基于 Huffman 编码的群落结构复杂性

金 森

(东北林业大学林学院, 哈尔滨 150040)

摘 要 Anand 和 Orlóci (1996) 提出用 Huffman 编码的平均码长来衡量群落结构总复杂性, 用 12 阶 Rényi 熵测度群落无序结构复杂性, 用两者之差测度群落有序结构复杂性。这是一种与基于生物多样性的复杂性测度完全不同, 至少在思路不同的结构复杂性测度, 但对于这种测度的性质研究还不多。该文模拟建立了具有代表性的各种结构的群落 400 多万个, 计算了这些群落的复杂性测度, 对其性质进行了研究。结果表明: 1) 群落结构总复杂性、无序结构复杂性和有序结构复杂性受群落多样性、变异程度、优势种组成等影响。其中, 群落结构总复杂性与 Shannon-Wiener 指数高度相关(决定系数 > 0.99), 完全可由群落组成的多样性决定。群落无序结构复杂性与群落优势种对数比例或变异系数关系最大, 与群落优势种对数比例之间的决定系数 > 0.99 。2) 该群落有序结构复杂性可表示为群落生物多样性和优势种对数比例的线性组合。在生态意义上可看作主要由群落中非优势种的组成比例决定, 而组成多样性能解释 41% ~ 46% 的群落有序结构复杂性。3) 群落组成物种总数的增加会导致群落复杂性的测度有所增加, 但没有像文献(Anand & Orlóci, 1996) 中描述的那么大, 且组成总数对结构复杂性与多样性和优势种的关系等没有影响。

关键词 Huffman 编码 Shannon-Wiener 指数 Rényi 熵 结构复杂性 有序结构复杂性 无序结构复杂性

PROPERTIES OF COMMUNITY STRUCTURAL COMPLEXITY DESCRIBED BY MEAN LENGTH OF HUFFMAN CODE

JIN Sen

College of Forest Science, Northeast Forestry University, Harbin 150040, China

Abstract Aims Anand & Orlóci (1996) proposed a method measuring the total, disordered and organized complexity of community structure using the mean length of Huffman code, the 12th order of Rényi entropy, and the difference between these two measures, respectively. This method is quite different from measures based on community biodiversity and is potentially of great significance. Although some properties of the measures were discussed by Anand & Orlóci (1996), a thorough study has not been conducted, especially on their ecological meaning. The purpose of the paper is to reveal: 1) factors affecting the measures and their ecological meaning, 2) relationship between complexity and biodiversity, and 3) effect of species number on structural complexity.

Methods First, $> 4\,000\,000$ simulated communities with different structures were constructed by two methods: evenly distributed mode and dominated mode. Then the three complexity measures were computed. Correlation analyses were conducted between these measures and potentially related factors such as Shannon-Wiener index, coefficient of variation and logarithmic proportion of most dominated species. Multiple linear regressions between the complexity measures and combination of these factors were carried out to explore the relationship between complexity and biodiversity. Communities with different species number were then constructed and tested to explore the effect of species number on complexity measures.

Important findings The three complexity measures are strongly correlated with the Shannon-Wiener index, coefficient of variation and logarithmic proportion of the most dominated species. For total complexity, the correlation coefficient between it and the Shannon-Wiener index is nearly 1 (0.996), so is that between the 12th order of Rényi entropy and logarithmic proportion of the most dominated species. The determining coefficient of multiple linear regression between organized complexity and Shannon-Wiener index and logarithmic proportion of the most dominated species is also near 1 (0.97). These findings suggest that all three measures of commu-

nity structural complexity are affected by the Shannon-Wiener index, coefficient of variation and logarithmic proportion of the most dominated species. For organized structural complexity, it can be expressed as a linear combination of Shannon-Wiener index and logarithmic proportion of the most dominated species. Biodiversity can explain 4146% of variation of organized structural complexity. Although the three complexity measures increase with number of species, they do not increase as much as described by Anand & Orlóci (1996), and it does not affect the relationship between the complexity measures and related factors.

