[Ecological resemblance] [https://www.davidzeleny.net/anadat-r/doku.php/en:similarity]

生态环境的resemblance包含样本间的相似性和距离, 是如何处理多重变量生态数据的基本工具. 两个共有相同丰都的物种的样本具有最高的相似性(最近的距离), 同时随着他们物种组成差异的增加, 其相似性降低(距离增加). 所有聚类或ordination方法都是处理样本间的相似性或距离. 即使是PCA, CA, 也是分别根据Euclidean和chi-square距离.

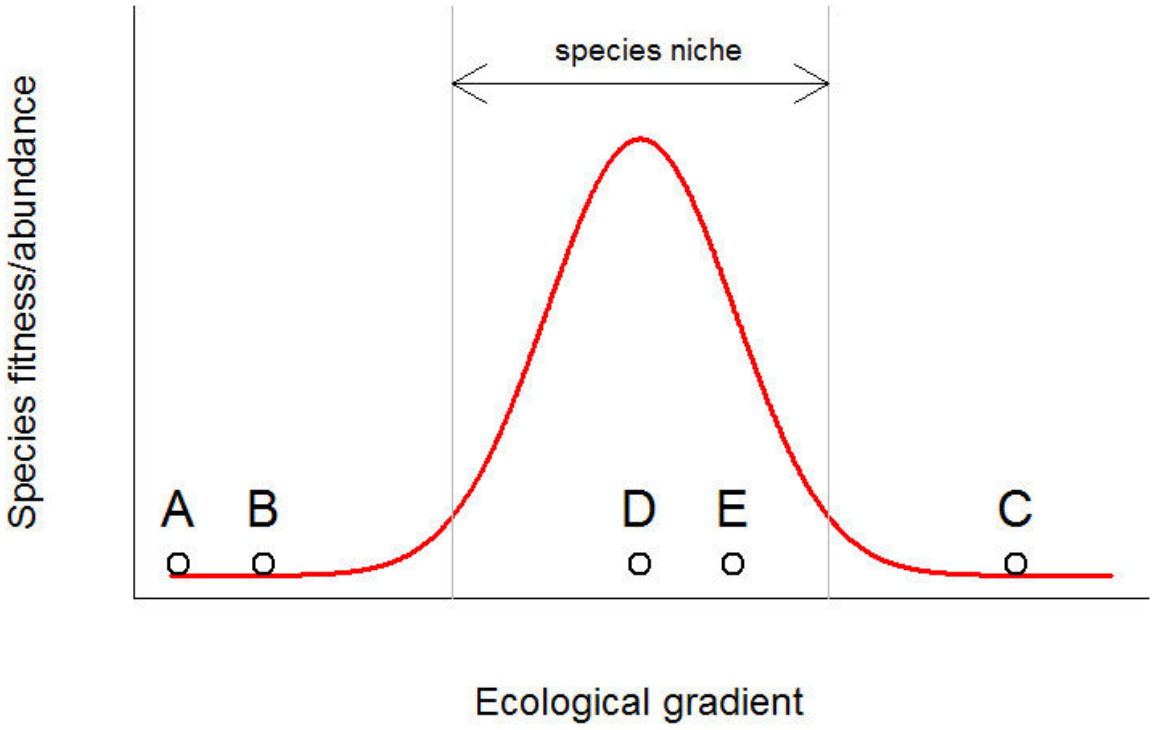
Similarity, dissimilarity and distance

针对物种组成数据, 使用相似性indices计算相似性, 范围从0到1. Ordination方法常基于距离(distances), 因为需要在多重维度空间定位样本; clustering方法常需要处理相似性或距离. 距离有两种类型: 差异性, 根据类似性的 相似性indices转换而来, 或是制定的距离测量, 例如Euclidean.同时, 所有相似性indices都可以转换为距离(distances), 但并非所有距离都可以转换为相似性(例如, Euclidean distance).

目前存在多种计算相似性或距离的方法. 首要选择是根据分析目的, 是R-mode分析还是Q-mode分析(R-mode针对物种间的差异, Q-mode针对样本间的差异). 由于一些方法不同于这两种modes(例如, Pearson's r correlation coefficient, 针对物种间的关联(R-mode), 但是不能作用于样本间的关系(Q-mode)), 相反, Sorensen index(Dice index R-mode)分析可针对Q/R-mode). 此外, 如果着重于样本间差异(Q-mode), 生态环境中最相关的测量为忽略double zeros条件下的非均匀indices(asymmetric indices). 然而, 分析方法也根据数据类型, 定性的(例如, binary, presence-absence)或定量的(物种丰度). 针对distance indices, 一个重要的标准是是否是metric(they can be displayed in Euclidean space) or not, 因为这将影响一些ordination或clustering方法的index选择.

Double-zero problem

'Double zero'为计算相似性/距离时, 明确物种在比较的community的样本中都缺失的情形. 在两个样本中同时缺失物种意味着: (1) 样本位于物种环境niche之外, 但是不能说这两个样本都在环境层级的同侧, 或相对侧; (2) 样本位于物种生态环境niche的内部, 但是给定样本中该物种都没发生, 因为其并没get there(dispersal limitation), 或者确实存在, 但是被忽视, 取样过程没有取到. 以上任意情况, double zero所代表的缺失信息, 不能给比较样本的生态提供判断信息.



相似性indices在处理double-zero问题时采用不同方式: symmetrical indices使用0-0来表示, 和存在表示一样; asymmetrical indices则会忽略double zero, 在评估样本相似性时着重于同时存在的情况, 针对物种组成数据, 这类indices更有意义.

Similarity indices

Symmetrical indices, 会考虑double zero为相关, 由于其针对生态环境数据分析不是很有用, 这里不会进一步分析. asymmetric similarity indices(ignoring double zero), 根据它所使用的数据类型分为两类: 定性和定量.

Similarity indices		How they deal with *double zero* problem?	
		symmetrical (treat double zeros as important information)	**asymmetrial (ignore double zeros)**
Which type of data indices use?	**qualitative (binary = presence absence data)**	not suitable for ecological data	Jaccard similarity, Sørensen similarity, Simpson similarity
	quantitative (species abundances)	not suitable for ecological data	Percentage similarity[1)]

Table 1: Similarity indices classified according to their properties.

number of species which are		**in sample 1**	
		present	absent
in sample 2	present	a	b
	absent	c	d

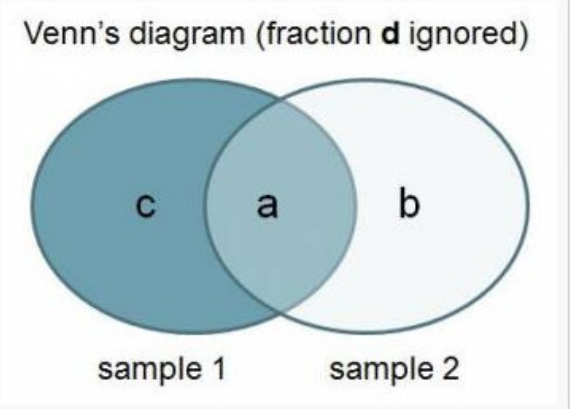


Table 2: The meaning of fraction a, b, c and d used in qualitative indices calculating similarity among two samples. In assymetric indices, the fraction d (double zero) is ignored.

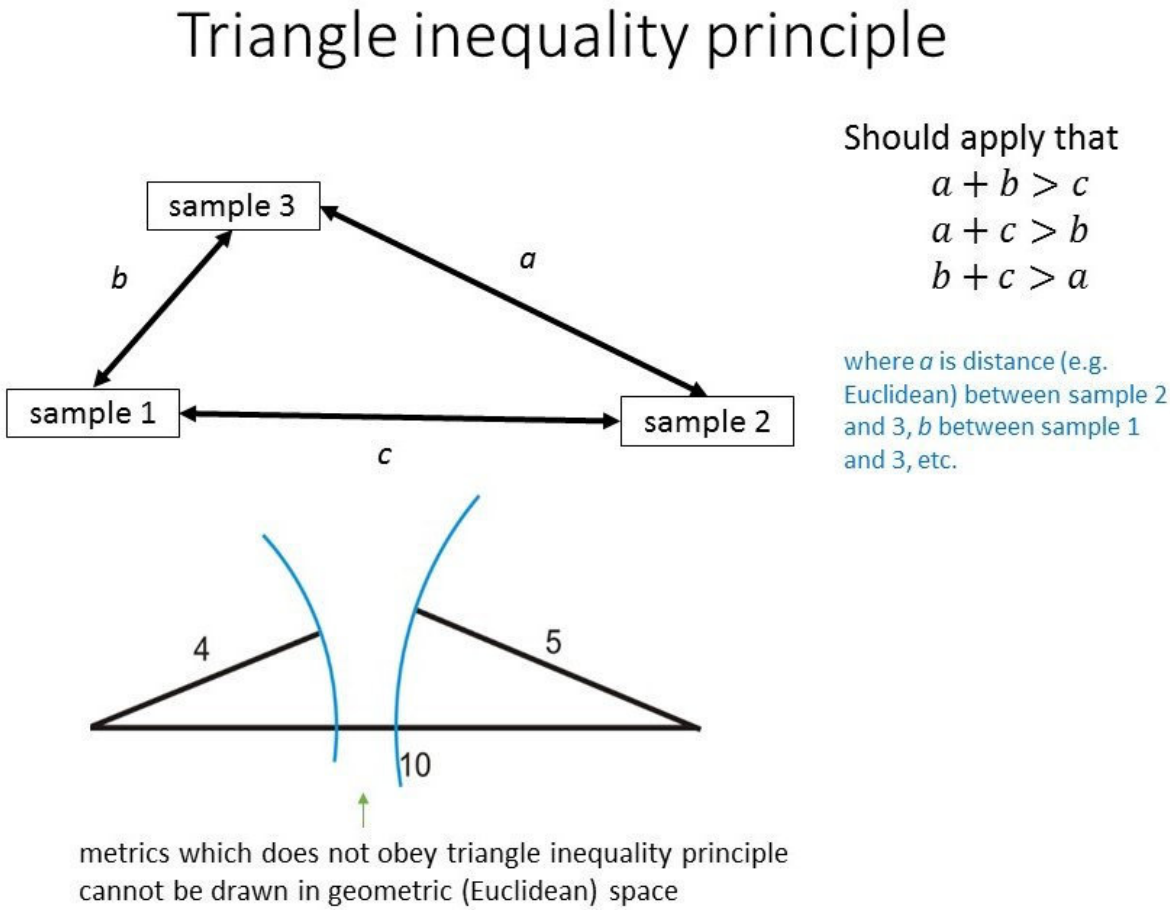
略!

