

当个体在群落中随机分布时,可以用 de Capraiiis 等^[36]的方法得到物种数目的准确估计。但是,当个体是集簇分布时,这个方法过高地估计了物种数目。de Capraiiis 和 Lindeman^[42]提出将独立变量从位置变换成物种的分布范围(species range),这样可以去除数据中明显的不一致性,即来自同一群落的不同样本,经过变换以后,就可给出一致的结果^[43]。

de Capraiiis^[43]比较了两种估计物种数目的方法:基于种-面积曲线(双曲线)的方法和基于由统计抽样理论导出的对数模型方法。其中,第二种方法是由 Root 和 Schuenemeyer 推导出的。对于他们的数据集,两种方法对物种数目给出同样准确的估计。他认为对数模型有如下优点:(1)它是基于统计抽样理论的;(2)当得到新的信息时,可以自动地调整其估计值;(3)不要求关于物种数目的极限值的假设。它的缺点是对参数 A_m (最大有效抽样面积)的估计常常有很大波动,直至抽到大的样本。双曲线模型的优点是:(1)容易使用;(2)只需抽取相对小的样本就可以得到准确的估计;(3)有时可以在斑块环境中应用。其缺点是:(1)它是经验性的;(2)要求关于物种数目的极限值的假设。这两个模型都没有关于群落结构的清楚的假设,但是,当个体的空间分布是均匀分布时,都可以对物种数目给出很好的估计。

类似的模型还有 Archibald^[44]提出的 logistic 曲线(9)(该模型等价于模型(1)、Miller 和 Wiegert^[45]使用的负指数模型(10)、Buys 等^[46]提出的模型(它实际上是 Michaelis-Menton 模型(11)的一个变形)以及 Williams^[47]的极值函数模型(12):

$$E(c^{(x)}) = \theta_1 / (x^{-\theta_2} + \theta_3) \quad (9)$$

$$E(c^{(x)}) = \theta_1 (1 - e^{-\theta_2 x}) \quad (10)$$

$$E(c^{(x)}) = \theta_1 - \theta_3 e^{-\theta_2 x} \quad (11)$$

$$E(c^{(x)}) = \theta_1 [1 - \exp(-\theta_2 x^{\theta_3})] \quad (12)$$

在这些模型中,(4)是(9)的特例(即在(9)中,令 $\theta_2 = 1$,再将 θ_3 换成 θ_2 即可),(10)是(11)的特例(即在(11)中令 $\theta_3 = \theta_1$ 即可)这些模型(9)~(12)给出的物种数目的估计量分别为:

$$\begin{aligned} \hat{C}_A &= \hat{\theta}_1 / \hat{\theta}_3 \\ \hat{C}_{MW} &= \hat{\theta}_1 \\ \hat{C}_B &= \hat{\theta}_1 \\ \hat{C}_W &= \hat{\theta}_1 \end{aligned}$$

Bunge 和 Fitzpatrick^[3]认为很难对曲线拟合方法的潜力过于乐观,因为如果函数 $f(x; \theta)$ 是从抽样模型导出的,那么抽样理论会对 C 给出更有效的估计;否则,其函数形式似乎很难得到证实。

3 讨论与结论

3.1 物种数目的很多估计方法都是基于对个体的随机抽样,如基于理论抽样的经典估计方法、基于个体抽样的贝叶斯方法和基于对数正态分布拟合的方法。这些方法对动物可能是合适的,因为它们都是离散的个体;而对植物是不适合的。而其它的方法,如再抽样方法、基于种-面积曲线拟合的方法以及 Mingoti 和 Meeden^[21]提出的贝叶斯估计方法,是基于样方抽样的,从而对植物是适合的^[24]。

在应用这些方法时一定要慎重,因为群落的一些特征(如均匀度、密度-单位面积中的个体数以及聚集水平)都会显著地影响估计的可靠性^[28]。

3.2 贝叶斯方法的一个优点就是,它以一种严密的方法将科学信息结合进统计推断之中。然而,在实际中,这一优点往往以牺牲数学上的可处理性为代价。在应用贝叶斯方法时的一个关键的问题就是,要指定给定 $C=c$ 时相对多度向量的先验分布。标准的方法都假定这个向量有一个特定的分布——对称 Dirichlet 分布。尽管这个分布在数学上是方便的,但其应用有两方面的问题。第一,它作为一个多度分布模型没有理论上或经验上的证实;第二,需要预先指定它的参数,这个参数本质上控制了多度分布的均匀程度。而 Solow^[20]用“分割线段”模型作为先验多度分布。这个模型没有显式表达式,而只提供了产生多度向量的算

