概述

生态系统多样性（ecosystem diverisity）可以定义为维持着复杂系统的遗传多样性、物种多样性、栖息地多样性以及功能的多样性。

物种多样性因空间而异，Whittaker提出alpha多样性是局部多样性，beta是多样性的空间变化，而gamma是区域多样性。

Alpha多样性

Alpha多样性是指一个特定区域或生态系统内的多样性，是反映丰富度和均匀度的综合指标。Alpha多样性主要与两个因素有关：一是种类数目，即丰富度；二是多样性，群落中个体分配上的均匀性。

群落丰富度（Community richness）

主要包括Chao1指数和ACE指数，均由Chao分别于1984和1993年提出。

- **Chao1**[[1]]：是用Chao1算法估计群落中含OTU数目的指数，在生态学中常用来估计物种总数。

  $$SChao1 = Sobs + n1(n1 - 1)/2(n2 + 1)$$

  其中：

  - $SChao1$为估计的OTU数，
  - $Sobs$为观测到的OTU数，
  - $n1$为只有一条序列的OTU数目，
  - $n2$为只有两条序列的OTU数目。

  Chao1指数越大，表明群落的丰富度越高。

- **Ace**[[2]]：是用来估计群落中含有OTU数目的指数，是生态学中估计物种总数的常用指数之一。默认将序列量10以下的OTU都计算在内，从而估计群落中实际存在的物种数。**ACE指数越大，表明群落的丰富度越高。**

群落多样性（Community diversity）

主要包括Shannon指数和Simpson指数

- **Shannon指数**[[3]]
  一个取样单元中物种多度的向量可以被看做是一个定性变量，其中每个物种是一个状态，多度分布是观测的频率分布。在这个逻辑下，这个定性变量的离散度（dispersion）可以通过此样方中q个物种每个物种的相对频度p_i来计算。

$$H = -\sum_{i=1}^{q} p_i log p_i$$

Shannon指数会随着物种数量的增加而增加，相对频度也在起作用。实际上Shannon指数是考虑两个组分贡献：

- 物种数（物种丰度度）
- 物种频度分布的均匀度（evenness）或均匀性（equitability）

对于任意数量个体数的样方，当所有物种多度一样的时候，H值最大：

$$H_{max} = -\sum_{i=1}^{q} \frac{1}{q} log \frac{1}{q} = logq$$

- **Pielou指数**[[4]] Pielou均匀度J可以定义为下面的公式：

$$J = \frac{H}{H_{max}}$$

其中H为Shannon指数，H_{max}为当所有物种多度一样时的Shannon指数。 Pielou均匀度数值与Shannon指数公式中对数的底（base）设为多少无关。尽管Pielou均匀度一直被认为是严重依赖于物种丰富度的不好的指标，但它的确是在生态学文献中使用最广泛的均匀度指数。

- **Simpson指数**[[5]] 是用从样方中随机抽取两个个体属于同一物种的概率来表示：

$$\lambda = \sum_{i=1}^{q} \frac{n_i(n_{i-1})}{n(n-1)} = \frac{\sum_{i=1}^{q} n_i(n_i-1)}{n(n-1)}$$

这里的q是物种的数量，当n很大时，n_i趋近于（n-1），公式可以简化为：

$$\lambda = \sum_{i=1}^{q} (\frac{n_i}{n})^2 = \sum_{i=1}^{q} p_i^2$$

事实上，当两个个体同属同一个物种的概率比较大的时候（即物种丰度度低的时候），这个数值就比较大。因此通常是使用了转换形式的指数：$D = 1 - \lambda$（**Gini-Simpson指数**[[6]]）或$D = 1/\lambda$（**逆Simpson指数**[[7]]）。逆Simpson指数通常对多度大的物种（一般数量很少）多度变化不敏感。Gini-Simpson指数也有$D = (1 - \lambda)/\lambda$版本，就是用随机抽取两个个体属于不同物种的概率与属于同一物种概率的比值来表示[[8]]。

Beta多样性

Beta多样性又称生境间的多样性(between-habitat diversity)，是指沿环境梯度不同生境群落之间物种组成的相异性或物种沿环境梯度的更替速率，用于研究群落之间的种多度关系。

群落的Beta多样性分析包括非限制性排序（如PCoA，NMDS等）、层次聚类、限制性排序等，且均以群落相似或距离为基础计算。

相似性和距离

生态相似性（Ecological resemblance）以计算不同样本群落组成相似程度或距离（相异程度）为基础，是处理多元生态数据的基本方法之一。在群落数据的分析中，常用其反映Beta多样性。

如在物种数据的分析中，对于两个群落，若它们共享相同的物种，并且所有物种的丰度也一致，那么这两个群落就具有最高的相似程度（或最低距离0）。**生态学数据分析中的很多统计方法都以样方之间的相似性或距离为基础**，例如上述提到的Beta多样性分析中的聚类、排序等，即使对于PCA实质上在计算时基于欧几里得（euclidean）距离考虑的。