Distance indices

存在两种计算距离或差异性indices的方法:

1. 根据相似性indices计算: D = 1 - S, 包含, 针对定性数据的Jaccard, Sorensen, Simpson dissimilarity; 针对定量数据的Bray-Curtis
2. 和相似性indices没有任何相似的距离计算: Euclidean, chord, Helligner, chi-square distance index

其中一个重要标准是, distance index是否是metric. 'metric'指得是indices能够展示在orthogonal Euclidean space, 因为它们满足所谓的'triangle inequality principle'



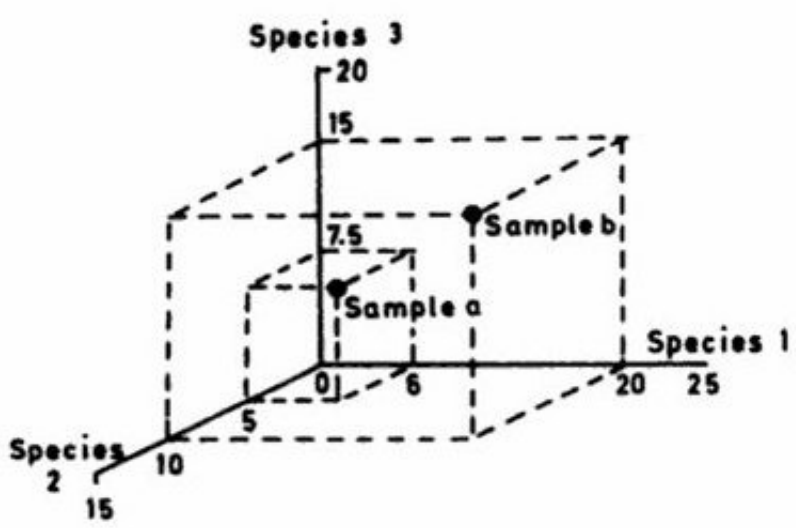
一些根据相似性计算得到的差异性为metric(Jaccard dissimilarity), 一些不是(Sorensen dissimilarity and it's a quantitative version called Bray-Curtis dissimilarity). 而非metric的indices会给依靠Euclidean space(PCoA, db-RDA)的ordination方法带来问题. 还有多个需要使用Euclidean space定位样本的clustering算法(Ward algorithm, K-means).

略!

[Ordination analysis] [https://www.davidzeleny.net/anadat-r/doku.php/en:ordination]

Theory

Ordination是一个多变量分析, 用于在多重变量数据中搜索连续的模式, 数据常是关于环境样本中的物种组成(sample x species matrix). 可以想象多重变量数据为位于超空间的多重维度的样本, 每个维度都被一种物种的丰度定义(例如, 两个样本和三个物种所组成的环境)



Ordination的主要假设是分析的数据是过剩的, 数据包含过多的变量, 多于必要的用于描述潜在信息的变量数目. 因此, 可在不丢失太多信息的情况下减少变量的数目(和维度). 例如, 在物种组成数据中, 常常一些物种的生态环境相似的(例如, 那些倾向于生长在潮湿环境而不是干燥环境的物种), 意味在描述相同的特征关系情况下,该数据集包含了多个冗余变量(物种). 或者, 根据物种的存在与否, 我们常可以用来预测多个其他物种的有无(例如, 假如样本包含适宜潮湿环境的物种, 我们可能预测偏好干燥环境的物种就不会存在, 同时, 其他喜湿的物种就可能存在).

由于多维度空间很难展示, 或者仅仅想象, 这就有必要将其缩减到少数几个主要的维度. 这也就意味着, 假如个别变量完全不依靠其他变量(例如, 有些物种拥有完全不同的喜好), 那么, ordination就不可能发现一些有理由的多维度空间缩减, 因为每个维度(物种)都有意义.

哪些ordination方式可以通过两种可变方式实施:

1. 搜索物种组成的等级关系(通过ordination坐标展示), 同时意图通过环境变量来解释这些等级关系.
2. 或者, 在缩减后的ordination空间内搜素搜样本的分布, 最大化地反映物种组成层面上的样本间的差异性

Types of ordination methods

	(a) Raw-data-based (classical approach)		(b) Transformation--based	(c) Distance-based
	Linear	**Unimodal**		
(1) Unconstrained (indirect) Shiny app (how to use, blog)	PCA Principal Component Analysis	CA & DCA Correspondence Analysis & Detrended Correspondence Analysis	tb-PCA Transformation-based Principal Component Analysis	PCoA, NMDS Principal Correspondence Analysis, Non-metric Multidimensional Scaling
(2) Constrained (direct, canonical) Shiny app	RDA Redundancy Analysis	CCA Canonical Correspondence Analysis	tb-RDA Transformation-based Redundancy Analysis	db-RDA Distance-based Redundancy Analysis

Table 1: Summary of ordination methods. Based on Lepš & Šmilauer (2003), extended for other methods from Legendre & Legendre (2012).

根据其算法在物种组成数据中是否包含环境变量分为: unconstrained do not, constrained do; 和用于分析的物种数据组成类型: raw data(物种组成的样本物种矩阵), pretransformed raw data(例如, 经过Hellinger transformation), 或距离矩阵(样本间一一对应的距离关系矩阵)

1. ordination算法是否也包含了环境变量

（1) Unconstrained ordination(indirect gradient analysis)

Ordination坐标并没有被环境变量所constrain. 该方法在于揭露物种组成数据主要的层级关系(改变方向), 并且返回unconstrained坐标, 该坐标对应为数据集中最大变化的方向. 根据个人意愿, 这些层级关系也可被事后通过环境变量解释. 环境变量并不进入到ordination算法中. Unconstrained ordination是主要的探索分析方法, 用于探索多重变量数据中的模型; 生成假设, 但是不检验假设.

(2) Constrained ordination(direct gradient analysis, canonical ordinations)

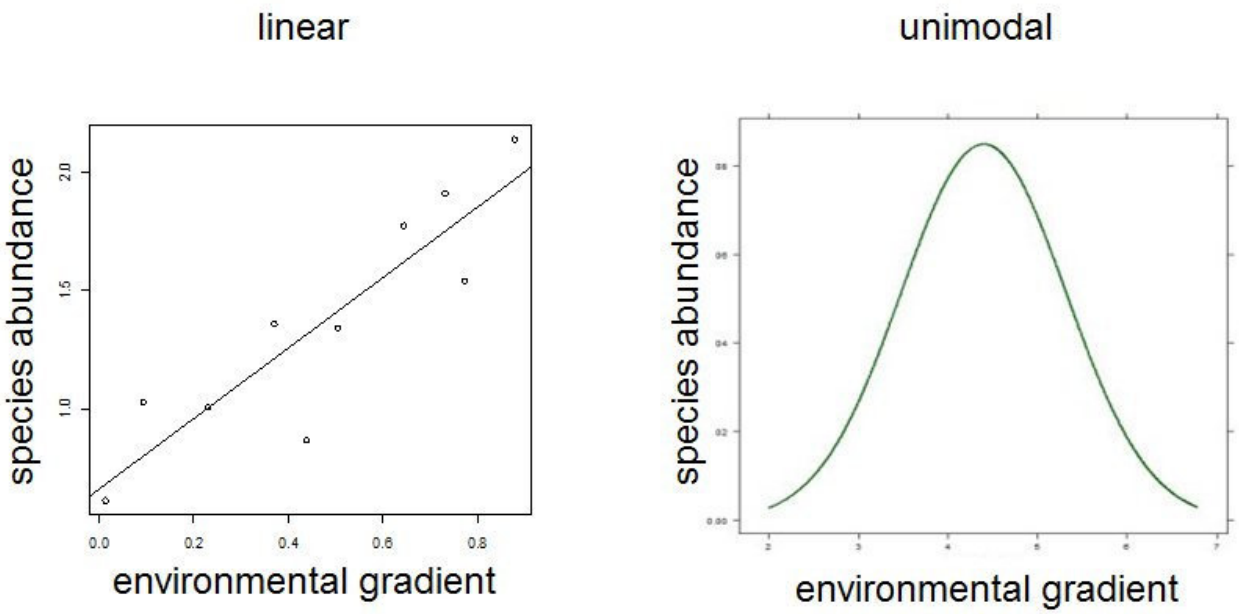
Ordination坐标被环境因素所constrain. 它直接关联物种组成和环境变量, 并且在直接关联环境的物种组成中提取变异. 环境变量直接进入算法, 并且constrained ordination坐标对应数据(可被环境变量解释的)中变异的方向. 该方法一般用于确认性分析, 例如, 能够检验关于环境因子在物种组成之间关系的假设(不同于unconstrained ordination, 仅是探索). 它将物种组成数据中的所有变化分解为能被环境变量所解释的比例部分(related to constrained ordination axes)和不能被环境变量所解释的比例部分(related to unconstrained ordination axes). 当涉及到可解释的变量时, 提供多个有趣机会: forward selection(通过排除和物种组成无关的环境变量来选择重要变量), Monte Carlo permutation test(能够环境因素解释的变量显著性的检测), variance partitioning(通过不同环境变量分组将解释的变化划分).