法。它除了有理论上和经验上的支持以外,一个最大的特点就是它不包含任何参数,从而避免了在用 Dirichlet 分布时产生的问题。然而,这些方法都是基于个体随机抽样的,向非个体随机抽样的扩展还是一个需要进一步研究的领域。同时,关于 C 的推断可以依赖于样本中不同种的观察多度,而不仅仅依赖于观察的种数。如果取对称的 Dirichlet 分布作为先验多度分布,则给定观察多度时 C 的后验分布仅仅依赖于观察的种数,还需要进一步的研究来建立这个结果成立的一般条件^[20]。

在贝叶斯方法中,值得推荐的是由 Mingoti 和 Meeden 提出的方法。他们的估计量只需存在-不存在数据,而不需要多度数据,又是基于样方抽样。所以,在实际情况下,很容易应用。同时,它们又能提供比 jack-knife 和 bootstrap 估计量还好的结果^[21]。

3.3 非参数方法(jackknife 和 bootstrap)的一个缺点是其估计的种数不超过观察种数的2倍,因此,只有在观察种数相当大,或者与总种数相比,稀疏种的数目不是太大的情况下,用这些方法才会得到好的结果。这些方法的优点是,它们只需要存在-不存在数据,而不是多度数据,同时又是基于样方抽样,并且也没有关于种在样方内相互作用的任何假设^[19]。

3.4 种-多度模型特别要求种间的独立性假设,并且对数据拟合正确的种-多度模型可能还存在着问题,同时适当地估计种-多度模型中的参数时可能还需要大量的样本,而获得这些样本是要付出相当大代价的^[29]。因此,这样的方法很不实用。

3.5 基于种-面积曲线的方法很容易应用,但它们往往缺乏理论依据,应当慎重使用。当个体在空间中随机分布或均匀分布时,这种方法可以得到物种数目的准确估计^[42,43]。还必须指出,基于种-面积关系的任何模型对群落物种数目所作的估计的精度都依赖于样本的代表性、抽样尺度以及所涉及的分群。当用观察数据去拟合这些模型并通过外推对物种数目所作的估计是在“群落特征和环境异质性保持不变,只是面积在扩大”这个条件下做出的^[44]。因此,在应用时应注意不要外推到抽样区域之外,并尽量采用随机抽样^[45]。

3.6 已经通过与一个或少数几个地方观察的物种数目(C)的比较或者通过模拟评价了一些估计量的行为。但第一种途径存在的问题是所用的样本量不足以说明一个估计量的效力;第二种途径存在的问题是计算机产生的格局可能没有真正模拟实际情况。可以通过在很多地点对真实群落大量地抽样来解决这些问题。通过将物种数目的一些估计值与真值相比较,可以确定估计量是否有偏或者估计量是否与真值高度相关^[24]。因此,建议采用下列3条途径来确定合适的方法:(1)在各种外貌、空间结构和物种组成的群落中进行比较研究;(2)在对其区系已相当熟悉的群落中评价各种方法;(3)采用模拟方法,选择各种不同的算法和参数对估计方法进行评价。

参考文献

- 1 Peet R K. The measurement of species diversity. *Annual Review of Ecology and Systematics*, 1974, **5**: 285~307
- 2 Mann C C. Extinction: are ecologists crying wolf? *Science*, 1991, **253**: 709~824
- 3 Bunge J and Fitzpatrick M. Estimating the number of species: a review. *Journal of the American Statistical Association*, 1993, **88**: 364~373
- 4 McNeil D. Estimating an author's vocabulary. *Journal of the American Statistical Association*, 1973, **68**: 92~96
- 5 Chao A and Lee S-M. Estimating the number of classes via sample coverage. *Journal of the American Statistical Association*, 1992, **87**: 210~217
- 6 Good I J. The population frequencies of species and the estimation of population parameters. *Biometrika*, 1953, **40**: 237~264
- 7 Good I J and Toulmin G H. The number of new species and the increase in population coverage, when a sample is increased. *Biometrika*, 1956, **43**: 45~63
- 8 Pielou E C. 卢泽愚译. 数学生态学(第二版). 北京: 科学出版社, 1988
- 9 Burnham K P and Overton W S. Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, 1978, **65**: 625~633
- 10 Burnham K P and Overton W S. Robust estimation of population size when capture probabilities vary among animals.