若两个对象在各属性上越近似，那么它们的相似性就越高。对于群落数据，这些属性一般就是物种组成，或者环境属性等。

通常使用物种组成数据，依据**相似性指数**（similarity indices）判断群落相似性，范围由0（两个群落不共享任何物种）到1（两个群落的物种类型和丰度完全一致）。所有相似性指数均可以转换为距离指数，转化公式为`距离指数 = 1 - 相似性指数`的关系。

- 可以转化为相似性指数的距离指数，例如定量数据的相异百分率（也称为Bray-Curtis距离）等。二者相互转换的公式通常表示为$$D=1-S或S=1-D$$其中S是相似性指数，D为距离指数。

- 无法转化为相似性指数的距离指数，例如欧几里得距离、卡方距离。

距离指数（distance indices）或称距离测度（distance measures），与相似性指数相反，距离数值越大表明群落间差异越大。存在多种距离类型，例如**欧几里得（Euclidean）距离**、**Bray-Curtis距离**、**UniFrac距离**等。对于物种组成数据，距离指数的最小值为0（两个群落的物种类型和丰度完全一致），最大取值取决于距离类型和数据本身。

常见的相似性/距离指数

- Jaccard相似性指数（Jaccard similarity index）[[9]]
  将两个样方共享的物种数量（a）除以两个样方中出现的所有物种的总和（a + b + c，其中b和c是仅在第一个和第二个样方中出现的物种数量）
- Bray-curtis距离（Bray-curtis distance）[[10]]
  或称Bray-curtis相异度（Bray-curtis dissimilarity）、相异百分率（Percentage

difference）。

- 欧几里得距离（Euclidean distance）[[11]]
  是多变量分析中经常使用的一种距离，如在线性排序方法PCA、RDA，以及某些层次聚类算法中。欧几里得距离没有上限，最大值取决于数据。
- Unifrac距离（Unifrac distance[[12]]）
  常用于微生物群落的研究中（例如，16S扩增子测序）。上述距离的计算方法，仅考虑了物种的存在与否及其丰度，没有考虑物种之间的进化关系，距离0表示两个群落的物种组成结构完全一致。在Unifrac距离中，除了关注考虑了物种的存在与否及其丰度外，还将物种之间的进化关系考虑在内，距离0更侧重于表示两个群落的进化分类完全一致。
  - 非加权Unifrac距离（Unweighted unifrac distance）
    只考虑了物种有无的变化，不关注物种丰度，若两个微生物群落间存在的物种种类完全一致，则距离为0。
  - 加权Unifrac距离（Weighted unifrac distance）
    同时考虑物种有无和物种丰度的变化，若两个微生物群落间存在的物种种类及丰度完全一致，则距离为0。

排序方法

排序过程是将样品或物种排列在一定的空间，在一个低维空间中，使相似的样品或物种距离相近，相异的样品或物种距离较远。也就是说排序可以揭示微生物-环境间的生态关系，降低维数，减少坐标轴的数目，使排序轴能够反映一定的生态梯度。降维的过程就像投影，找到最好的角度使投影后的物种或者样品的位置关系尽量保持原始的位置关系。

PCA（Principal Component Analysis）[[13]]

即主成分分析方法，是一种使用最广泛的数据降维算法。PCA的主要思想是将n维特征映射到k维上，这k维是全新的正交特征也被称为主成分，是在原有n维特征的基础上重新构造出来的k维特征。

PCA的工作就是从原始的空间中顺序地找一组相互正交的坐标轴，新的坐标轴的选择与数据本身是密切相关的。

- 第一个新坐标轴选择是原始数据中方差最大的方向，
- 第二个新坐标轴选取是与第一个坐标轴正交的平面中使得方差最大的，
- 第三个轴是与第1,2个轴正交的平面中方差最大的。
  依次类推，可以得到n个这样的坐标轴。通过这种方式获得的新的坐标轴，我们发现，大部分方差都包含在前面k个坐标轴中，后面的坐标轴所含的方差几乎为0。于是，我们可以忽略余下的坐标轴，只保留前面k个含有绝大部分方差的坐标轴。事实上，这相当于只保留包含绝大部分方差的维度特征，而忽略包含方差几乎为0的特征维度，实现对数据特征的降维处理。

PCoA（Principal Coordinate Analysis）[[14]]

即主坐标分析，可呈现研究数据相似性或差异性的可视化坐标，是一种非约束性的数据降维分析方法，可用来研究样本群落组成的相似性或相异性。

PCoA和PCA的不同之处是PCA是基于OTU table也就是基于欧式距离，而PCoA是基于两两样品之间的距离矩阵（可以是任何一种距离），**如果PCoA也使用欧式距离矩阵的话，那么PCA**