What type of species composition data is used for analysis

(a) Raw-data-based methods(classical approach)

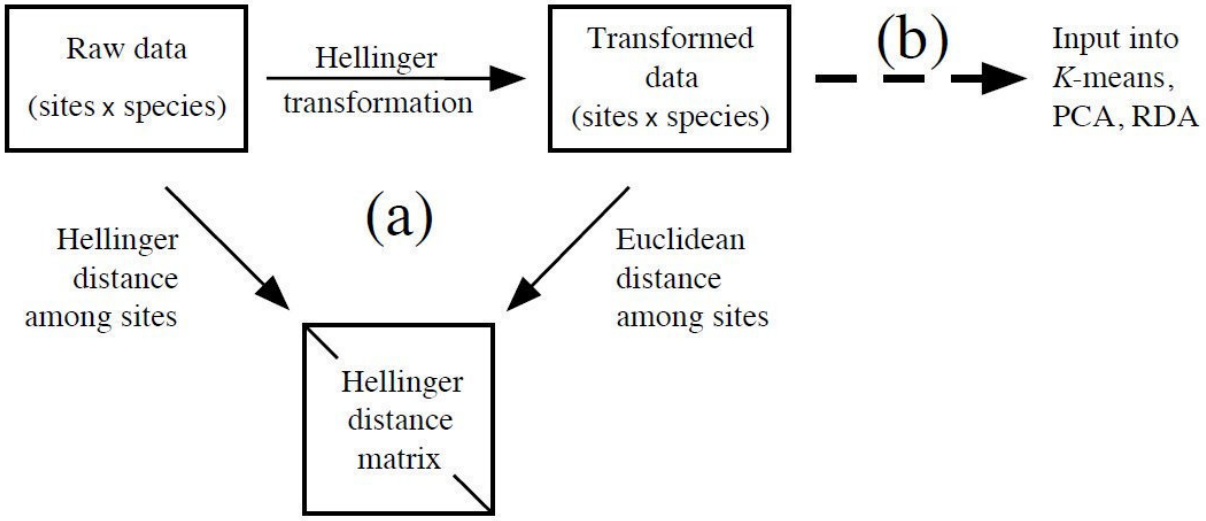
根据包样本物种丰都或存在与否的矩阵的分析, 该类方法中, 为传统所认可的有两类, 其差异表现为物种是否相应环境的层级

- linear(线性), 假设物种线性相应环境层级变化, 即使针对同种生态数据, 其环境层级相当小时, 也可能真是存在
- unimodal(单峰曲线), 物种相应时沿着层级单峰变换, 在某一明确层级位置拥有最大值; 该模型更接近于实际生态数据, 跟适合于杂合型数据集(由较强或更长的生态层级构成, with high species turnover and many zeroes in the species matrix), 倾向更长的环境层级.



(b) Transformation-based methods(tb-PCA and tb-RDA)

该分类包含线性的raw-data-based ordination方法(PCA, RDA), 适用于通过Hellinger转换得到的sampeXspecies数据. The Euclidean distance, 当用于Hellinger-transformed 物种组成数据时, 生成为Helliger distance, 该数据更适合于生态数据, 因为(相对于Euclidean distance)它是非对均匀的, 基于这点更适合分析杂合型数据. 此外, 针对Hellinger transformation, 其他合适的转化为chord transformation, 和其他物种profile transformation, Chi-square distance和chi-square metric transformtion.



a. Hellinger distance可直接通过原始物种组成数据计算而来, 或首先通过Hellinger transformation(standardization)转换物种组成数据, 然后使用Euclidean distance计算转换后数据

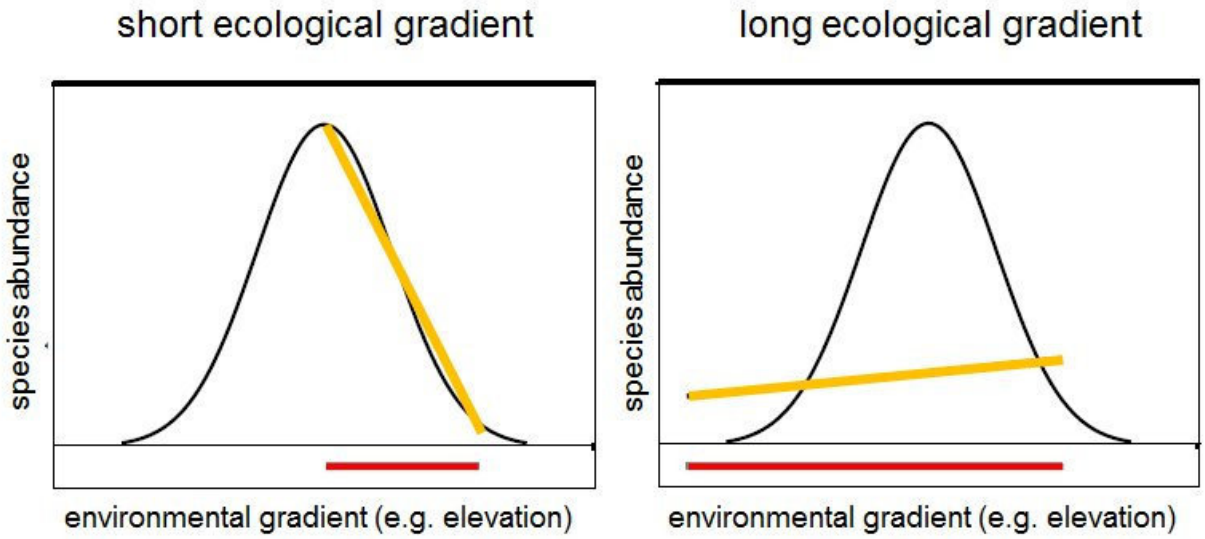
b. Helinger 转换后数据可通过Euclidean distance用于ordination methods, 例如, PCA, RDA, K-means clustering), 这些方法都使用Helligner distances(Hellinger transformed data + Euclidean distance = Hellinger distance)

(c) Distance-based methods

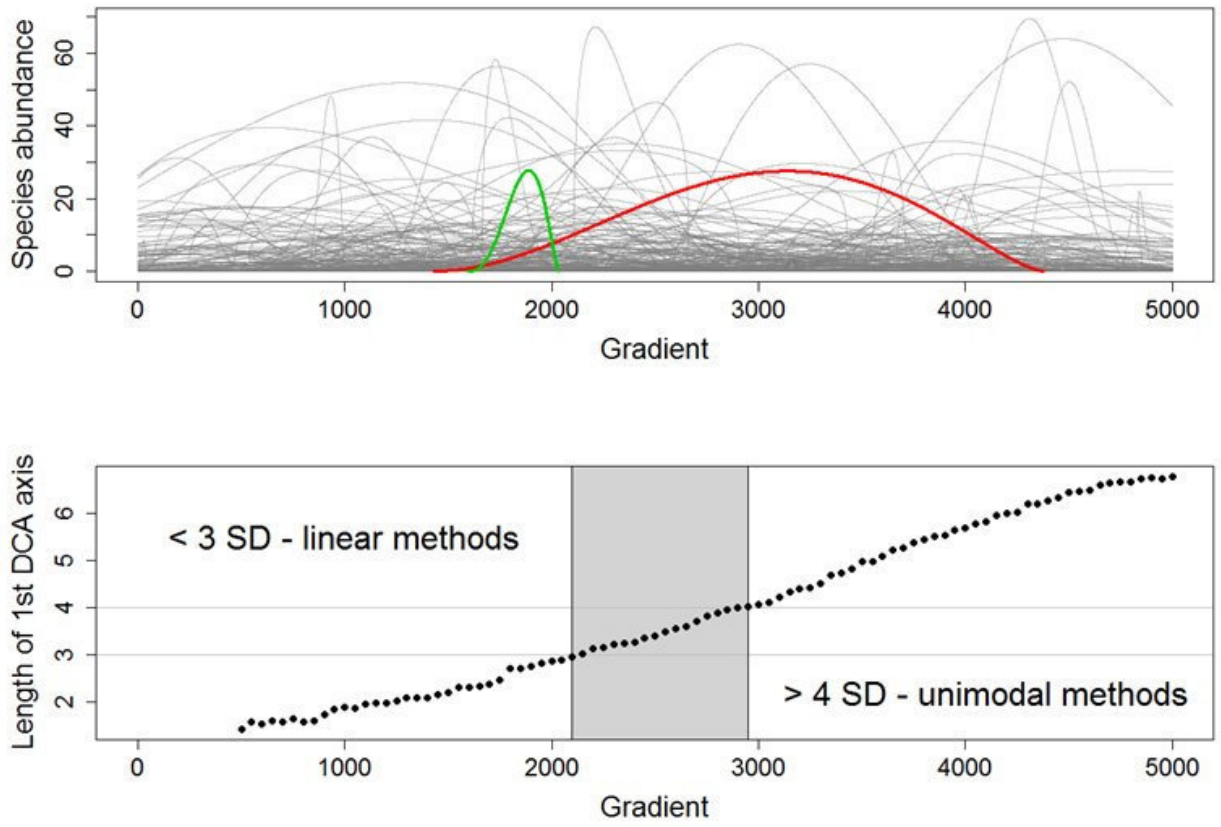
根据distance coefficients计算而来的样本间的距离矩阵, 并且投射这些距离到二维或多维ordination 图表中. Distance-based RDA(db-RDA) 为PCoA的合并, 将选择的距离测量值用至原始数据, 同时RDA使用来自PCoA的eigenvectors.