- Ecology*, 1979, **60**: 927~936
- 11 Chao A. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, 1987, **43**: 783~791
 - 12 Yip P. A martingale estimating equation for a capture-recapture experiment in discrete time. *Biometrics*, 1991, **47**: 1081~1088
 - 13 Doss H and Sethuraman J. The price of bias reduction when there is no unbiased estimate. *The Annals of Statistics*, 1989, **17**: 440~442
 - 14 Esty W W. The efficiency of Good's nonparametric coverage estimator. *The Annals of Statistics*, 1986, **14**: 1257~1260
 - 15 Chao A and Lee S-M. Estimating the number of unseen species with frequency counts. *Chinese Journal of Mathematics*, 1990, **18**: 335~351
 - 16 Efron B and Tibshirani R. Estimating the unseen species, how many words did Shakespeare know? *Biometrika*, 1976, **63** (3): 435~447
 - 17 Hill B M. Posterior moments of the number of species in a finite population and the posterior probability of finding a new species. *Journal of the American statistical Association*, 1979, **74**: 668~673
 - 18 Lewins W A and Joanes D N. Bayesian estimation of the number of species. *Biometrics*, 1984, **40**: 323~328
 - 19 Keener R, Rothman E and Starr N. Distributions of partitions. *The Annals of Statistics*, 1987, **15**: 1466~1481
 - 20 Solow A R. On the Bayesian estimation of the number of species in a community. *Ecology*, 1994, **75** (7): 2139~2142
 - 21 Mingoti S A and Meeden G. Estimating the total number of distinct species using presence and absence data. *Biometrics*, 1992, **48**: 863~875
 - 22 Preston F W. The commonness and rarity of species. *Ecology*, 1948, **29**: 254~283
 - 23 Slocumb J, Stauffer B and Dickson K L. On fitting the truncated lognormal distribution to species-abundance data using maximum likelihood estimation. *Ecology*, 1977, **58**: 693~696
 - 24 Palmer M W. The estimation of species richness by extrapolation. *Ecology*, 1990, **71** (3): 1195~1198
 - 25 Palmer M W. Estimating species richness: the second-order jackknife reconsidered. *Ecology*, 1991, **72** (4): 1512~1513
 - 26 Baltanas A. On the use of some methods for the estimation of species richness. *Oikos*, 1992, **65** (3): 484~492
 - 27 Heltche J F and Forrester N E. Estimating species richness using the jackknife procedure. *Biometrics*, 1983, **39**: 1~11
 - 28 陈华豪. 刀切法在生态学中的应用. *东北林学院学报*, 1984, **12** (增刊): 134~143
 - 29 Smith E P and van Belle G. Nonparametric estimation of species richness. *Biometrics*, 1984, **40**: 119~129
 - 30 Stout J and Vandermeer J. Comparison of species richness for stream-inhabiting insects in tropical and midlatitude streams. *American Naturalist*, 1975, **109**: 263~280
 - 31 Jenzen D J. Sweep samples of tropical foliage insects, effects of seasons, vegetation types, elevation, time of day, and insularity. *Ecology*, 1973, **54**: 687~708
 - 32 Pielou E C. *Ecological diversity*. New York: Wiley, 1975
 - 33 Goodell D W. Quantitative aspects of plant distribution. *Biological Review*, 1952, **27**: 194~245
 - 34 Hopkins B. The species-area relations of plant communities. *Journal of Ecology*, 1995, **43**: 409~426
 - 35 Kilburn P D. Analysis of the species area relation. *Ecology*, 1996, **47**: 831~843
 - 36 Connor E F and McCoy E D. The statistics and biology of the species-area relationship. *American Naturalist*, 1979, **113**: 791~833
 - 37 Minshall G W, Petersen R C and Nimz C F. Species richness in streams of different size from the same drainage basin. *American Naturalist*, 1985, **125**: 16~38
 - 38 de Capraia P, Lindemann R H and Collins C M. A method for determining optimum sample size in species diversity studies. *Mathematical Geology*, 1976, **8** (5): 575~581
 - 39 de Capraia P, Lindemann R H and Haines R. A relationship between sample size and accuracy of species richness predictions. *Mathematical Geology*, 1981, **13** (4): 351~355
 - 40 Kalantar A H. Using the hyperbolic model to estimate species richness. *Mathematical Geology*, 1987, **19** (2): 151~154

- 41 Dowd J E and Riggs D S. A comparison of estimates of Michaelis-Menton kinetic constants from various linear transformations. *Journal of Biological Chemistry*, 1965, **240**(2): 863~869
- 42 de Capraais P and Lindemann R H. Speices richness in patchy environments. *Mathematical Geology*, 1978, **10**(1): 73~90
- 43 de Capraais P. Estimating species diversity: comparison of two algorithm. *Mathematical Geology*, 1984, **16**(3): 237~248
- 44 Archibald E E A. The specific character of plant communities. II. A quantitative approach. *Journal of Ecology*, 1949, **37**: 260~274
- 45 Miller R L and Wiegert R G. Documenting completeness, species-area relations and the species-abundance distribution of a regional flora. *Ecology*, 1989, **70**(1): 16~22
- 46 Buys M H, Maruz J S, boucher C and et al. A model for species-area relationships in plant communities. *Journal of Vegetation Science*, 1994, **5**: 63~66
- 47 Williams M R. An extreme value function model of the species incidence and species-area relations. *Ecology*, 1995, **76**(8): 2607~2616
- 48 He F and Legendre P. On species-area relations. *American Naturalist*, 1996, **148**(4): 719~737
- 49 Evans F C, Clark P J and Brand R H. Estimation of the number of species present on a given area. *Ecology*, 1955, **36**(2): 342~343