Key words complexity, structural complexity, Huffman code, Shannon-Wiener index, Rényi entropy, organized complexity, disordered complexity

复杂性的测度方法是进行复杂性研究的基础。现有生态复杂性的测度多数基于多样性 (McElhinny *et al.*, 2005)。一些学者意识到这一点,指出这些测度虽可提供关于复杂性的某些认识,但不充分 (Anand & Orlóci, 1996; 李镇清, 2000)。因此, Anand 和 Orlóci (1996) 以群落结构为研究对象, 用最小剩余码的平均码长测度结构的总复杂性, 具体采用 Huffman 编码, 用 12 阶 Rényi 熵表示群落无序结构的复杂性, 用两者之差代表有序结构复杂性。这是一种至少在思路与基于多样性的复杂性测度完全不同的复杂性测度方法。Anand 和 Orlóci (1996) 采用该方法对不同环境梯度和不同时空尺度上的群落结构复杂性进行了研究, 李镇清 (2000) 用相似的方法对中国东北部样带上的植物群落结构复杂性进行了研究。

该测度本质上是一种科尔莫哥洛夫复杂性测度, 即构造一个描述某个对象的算法, 用该算法的计算量作为该对象的复杂性 (李镇清, 2000; 吴彤, 2004)。计算复杂性在概念上完全不同于多样性, 将其引入生态结构复杂性的测度中十分必要。Anand 和 Orlóci (1996) 及李镇清 (2000) 等工作使复杂性与多样性至少从概念上有了区分, 同时表明, 用计算测度来衡量群落结构的复杂性是可行和有意义的。

Anand 和 Orlóci (1996) 曾对该测度的一些性质, 如有序结构复杂性与群落中物种数量之间的关系进行了研究。但目前关于该测度性质的系统研究还很少。本文希望回答以下问题:

1) 影响这些测度, 特别是群落有序结构复杂性测度的因子有哪些? 在确定影响因子后, 若这些测度可用一些已知指标来表示, 则其生态内涵就明确了。

2) 这些测度与多样性的关系如何? 多样性和复杂性的关系一直为人们关注, 基于多样性的复杂性测度与多样性关系自然密切。基于信息编码的复杂性与多样性在概念上不同, 与多样性的关系, 如相关程度、多样性对复杂性的影响等还需要研究。

3) 群落物种数对结构复杂性的影响如何?

Anand 和 Orlóci (1996) 曾指出, 随着物种数的增加, 有序结构复杂性将显著增加, 但本研究计算的结果与他们不同, 需要进一步研究验证。

1 研究方法

首先模拟构造若干具有代表性的不同结构的群落, 然后对其结构复杂性的测度进行分析, 找出影响因子, 建立复杂性与这些因子的关系模型, 据此来回答上述问题。

1.1 不同结构群落的模拟构建方法

Anand 和 Orlóci (1996) 的结构复杂性测度方法是针对结构特征序列 $P = (p_1, p_2, \dots, p_n)$, 其中 n 为群落的物种数, $p_i < p_j, i < j$, 且 $\sum_{i=1}^n p_i = 1$, p_i 为物种 i 的一相对指标 (相对多度、相对重要值、株数比例等), p_i 的分布决定了群落结构复杂性。设群落个体总数为 S , 由序列 $(p_1, p_2, \dots, p_n)$ 表示的群落结构是对 S 的一个 n 无序分拆。为搞清该复杂性的性质, 须建立完备的群落结构集, 即 S 的全部 n 无序分拆。但在 S 较大时, 其 n 无序分拆的数量巨大, 计算机实现起来也十分困难, 因此, 本研究根据现实群落的结构特点分别构建均匀型和优势型两类群落。均匀型群落的各物种出现的概率相等, 优势型群落中一些种出现的概率要比其它种大很多。这两种类型基本代表了实际群落的特点。

具体构建方法如下:

1) 均匀型 设群落个体数为 S , 物种数为 n , 每次产生 1 个在 $[1, n]$ 内的随机整数 i , 然后物种 i 的个体数加 1, 共产生 S 个随机数, 然后计算 p_i , 形成序列 $(p_1, p_2, \dots, p_n)$ 。