Linear or unimodal ordination method?

当取样是一个环境层级中的一个短的部分时, 我们可能将假设物种对环境的相应时线性的; 当取样为一个长的环境层级时, 模拟的到的物种响应就不再是线性的了.



为判断到底使用线性还是单峰型ordination method, 可以使用公推的方法: 首先, 计算数据的DCA(detrened by segments), 然后查看第一个DCA坐标的长度(which is scaled in units of standard deviation, S. D.). 当第一个DCA坐标长度大于4, 表明为杂合型数据集, 应使用unimodal methods, 当第一个DCA坐标小于3, 标明为同类型数据集, 采用linear mthods. 若该值位于3和4之间, 那么使用哪种都可以. 需要注意的是, 线性模型不能用于杂合型数据, 而单峰模型可用于同类型数据集. 此外, 如果数据为杂合型, 任然希望使用线性模型方法(PCA, RDA), 那么应用到经过Hellinger transformed 物种组成数据上来计算ordination(based on Hellinger distances)

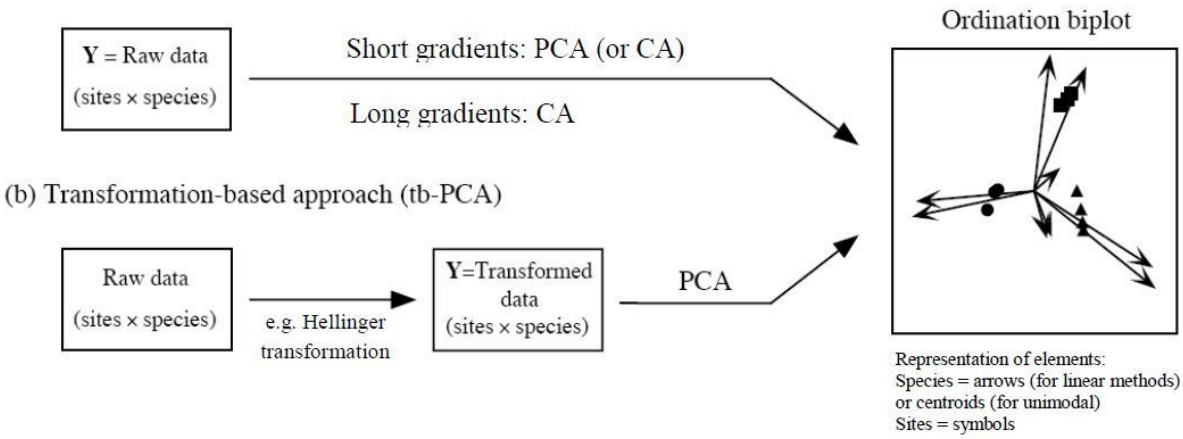


图示如何选择何种ordination方式(线性, PCA, RDA; 单峰型, CA, DCA, CCA). 上图展示了模拟的被单个环境层级所代表的环境结构, 同时包含物种数目响应曲线. 下图展示了在模拟环境中取样的层级长度和第一个DCA坐标长度的关系. 根据DCA, 该数据集倾向于同种类型(SD < 3). 当随着层级的增加, 进入杂合型数据, 此时线性模型就不再合适了.

Summary: Three alternative approaches for ordination

(1) Unconstrained ordination analysis

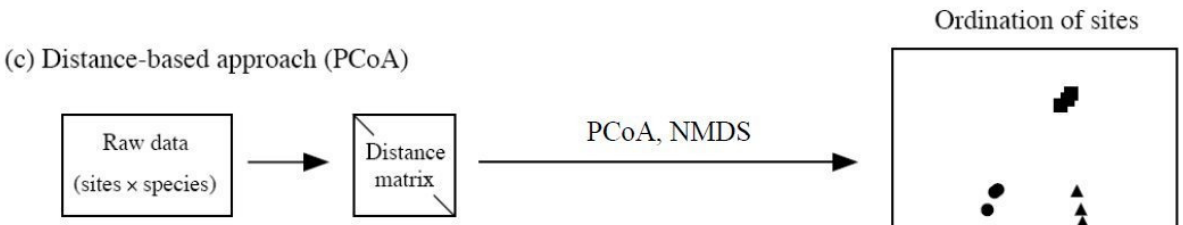
(a) Classical approach



(b) Transformation-based approach (tb-PCA)

Representation of elements:
Species = arrows (for linear methods) or centroids (for unimodal)
Sites = symbols

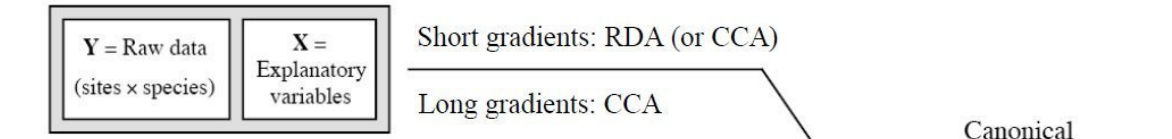
(c) Distance-based approach (PCoA)



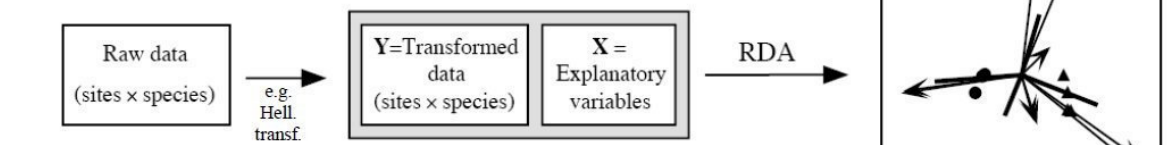
Representation of elements:
Sites = symbols
(Species could be added e.g. as weighted averages)

(2) Constrained ordination analysis

(a) Classical approach: RDA preserves the Euclidean distance, CCA preserves the chi-square distance

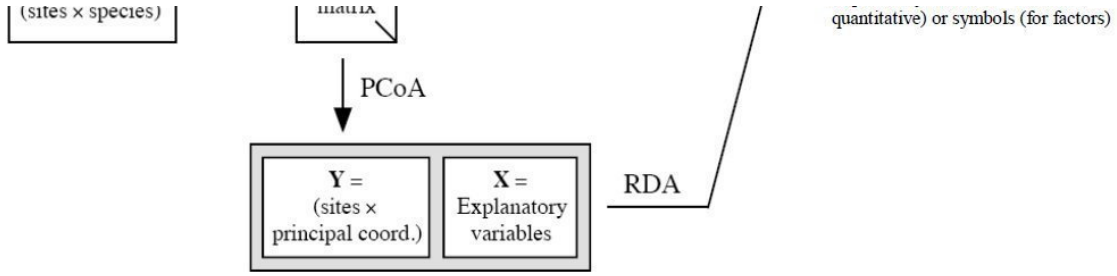


(b) Transformation-based RDA (tb-RDA) approach: preserves a distance obtained by data transformation



(c) Distance-based RDA (db-RDA) approach: preserves a pre-computed distance

Representation of elements:
Species = arrows (for linear methods) or centroids (for unimodal)
Sites = symbols
Explanatory variables = lines (for