和PCoA的分析结果是一样的。

PCoA是基于距离矩阵，它的排序的目的是将N个样品排列在一定的空间，使得样品间的空间差异与原始距离矩阵保持一致，这类排序方法也称作多维标定排序（Multi—dimensional scaling）。

- 有度量多维标定法（Metric Multi—dimensional Scaling）
  排序依赖于相异系数的数值
- 无度量多维标定法（Non—Metric Multi—Dimensional Scaling；**NMDS**）
  排序仅仅决定于相异系数的大小顺序（秩次排序）

无度量多维标定法（Non—Metric Multi—Dimensional Scaling；NMDS）[[15]]
非度量多维尺度法是一种将多维空间的研究对象（样本或变量）简化到低维空间进行定位、分析和归类，同时又保留对象间原始关系的数据分析方法。
适用于无法获得研究对象间精确的相似性或相异性数据，仅能得到他们之间等级关系数据的情形。
其基本特征是将对象间的相似性或相异性数据看成点间距离的单调函数，在保持原始数据次序(秩)关系的基础上，用新的相同次序的数据列替换原始数据进行度量型多维尺度分析。换句话说，当资料不适合直接进行变量型多维尺度分析时，对其进行变量变换，再采用变量型多维尺度分析，对原始资料而言，就称之为非度量型多维尺度分析。
其特点是根据样品中包含的物种信息，以点的形式反映在多维空间上，而对不同样品间的差异程度，则是通过点与点间的距离体现的，最终获得样品的空间定位点图。
NMDS过程是迭代的，并且分几个步骤进行：

- 在多维空间中定义群落的原始位置；
- 指定降低维度的数量（通常为2个维度）；
- 二维构造样本的初始配置；
- 该初始配置下的距离相对于观察到的（测量的）距离进行回归；
- 根据回归确定应力（stress）或二维构造与预测值之间的差异；

如果应力较高，则按减小应力的方向重新定位2维中的点，然后重复进行直到应力低于某个阈值。**经验法则：应力<0.05可很好地表示尺寸减小，<0.1非常好，<0.2 还不错，而应力<0.3 有待提高。**

附加说明：**最终结果可能会因初始配置（通常是随机的）和迭代次数而有所不同，因此建议多次运行NMDS并尽可能减降低应力值。**

首先，NMDS需要距离矩阵或相异矩阵。原始欧几里得距离并不是达到此目的的理想方法：它们对总丰度敏感，因此即使物种的标识不同，也可能将具有相似数量物种的站点(site)视为相似物种。它们对物种的缺失也很敏感，因此可以将缺少相同物种数的站点视为相似物种。

因此，生态学家使用Bray-Curtis相异性计算，该计算具有许多理想属性：

- 它不随单位的变化而变化；

- 它不受添加/删除两个群落中不存在的物种的影响；
- 它不受新增群落的影响；
- 它可以识别总丰度的差异。

参考资料

1. Chao A. Nonparametric estimation of the number of classes in a population[J]. Scandinavian Journal of statistics, 1984: 265-270.
2. Chao A, Yang M C K. Stopping rules and estimation for recapture debugging with unequal failure rates[J]. Biometrika, 1993, 80(1): 193-201.
3. Shannon C E. A mathematical theory of communication[J]. The Bell system technical journal, 1948, 27(3): 379-423.
4. Pielou E C. The measurement of diversity in different types of biological collections[J]. Journal of theoretical biology, 1966, 13: 131-144.
5. Simpson E H. Measurement of diversity[J]. nature, 1949, 163(4148): 688-688.
6. Greenberg J H. The measurement of linguistic diversity[J]. Language, 1956, 32(1): 109-115.
7. Hill M O. Diversity and evenness: a unifying notation and its consequences[J]. Ecology, 1973, 54(2): 427-432.
8. Margalef R, Gutierrez E. How to introduce connectance in the frame of an expression for diversity[J]. The American Naturalist, 1983, 121(5): 601-607.
9. Jaccard P. The distribution of the flora in the alpine zone[J]. New phytologist, 1912, 11(2): 37-50.
10. Bray J R, Curtis J T. An ordination of the upland forest communities of southern Wisconsin[J]. Ecological monographs, 1957, 27(4): 326-349.
11. Ratcliffe J G, Axler S, Ribet K A. Foundations of hyperbolic manifolds[M]. New York: Springer, 1994.
12. Lozupone C, Knight R. UniFrac: a new phylogenetic method for comparing microbial communities[J]. Applied and environmental microbiology, 2005, 71(12): 8228-8235.
13. Pearson K. LIII. On lines and planes of closest fit to systems of points in space[J]. The London, Edinburgh, and Dublin philosophical magazine and journal of science, 1901, 2(11): 559-572.
14. Gower J C. Some distance properties of latent root and vector methods used in multivariate analysis[J]. Biometrika, 1966, 53(3-4): 325-338.
15. Clarke K R. Non-parametric multivariate analyses of changes in community structure[J]. Australian journal of ecology, 1993, 18(1): 117-143.