2) 优势型 设群落 n 个物种中有 m 个优势种组, $m < n$, 每个种组内有 n_i 个个体 ($i = 1, 2, \dots, m$), $\sum_{i=1}^m n_i < n$, 这些种的优势度为 d_i , 且 $d_i < d_j (i < j, i, j \in [1, m])$ 。按优势度 d_i 大小顺序为这些种组分配个体数, 第一优势度的个体数为 Sd_1 , 第二优势

度的个体数为 $S(1 - d_1)d_2$,依次类推。所余个体数在剩余的 $n - \sum_{i=1}^m m_i$ 个种中按 1) 中均匀型方法分配。采用优势种组是因为群落中常有优势度相近的多个物种 ,这等效于设定 $\sum_{i=1}^m m_i$ 个优势种 ,但可避免因随机数产生方法所造成的过于均匀。

1.2 群落结构复杂性的影响因子分析及其与多样性的关系

因为结构复杂性的计算采用相对比例 ,所以群落个体总数对复杂性影响不大。根据文献(Anand & Orlóci ,1996)设 $n = 20$, $S = 2\,000$, $m = 4$ ($m_i \in [0, 4]$) 这样对于 20 个种的群落 ,超过 16 个优势种的机会很小 ,即使有 ,也演变成均匀型了。这样 ,共有 25 个优势种组合 ,其中 $m_i = 0$ 就是均匀型。优势度 d_i 设为 0.1 ~ 0.9 ,步长为 0.1。最后共有 $5^4 \times 9^4 = 4\,100\,625$ 个群落。计算这些群落的总复杂性即 Huffman 编码的平均长度、无序结构复杂性既 12 阶 Rényi 熵及有序结构复杂性即前两者之差。

Huffman 编码的算法如下(肖自美 ,2000) : 1) 把各物种按相对比例从大到小排成一行 ; 2) 把比例最小的两个种合在一起成为一个新种 ,上面一个编为 0 ,下面一个编为 1 ,将这两个种的比例和做为新种的比例 ,再与尚未处理过的物种按比例大小重新排序 ; 3) 重复步骤 2) ,直到所有物种都处理完 ; 4) 从右向左依据编码路径返回 ,就得到各物种的码字和码长。

群落结构复杂性取决于结构指标在不同组分(物种等) 的分布 ,能够反映这些分布特点的因子对结构复杂性都可能影响 ,因此 ,影响因子取 Shannon-Wiener 指数、变异系数、最大值、对数最大值、最大值和对数最大值的乘积。用 Shannon-Wiener 指数代表群落的多样性。计算 3 种结构复杂性和这些因子之间的相关系数 ,建立上述因子的不同组合与有序结构复杂性的多元线性模型 ,通过模型分析多样性对有序结构复杂性的影响。

在多元线性回归中 ,标准回归系数反映了各自变量对因变量的影响程度 ,半偏相关系数(Semi-partial correlation coefficient) 的平方反映了预测变量对因变量变异的影响比例。这两个指标都可用来刻画多样性对复杂性的绝对贡献率。为方便不同影响因子之间的比较 ,本文采用相对贡献值 m 来代替 ,其定义为 :

$$m_i = p_i / \sum_{i=1}^n p_i$$

式中 , n 为模型中使用的自变量数目 , m_i 为第 i 个变量的相对贡献值 , p_i 为第 i 个变量的标准回归系数或半偏相关系数平方。采用不同回归参数所得结果会有所不同。

1.3 群落物种数对复杂性的影响

物种数 n 设为 10 ~ 50 , $S = 2\,000$ 。为避免计算量过大 ,设 $m = 3$ ($m_i \in [0, 3]$)。优势度 d_i 的值为 $\max(m/n, 0.1)$ 到 0.9 ,步长为 0.1。当 m_i 或超过 $\sum_{i=1}^m m_i$ 时 ,不进行分配。计算这些不同物种数的群落的各种结构复杂性 ,分析其与群落物种数的相关性。另按 Anand 和 Orlóci (1996) 的方法构造群落 ,即采用 20 个种 2 000 株 ,群落的结构从均匀(每种 100 株) 到最大优势(1 个种 1 881 株 ,其它 19 种各 1 株) ,计算群落复杂性 ,并与文献(Anand & Orlóci ,1996) 的结果比较。