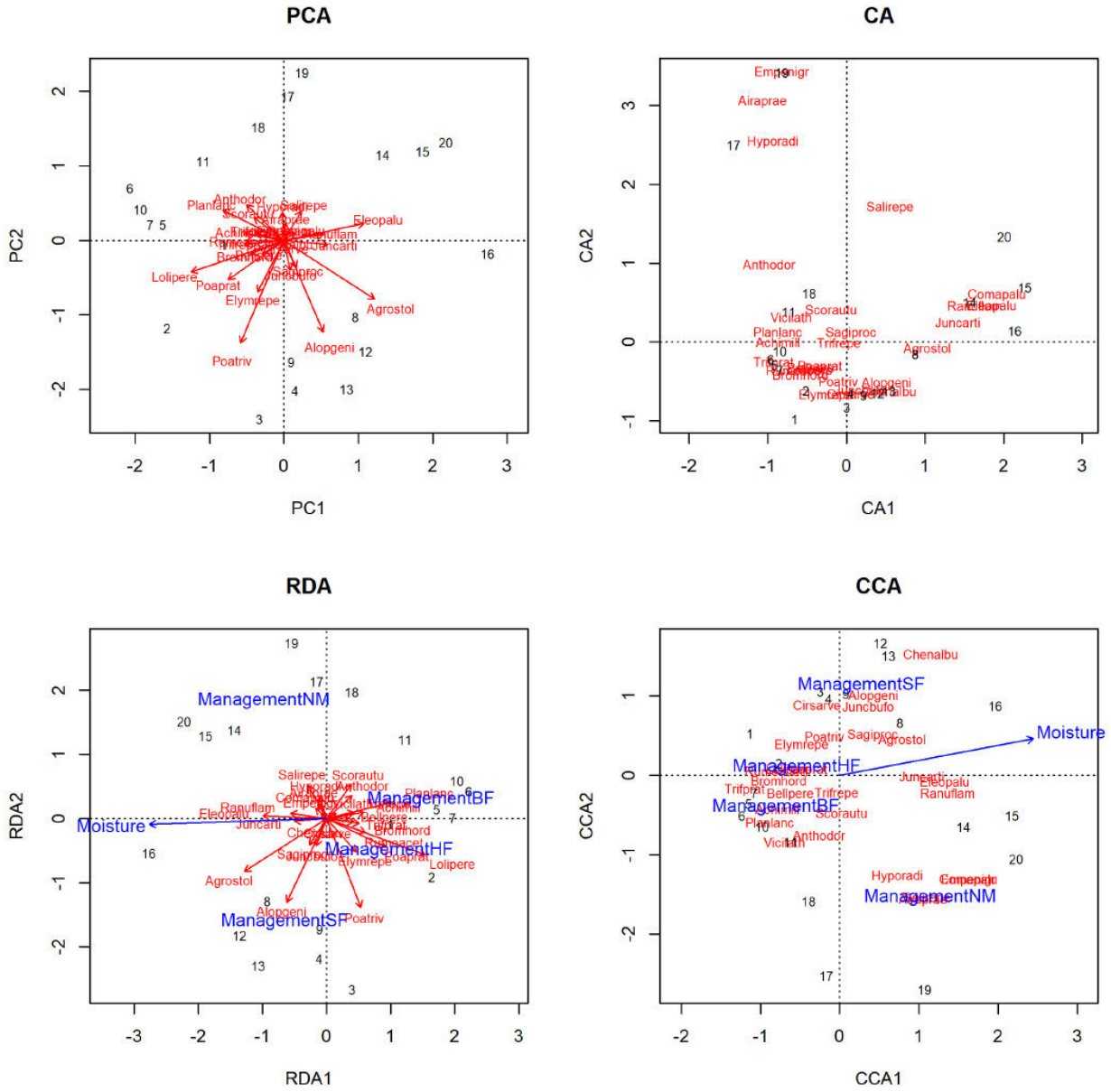
(slightly) modified from Legendre & Legendre (2012)

[Ordination diagrams] [https://www.davidzeleny.net/anadat-r/doku.php/en:ordiagrams]

Ordination diagrams常为二维图表来表示ordination分析结果. 不同的ordination方法可能在其展示的结果内容和方法上有区别(PCA, CA, RDA, CCA ordination diagrams).

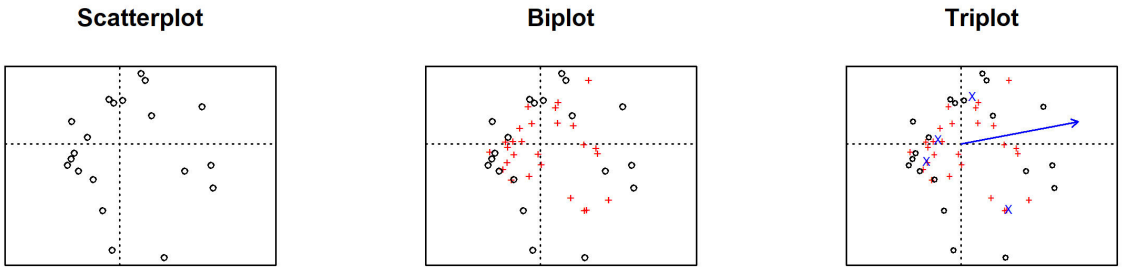
展示ordination diagrams的标准形式:

- **样本(sites) - points; 点之间的距离近似于样本间的组成差异(通过相对距离表示, 例如PCA/RDA中的Euclidean, CA/DCA/CCAZ中的chi-square)**
- **物种(species) - 箭头针对线性, centroids针对单峰形式以及根据距离的方法: 箭头指明物种丰度增加的方向; 在给定物种丰度在所有方向都降低的条件下, centroids表明ordination diagram中物种最佳(species optima)的位置.**
- **Ordination axes - 水平坐标往往比垂直坐标更有意义(高一个数量级, 例如, 水平=1s axis相当于垂直=2nd); axes代表了物种组成的主要层级.**
- **Environmental variables - 箭头针对定量, centroids用于定性; 箭头制定环境值增加的方向; centroids表明定性变量位于其所在的那部分diagram.**



Type of ordination plots

根据展示的变量类型和数目, 会采用多个不同的ordination diagrams: 'scatterplot', 一类数据, 仅有物种值或位置值; 'biplot', 两类数据, 物种值和环境变量; 'triplot', 三类数据, 物种, 样本值和环境变量.



(o = sample scores)

(o = sample, + = species scores)

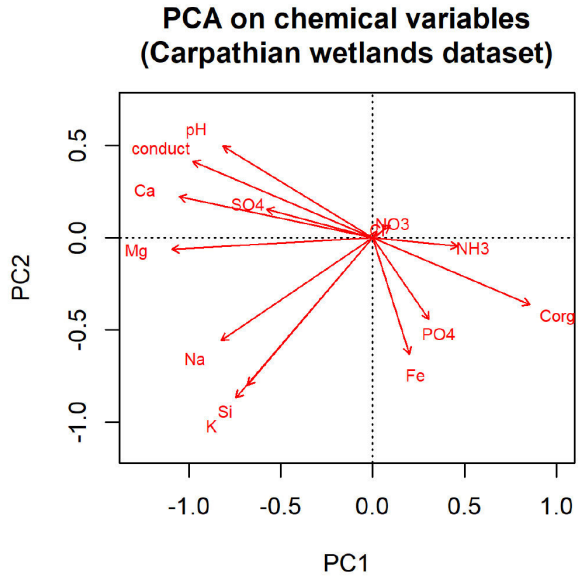
(o = sample, + = species scores, vector & centroid = env. vars)

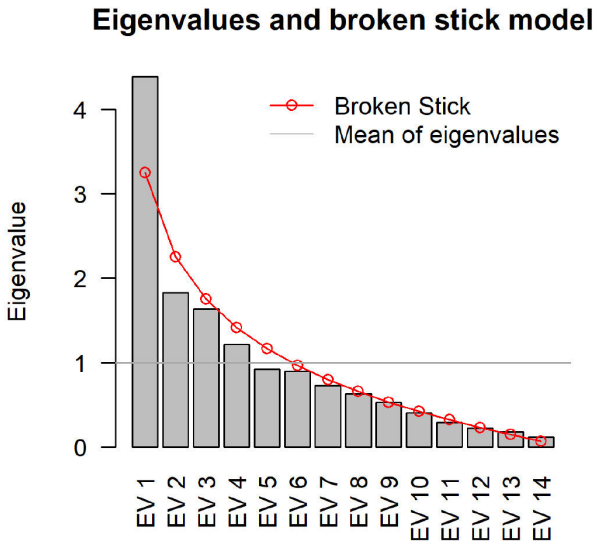
How many ordination axes to consider?

大部分情况下是2个, 有时3个, 罕有情况下4个. 如果熟悉自己数据(例如, 取样深度, 环境的变化性), 经会知道需要考虑多少个生态维度.

在使用unconstrained ordinations时, 可能绘制eigenvalues的scree plot(barplot). Eigenvalues代表了数据中被给定ordination axis所捕获的变异, 并且ordination axes是根据eigenvalues的顺序排列的(从高到低). 从scree plot, 可看到是否存在主要的非联系性(major discontinuity), 和使用这些axes能够获得相当多的变异数目. 根据以下来判断保留axes数目:

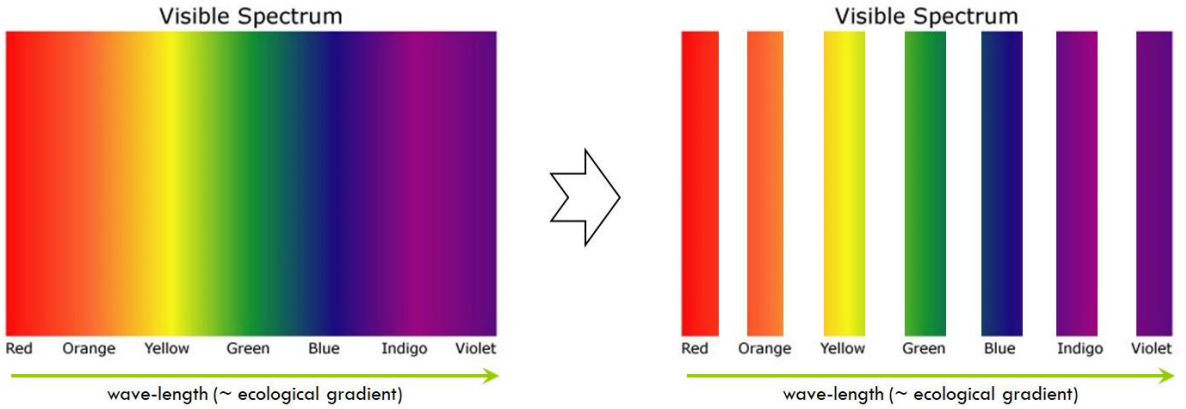
- Kaiser-Guttman criterion - 就是那所有eigenvalues的平均值, 解释大于该均值的eigenvalues
- broken stick model - randomly divides the stick of unit length into the same number of pieces as there are PCA axes and then sorts these pieces from the longest to the shortest. Repeats this procedure many times and averages the results of all permutations (analytical solution to this problem is also known). *Broken stick model* represents a null model and generates values of eigenvalues, which would occur at random. One may want to interpret only those PCA axes with eigenvalues larger than values generated by *broken stick model*.