群落构造和复杂性计算由 VB 6.0 编程实现 ,统计绘图由 Statistica 6.0 (StatSoft ,2001) 完成。

2 结果与分析

2.1 所构造群落的基本特征

图 1 是 1.2 中构造的 4 100 625 个群落的优势种株数和变异系数的分布。这些群落结构的变异系数分布是正态的 ,避免了因人为构造因素造成的群落结构中指数型分布过多 ,使优势种的作用被夸大的现象。

图 2 是 1.3 中构造的不同物种数的群落结构有序复杂性随优势种株数变化的情况。复杂性在不同群落中分布的随机性保证了下面分析结果的代表性。

2.2 群落结构复杂性的影响因子分析

表 1 给出了 3 种结构复杂性和影响因子的关系。其中群落结构总复杂性与 Shannon-Wiener 指数相关性接近 1 ,说明群落总复杂性受多样性影响很大 ,几乎等同于多样性。该复杂性与群落组成比例的变异系数、12 阶 Rényi 熵、最大值等的相关性也很大。群落无序结构复杂性与群落最大优势种比例关系最密切 ,与其对数比例的相关系数几乎达到 1 ,与变异系数和 Shannon-Wiener 指数的相关性也较高。群落有序结构复杂性与上述因子的相关性显著下降 ,与群落优势种的对数比例相关性最高 ,但也只有 0.55。

图1 所构建的群落结构优势种株数和变异系数的分布
Fig.1 Distribution of stem numbers of most dominant species and coefficients of variation

图2 不同结构群落的有序结构复杂性分布
Fig.2 Distribution of organized complexity of communities of different structures

2.3 组成多样性对结构复杂性的影响

以 Shannon-Wiener 指数和优势种的对数比例为自变量,以结构有序复杂性为因变量所做的多元线性回归的决定系数最大,回归效果最好,结果为:

$$C_o = L_o - R_o \tag{2}$$

$$C_o = 0.842\ 94\ SV + 0.979\ 77\ \log_2 Max + 0.379\ 77 \tag{3}$$

式中: C_o 为群落有序复杂性, SV 为 Shannon-Wiener 指数, Max 为优势种的比例。

调整后的 $R^2 = 0.973\ 13$ 。这说明方程拟合极好,接近确定性关系。

表 2 是对上述方程进行偏相关分析的结果。从标准回归系数看,多样性对结构有序复杂性的贡献为 45.62%,从半偏相关系数上看,多样性可解释为 40.99% 的结构有序复杂性变异,而优势种比例可解释为 59.01% 的变异。可以认为,组成多样性对有序结构复杂性的贡献率为 41% ~ 46%。

用 Shannon-Wiener 指数和优势种的对数比例几乎能解释 100%(> 97%) 的有序结构复杂性,这说明用这两个变量完全可以解释群落的有序结构复杂性。同样,从上述相关分析(表 1)可知,生物多样性几乎可以解释群落结构总复杂性的全部变化,而优势种可以表征群落无序结构复杂性。

2.4 群落物种数量对复杂性的影响

图 3 给出了不同物种数群落的 3 种复杂性的分布。随物种数的增加,群落结构的各种复杂性总体有增加的趋势,但远没有象 Anand 和 Orlóci(1996)所指出的那样大(图 4)。究其原因,Anand 和 Orlóci(1996)所用的编码算法有误。在 Anand 和 Orlóci(1996)的算法中,新构成的种(成分)永远是作为下一个两两合并中的一个,直接与序列中下一个未处理的物种进行合并,并没有与其它物种根据新种的比例重新排序。正确的算法要求每次产生的新种(成分)都要根据其重新计算的比例与序列中未处理的物种进行重新排序,然后由其中比例最小的两个物种再组成新的物种。

另外,对不同物种数的群落结构复杂性和影响因子的回归分析结果与 2.3 中结果相似,说明结构复杂性受生物多样性和优势种对数比例的影响,而与群落物种总数无关。

表 1 3 种群落结构复杂性与影响因子的相关性
Table 1 Correlation coefficients between structural complexities of community and effecting factors