[Numerical classification] [https://www.davidzeleny.net/anadat-r/doku.php/en:classification]

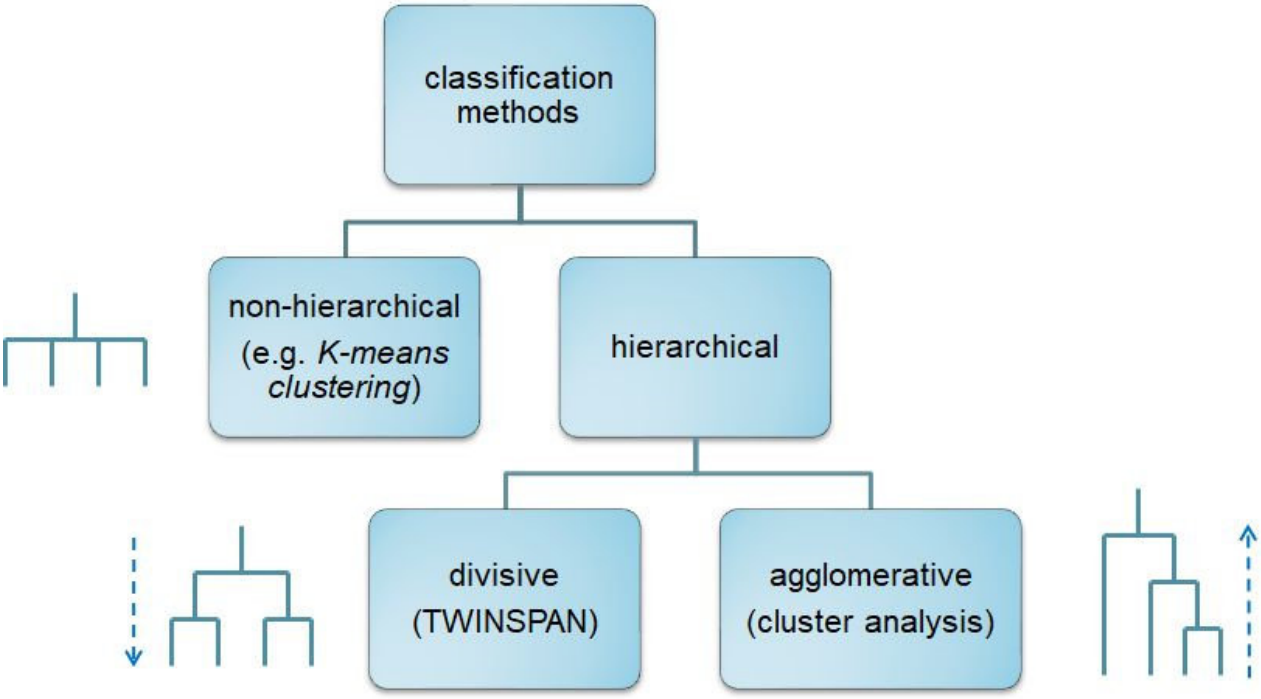
基于数字的分类方法目的在于在community数据中发现不连续性(也可能是更连续或不连续的点)并且命名, 例如为了方便交流或清晰地查看组成模型. 该分析通过将类似对象(samples, species)聚集到一个群(group)内, 使得该群内对象同源且容易和其他群区分开. 在一些特殊的community数据情况下(sample X speceis matrix), 该分类可以根据样本(产生的群包含类似物种组成的样本, 由于communities来源于类似的栖息地), 或根据物种(每个群包含的物种具有类似的生态行为).



即使真正的连续的对象, 也可能被很好的区分来, 并且给定名称.

Types of classification methods

数字分类方式的简单分类如下图, 该方法根据样本的结果群(groups)是否拥有等级关系分为等级型和非等级型(一些更类似另一些时, 可通过系统发育树来表示). 针对等级型方法, 存在两个选择: divisive algorithms, 将整个数据集切割成小的数据集, 再切割为更小的数据集...(in to down direction); 和 agglomerative algorithms, 从单个样本水平出发, 知道将其合并到一个大的群位置(bottom-up direction)



Unsupervised vs supervised classification

Unsupervised方法在物种组成中搜索主要的层级关系, 主要是不连续性或同源性样本, 且返回结果仅仅依赖与所选择的方式和数据集的内部结构. 相反, Supervised分类方法使用额外的标准来分来数据集, 可提供关于如何处理分类的信息, 同时这会将其应用到现存数据集中. 针对非监督性分类, 可以根据主观选择来修改结果(例如, 聚类算法, distance metric, 形成groups的阈值), 但是其主要的结果依靠数据集的内部结构, 同时即使稍微改变数据集就可能改变样本在组(groups)间的分配(例如, 增加样本量). 相反, 监督型分类, 根据外部提供的标准完成分类, 不会因为数据结构的改变而改变样本的分配.

Examples of unsupervised methods are TWINSPAN or cluster analysis, supervised methods (not discussed in detail on this website) include artificial neural networks (ANN), classification and regression trees (CART), random forests, COCKTAIL (logical formulas, designed for veg. data). Some methods, like K-means clustering, can run in either unsupervised or supervised mode – in the unsupervised mode the method first searches for the centroids of the predefined number of groups and assigns individual samples to these groups, while in the supervised mode the centroids are defined by user and the method just assigns the samples into these predefined groups.

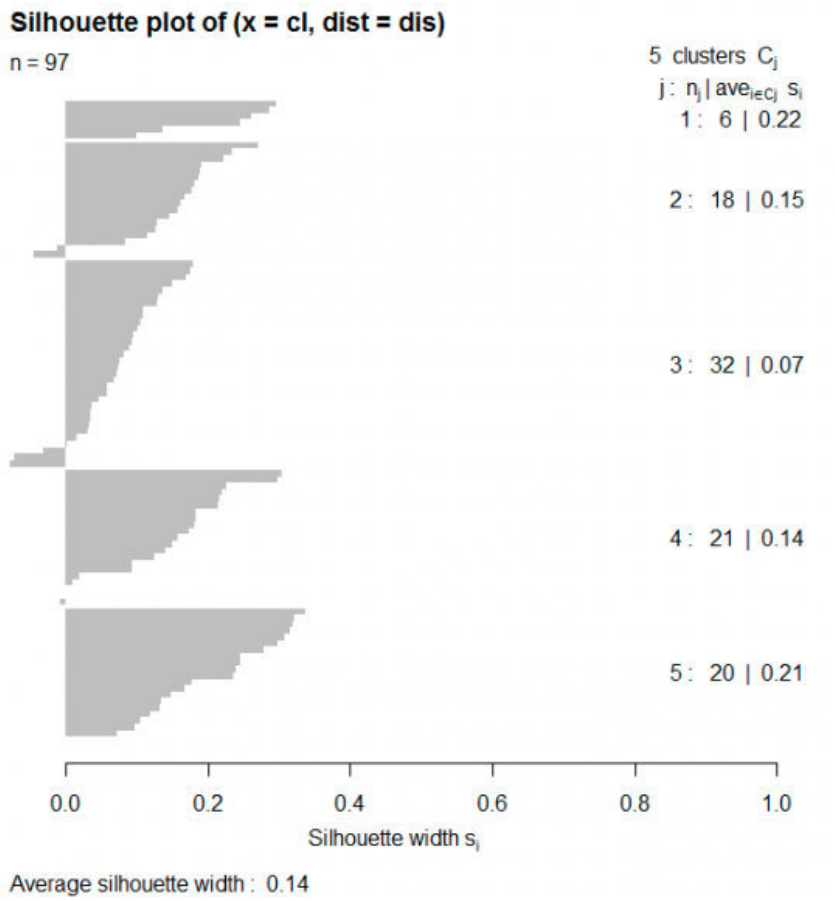
非监督型分类方法, 例如TWINSPAN或聚类分析, 监督型分类方法包括ANN(aritifical neural networkds), CART(classification and regression trees), random forests, COCKTAIL(logical formulas, designed for veg. data). 一些方法, 类似K-means clustering, 可在两种模式下运行, 非监督模式时根据定义的groups数目先搜索centroids, 然后完成分配; 监督模式下, 用户定义centroids, 然后分析完成分配样本到制定的groups即可.