	总复杂性 Total complexity	无序结构复杂性 Disordered structural complexity	有序结构复杂性 Organized structural complexity
平均编码长度 Mean code length	1.00	0.89	- 0.08
Shannon-Wiener 指数 Shannon-Wiener index	0.99	0.88	- 0.08
12 阶 Rényi 熵 12th order of Rényi entropy	0.89	1.00	- 0.53
平均编码长度与 SV 指数差 Difference between Row 1 and Row 3	- 0.64	- 0.58	0.07
有序结构复杂性 Organized structural complexity	- 0.08	- 0.53	1.00
变异系数 Coefficient of variation	- 0.96	- 0.96	0.30
最大比例 Proportion of dominant species	- 0.84	- 0.95	0.50
对数最大比例 Logarithmic Row 7	- 0.87	- 0.99	0.55
最大比例与其对数之积 Production of Row 6 and Row 7	- 0.07	- 0.02	- 0.07

表中所有相关系数都在 0.000 1 水平上显著 All the coefficients are significant at 0.000 1 level

表 2 对方程进行偏回归分析的结果
Table 2 Results of semi partial correlation analysis

	标准回归系数 Standardized regression coefficient	半偏相关系数 Semi partial correlation coefficient	显著水平 Significant level
Shannon Wiener 指数 Shannon Wiener index	1.68	0.82	0.00
优势种对数比例 Logarithmic proportion of dominant species	2.01	0.98	0.00

图 3 群落物种总数和群落结构复杂性的关系
Fig.3 Relationship between species number and complexity of community

3 结 论

1)群落结构的总复杂性、无序复杂性和有序复杂性受群落组成多样性、变异程度、优势种组成比例等影响。其中,总复杂性与 Shannon-Wiener 指数高度相关,表明其由群落组成的多样性决定。无序复杂性与群落优势种的对数比例或变异系数关系最大,特别是与群落优势种对数比例的关系几乎是确定性的。

2)群落有序结构复杂性可表示为群落生物多样

性和优势种对数比例的线性组合,即:

$$C_o = - 0.842\ 94 \sum_{i=1}^n p_i \log_2 p_i + 0.979\ 77 \log_2 p_i + 0.379\ 77$$

(4)

这说明该复杂性主要由群落中非优势种的组成比例决定,这可以看作是該有序结构复杂性的一种生态内涵。组成多样性能解释 41% ~ 46% 的群落有序结构复杂性。

3)群落物种数量的增加会导致群落复杂性的测度增加。但没有像文献(Anand 和 Orlóci ,1996)上描述的那么大,且组成物种总数对结构复杂性与多样

性和优势种之间的关系没有影响。

4 讨 论

1) 一般情况下 , 用 Huffman 编码描述的群落结构总复杂性、Shannon-Wiener 指数、12 阶 Rényi 熵都随群落优势种的优势度增加而增加(图 4), 但 Huffman 编码与 12 阶 Rényi 熵之差所描述的群落有序结构复杂性与之不同 , 在其它特征一致的条件下 , 从均匀分布到最极端优势分布的群落 , 有序复杂性随优势度增加而增加 , 达到极大值后随优势度增加而下降(图 2 图 4)。这是一个十分重要的性质 , 复杂性的本质是事物的可预测性。事物越易于描述 , 也易于预测。编码是描述复杂性的一种方法 , 可压缩性也是描述复杂性的一种方法。假设有两个数字序列 , 一个很长 , 如 100 个 9 , 但它能用一个简单的公式来描述 ; 另一个相对短 , 如只有 30 个数字 , 但十分随机 , 却无法用一个公式描述 , 那么后者比前者复杂 , 因为它难以压缩。群落的无序结构复杂性反映的就是群落结构特征的编码描述复杂性 , 越均匀分布 , 其无序性越强。而有序结构复杂性反映的是群落结构特征的可压缩性复杂性。平均分布和极端优势分布都可用相对简单的公式来描述 , 压缩性较大 , 复杂性小 , 而中间的一些分布随机性强 , 可压缩性小 , 复杂性大。受群落生物多样性和优势种对数比例两个因素控制是有序结构复杂性具有这种分布模式的原因。