[Evaluation of classification results] [https://www.davidzeleny.net/anadat-r/doku.php/en:class-eval]

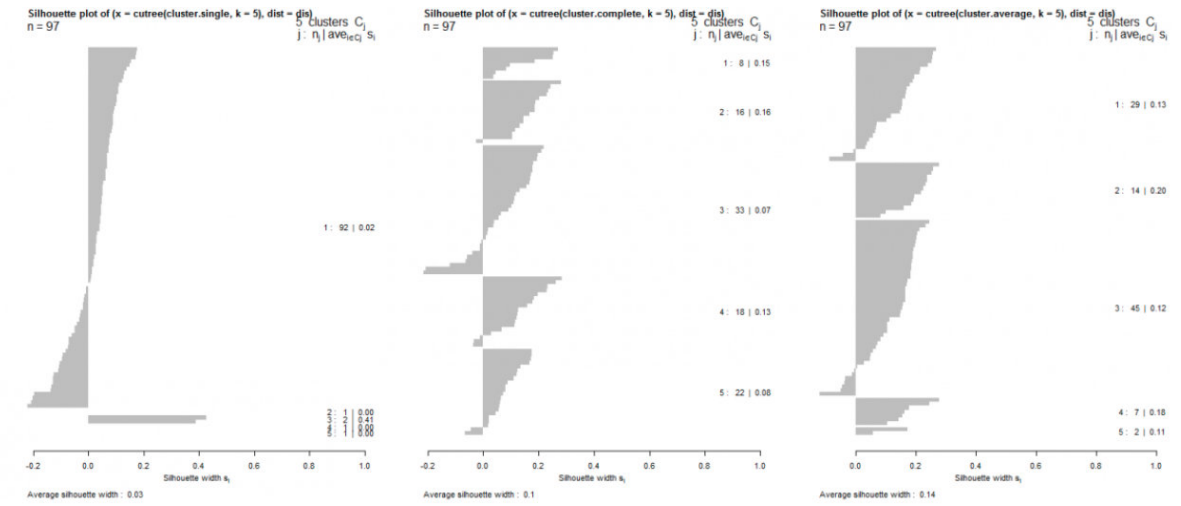
silhouette(library cluster)

若样本含有相近的group memebership. 样本具有较高的的s值表示样本很好的聚集, s值在0附近表示样本位于两个cluster之间, 负值则表示样本被错误的分类了.

```
## Example of silhouette function

## Following code is not necessary, if you already used examples above...
# library (cluster)
# dis <- vegdist (sqrt (vltava.spe), method = 'bray') # percentage cover data
are transformed by square root
# cluster.flexible <- agnes (x = dis, method = 'flexible', par.method = 0.625)
# cluster.flexible.hclust <- as.hclust (cluster.flexible)

cl <- cutree (cluster.flexible.hclust, k = 5)
si <- silhouette (cl, dis)
plot (si)

# Group 3 has the highest number of missclassified samples, on the other hand
groups 1, 2 and 5 are well defined.
```




```
# Comparison of silhouettes for single linkage, complete average linkage
method.
# dis <- vegdist (sqrt (vltava.spe), method = 'bray') # percentage cover data
are transformed by square root
# cluster.single <- hclust (d = dis, method = 'single')
# cluster.complete <- hclust (dis, 'complete')
# cluster.average <- hclust (dis, 'average')

par (mfrow = c(1,3))
plot (silhouette (cutree (cluster.single, k = 5), dis))
plot (silhouette (cutree (cluster.complete, k = 5), dis))
plot (silhouette (cutree (cluster.average, k = 5), dis))
```

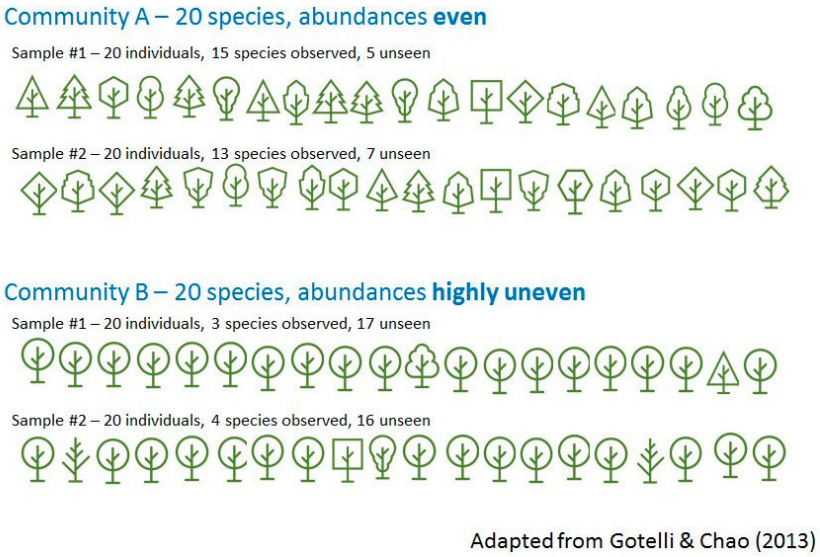


[Diversity analysis][https://www.davidzeleny.net/anadat-r/doku.php/en:diversity_analysis]

Theory

一般而言, 多样性是一个系统内的不同状态的测量量化数目. 针对生态环境时, 这些状态常指的是物种, 但是也可能是genera, family, OTU's或 功能类型. 许多重要的生态理论在一个community中预测物种数目. 多样性被认为是一个community的'emergent property', 在community水平作用, 而不是在个体物种水平. 多样性也是用于保护管理的一个重要指标, 作为生态环境'well-being'的指标.

Diversity包含两个组成: species richness(community中的物种数目)和evenness(物种丰度分布(SAD)的形状, 表现为一些物种常见而另一些罕见的现实情况)



Adapted from Gotelli & Chao (2013)

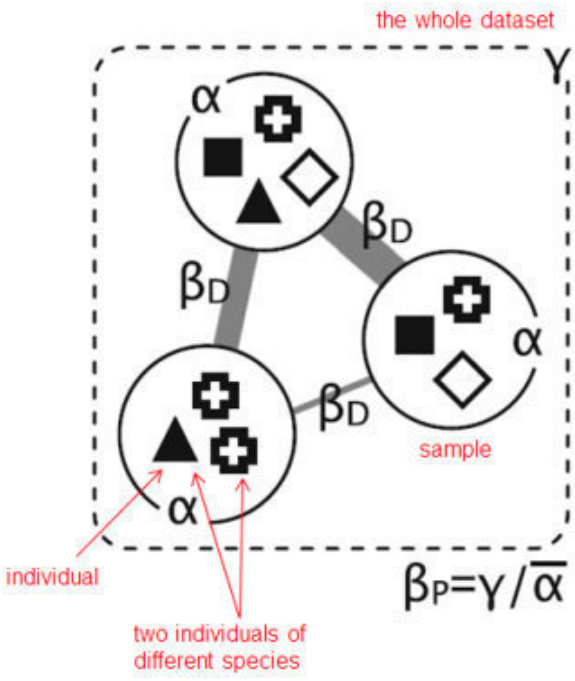
由于采用取样的方法来评估community的多样性, 同时取样往往是不完整的. 来自取样数据的多样性评估依靠与取样效果, 并且假如多样性(alpha, beta, gamma)需要在不同的community间比较, 那么, 取样效果就应该被标准化. 可通过rarefaction curves来完成, 允许根据相同的个体数目或相同的样本数目比较多样性. 另一种使用diversity estimators的是评估没有被取样采集到的物种数目, 同时假设在取样效果增加时就会看到这些物种.

两类diversity estimators存在, 第一种是根据丰度数据(一个样本中物种的丰度, 根据个体数据或biomass量体现); 第二种是incidence data(物种存在某一组样本中的频率, 在每个样本中只含有species incidence, 例如, 存在-缺失信息被记录).

存在多种用于指定多样性差异形式的概念. `alpha`, `beta`和`gamma`多样性.

Whittaker构建了根据Fisher's alpha, 并且延伸局部物种richness(alpha diversity)的概念到区域物种richness(gamma diversity), 和样本间物种组成的改变(beta diversity).

Beta divesity是一个和alpha, gamma diversity根本上不同的概念, 并且它自身代表一个复杂的方面. Beta diversity可看作species turnover(directional exchange of species among pair of samples or along spatial, temporal or environmental gradient)或者物种组成的变化(no-directional description of heterogeneity in species composition within the dataset). 或者说, beta diversity既可以被看作是不同的多样性(考虑物种组成的差异), 或是成比例的多样性(在一个区域或局部水平上的物种比例, gamma vs alpha diversity).



alpha diversity, 样本中物种数目; gamma diversity, 数据集, community或区域中物种数目; beta d diversity, 差异性多样性, 成对样本间的相似性(jaccard similarity index), 灰色带更宽, 相似度越高; beta p diversity, 成比例多样性, 两个被调查考虑alpha或gamma水平上的物种数目水平之间的关系.

[Diversity indices][https://www.davidzeleny.net/anadat-r/doku.php/en:div-ind]

species richness

表示为: S, 是最直观的和最自然的多样性指针(index). 由于它平等的评估所有物种, 独立与它们的相对丰度, 也是对不同取样效果最敏感的.

Shannon index 又称为Shannon entropy, Shannon-Wiener, 考虑了物种的richness和evenness. 从信息理论推导而来, 并且代表了哪一些物种通过从community内随机选择而能被预测的不确定性. 假如, community仅包含一个物种, 那么不确定性为0. community中包含越多的物种, 不确定性就越高; 在一个多样性的community, 我们无法猜测哪些物种将会被随机选择出来. 然而, 假如community包含很多物种, 但是其中一个或几个显著性存在其中(prevails), 那么不确定性就不会很高, 因为我们有很高的概率随机选择的个体就是丰度最高的物种. 这就是为什么Shannon index随着richness和evenness而增加, 同时richness给予的权重高于evenness.