图 4 不同编码方法计算的群落各种复杂性
Fig.4 Complexities of community computed by different methods
文献指 Anand 和 Orlóci (1996) The literature here is Anand & Orlóci (1996) TC : 总复杂性 Total complexity OC : 有序复杂性 Organized complexity

2) 上述分析是从数值关系和描述角度进行的。结构服务功能和结构复杂性的意义只有与生态过程相联系 , 其生态意义和应用价值才能进一步明确。生态过程十分复杂 , 其中一些我们比较清楚 , 可用模型来描述 , 而另一些则不清楚 , 只能采用黑箱的办法来处理。对于前类生态过程 , 当群落的初始结构状态有序时 , 其动态过程简单、易于预测 ; 当初始结构有序性低或随机性强时 , 其动态往往不确定 , 会形成复杂的格局乃至混沌。例如 , 用 LV 方程组预测的多物种的种群数量动态过程时 , 当群落中优势种群优势度很大或各种群数量比较均匀时 , 各种群数量的动态是简单的。但各种群的初始数量比较复杂时 , 其动态过程十分复杂 , 平衡性差。群落有序结构复杂性反映的就是与这类生态过程相关联、能够导致复杂格局的群落状态。有序复杂性越高 , 过程动态越复杂 , 其中的极值点可能是导致混沌等复杂现象的临界值 , 值得深入研究。对于第二类过程 , 无序结构复杂性大的群落的动态过程更难以预测 , 如各个物种数量均匀分布的群落比单优势种的群落在种间作用、对环境的适应和稳定性等生态过程和状态上要复杂得多。因此 , 其动态发展等更具不确定性 , 这就是结构无序复杂性所反映的结构特点。群落结构总复杂性是两个复杂性的总和 , 是群落对上述两类生态过程响应的可预测性的总和。

3) Shannon-Wiener 指数本身是信息编码理论上的最小平均码长 , 而 Huffman 编码的平均码长是可实现的最短平均码长(肖自美 , 2000)。从理论上讲 , 两者之间的差由群落的结构造成 , 它可能更适合于衡量群落的结构复杂性。但由于两者高度相关 , 两者的差过小。Huffman 编码的平均码长和 Rényi 熵指数的相关性相对较低 , 用其描述有序复杂性效果要好些。图 2 表明 , 有序结构复杂性的测度与结构的特征对应关系过于广泛 , 存在明显的一值多构现象 , 从数值上判断结构复杂性的具体情况还很难 , 与生物多样性的关系也十分密切 , 在此意义上讲 , 用 Huffman 编码的平均码长来描述的计算复杂性在数量关系上还没有完全脱离生物多样性 , 尽管两者的概念差别很大。因此 , 寻求在概念和数量关系上与生物多样性都有差别的计算复杂性的描述方法十分必要。

4) 本文虽然没有对可能出现的全部群落结构进行模拟 , 所得结论对某些特殊结构可能不适应 , 但由于对生物群落可能出现的大多数结构都进行了模拟 , 考虑到 4 100 625 个样本的巨大性 , 所得结论是

可靠的。

参 考 文 献

Anand M, Orlóci L (1996). Complexity in plant communities: the notion and quantification. *Journal of Theoretical Biology*, 179, 170 – 186.

Li ZQ (李镇清) (2000). The complexity and diversity of typical plant communities along the Northeast China Transect (NECT). *Acta Botanica Sinica* (植物学报), 42, 971 – 978. (in Chinese with English abstract)

McElhinny C, Gibbons P, Brack C, Bauhus J (2005). Forest and

woodland stand structural complexity: its definition and measurement. *Forest Ecology and Management*, 218, 1 – 24.

StatSoft, Inc. (2001). *STATISTICA (data analysis software system)*, version 6. www.statsoft.com

Wu T (吴彤) (2004). On concepts of complexities and its meaning in research. *Journal of Renmin University of China* (中国人民大学学报), 5, 1 – 9. (in Chinese with English abstract)

Xiao ZM (肖自美) (2000). *Theory of Image Information and Technology of Compressing and Encoding* (图象信息理论与压缩编码技术). Sunyat-Sen University Press, Guangzhou, 149 – 152. (in Chinese)

责编编委 : 周广胜 责任编辑 : 张丽赫