真实的生态数据, H值一般为1.5-3.5(the units are bits of information); 同时该值根据底数不同变化. 给定richeness条件下, community的最大的H index发生在所有物种都是均匀分布的完美情况下(all species have the relative proportion).

Shannon index

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

$$H_{max} = \log S$$ [1)]

where
S = species richness,
p_i = relative abundance of species i,
log = usually natural logarithm (i.e. log_e or *ln*)

Simpson index 又称为Simpson concentration index, 也会考虑richness和evenness, 但是相对于Shannon, evenness对其影响超过richness. 它代表两个随机被选择个体是相同species的概率. 由于该概率随着物种richness升高而降低, Simpson index也会随之而降低, which is not too intuitive. 因此, 使用Goni-Simpson index 更有意义, 1 - Simpson index, 该值随着richness增加而增加.

其值D的范围为0-1, 单位为一个probability. 当community中物种richness超过10, Simpson index就会主要被evenness所影响.

Simpson index

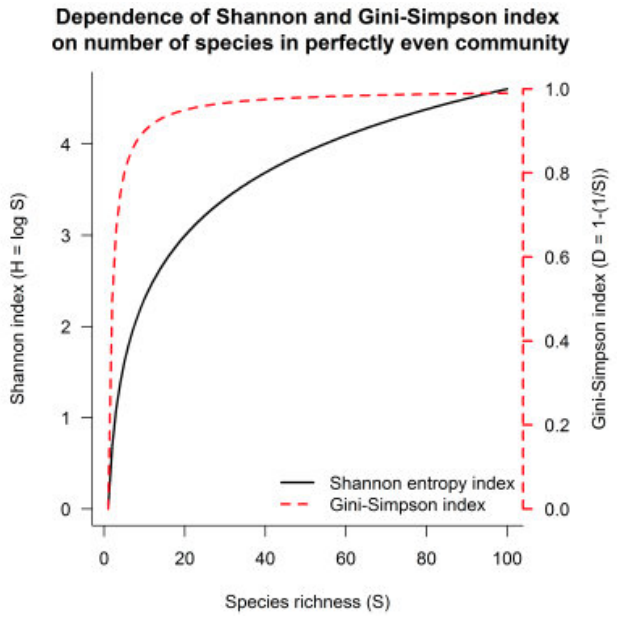
$$D = \sum_{i=1}^{S} p_i^{\,2} \quad D_{max} = \frac{1}{S}$$ [3)]

Gini-Simpson index

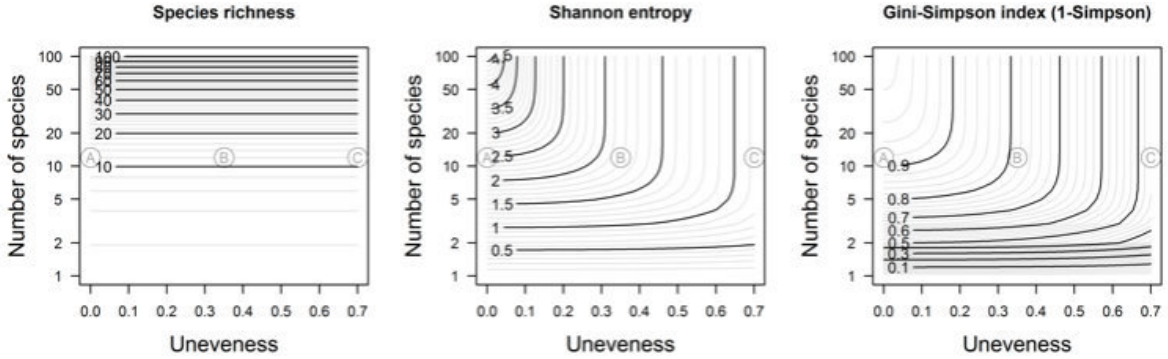
$$GS = 1\text{-}D$$

where
S = species richness,
p_i = relative abundance of species i,

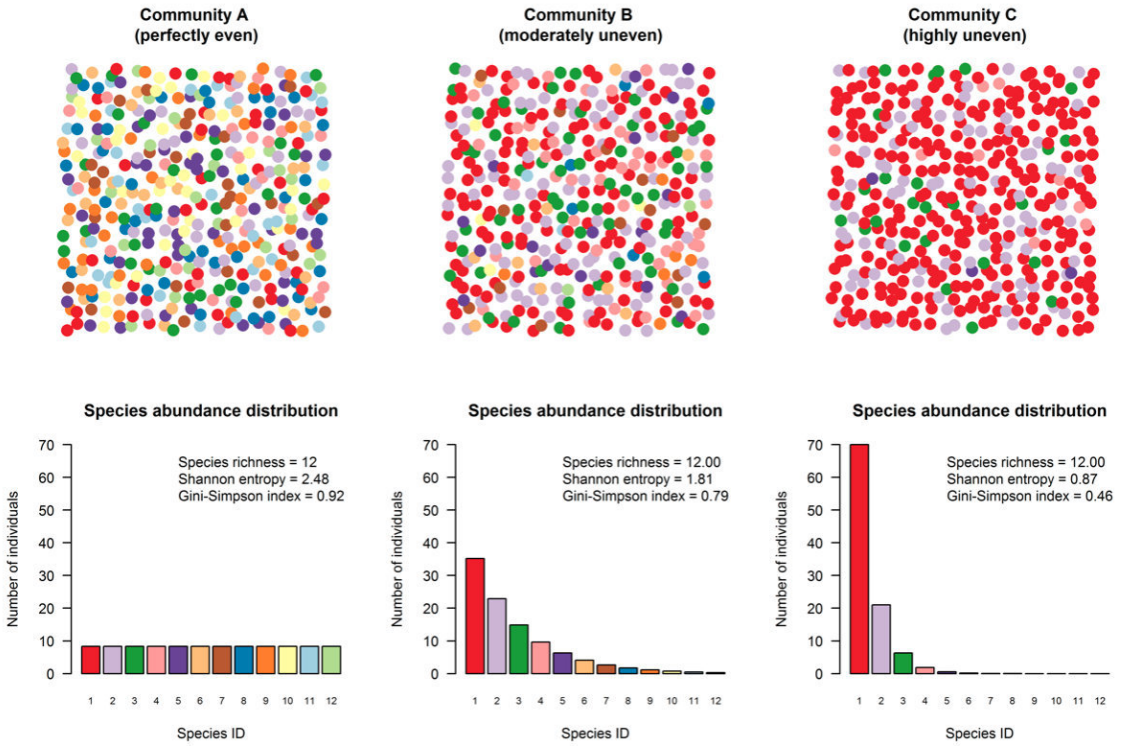
Comparison of species richness, Shanon index and Simpson index 在一个完美均匀的community条件下, Shannon和Simpson index随之community中物种数目的增加而非线形增加; Gini-Simpson index增加更快.



三个diversity indices(richness, Shannon, Simpson)针对不同均匀的多样性分布如下:



A/B/C标记对应了三个不同均匀程度的community, 每个community包含12个物种(species richness=12)



随着evenness下降, Shannon entropy和Gini-Simpson index都是下降的.

evenness 是一个认为描述community中物种相对丰度的模型. 有多种方式可以计算evenenss.

Shannon' evenness, 又称为Pielou's J, 被计算为Shannon index的比例(from real community). S species, p1,p2,p3..pi为物种的相对丰度. 针对相同丰度下最大Shannon index是1, 此时所有物种都拥有相同的丰度(p1,p2,p3...pi = 1/S).

Shannon's evenness

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

Simpson's evenness, 又称为均匀度(equitability), 是除以物种观察数目得到的Simpson's有效物种数目. 物种有效数目(ENS, effective number of species)为相同丰度物种的数目. 针对Simpson's D, 物种有效数目为1/D.

Simpson's evenness

$$equitability = \frac{\frac{1}{D}}{S} = \frac{1}{DS}$$

Effective numbers of species(ENS)

Effective number of species(ENS), i.e. number of species in equivalent community (i.e. the one which has the same value of diversity index as the community in question) composed of equally-abundant species.

在完美均匀community条件下, ENS等同于物种richness; 针对不均匀community, ENS常小于S. 转换公式如下:

Effective number of species

- for species richness = S
- for Shannon index = e^H (exponential of Shannon entropy index)
- for Simpson index = $1/D$ (reciprocal of Simpson concentration index)

Hill numbers

Hill numbers

$$ {}^{q}D = \left(\sum_{i=1}^{S} p_i^{\,q} \right)^{\frac{1}{(1-q)}} $$

For q = 0, 1 and 2 (also noted as N_0, N_1 and N_2):

$ {}^{0}D = S $ (species richness)

$ {}^{1}D = e^{H} $ (exponential of Shannon entropy)

$ {}^{2}D = \frac{1}{D} $ (reciprocal of Simpson index)

物种richness, Shannon entropy和Simpson' concentration index都是相同多样性indices家族成员, 后称之为Hill numbers. 个体Hill nubmer之间差异为参数q, which quantifies how much the measure discounts rare species when calculating diversity. q=0, 为简单的物种richness; q=1, 为Shannon diversity; q=2, 为Simpson diversity. 当q>0, indices不考虑罕有物种, 当q<0, indices忽略常见物种而仅考虑罕有物种.

Diversity profiles 绘制coefficient q绘制物种有效数目图, 增加q会降低罕有物种在多样性测量方面的影响. q=0等同于物种richness, q=1等同于Shannon diversity, q=2等同于Simpson diversity.