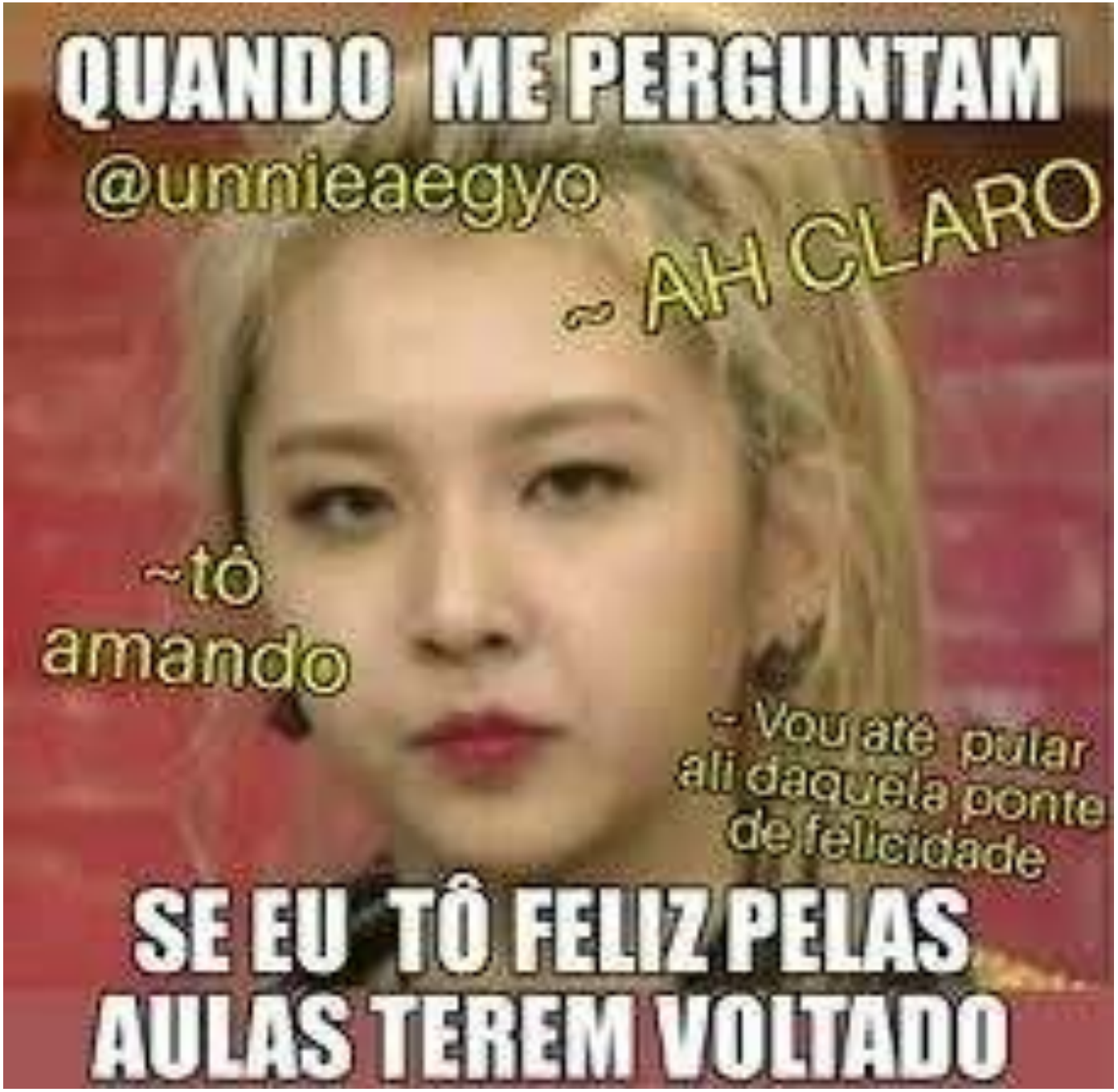
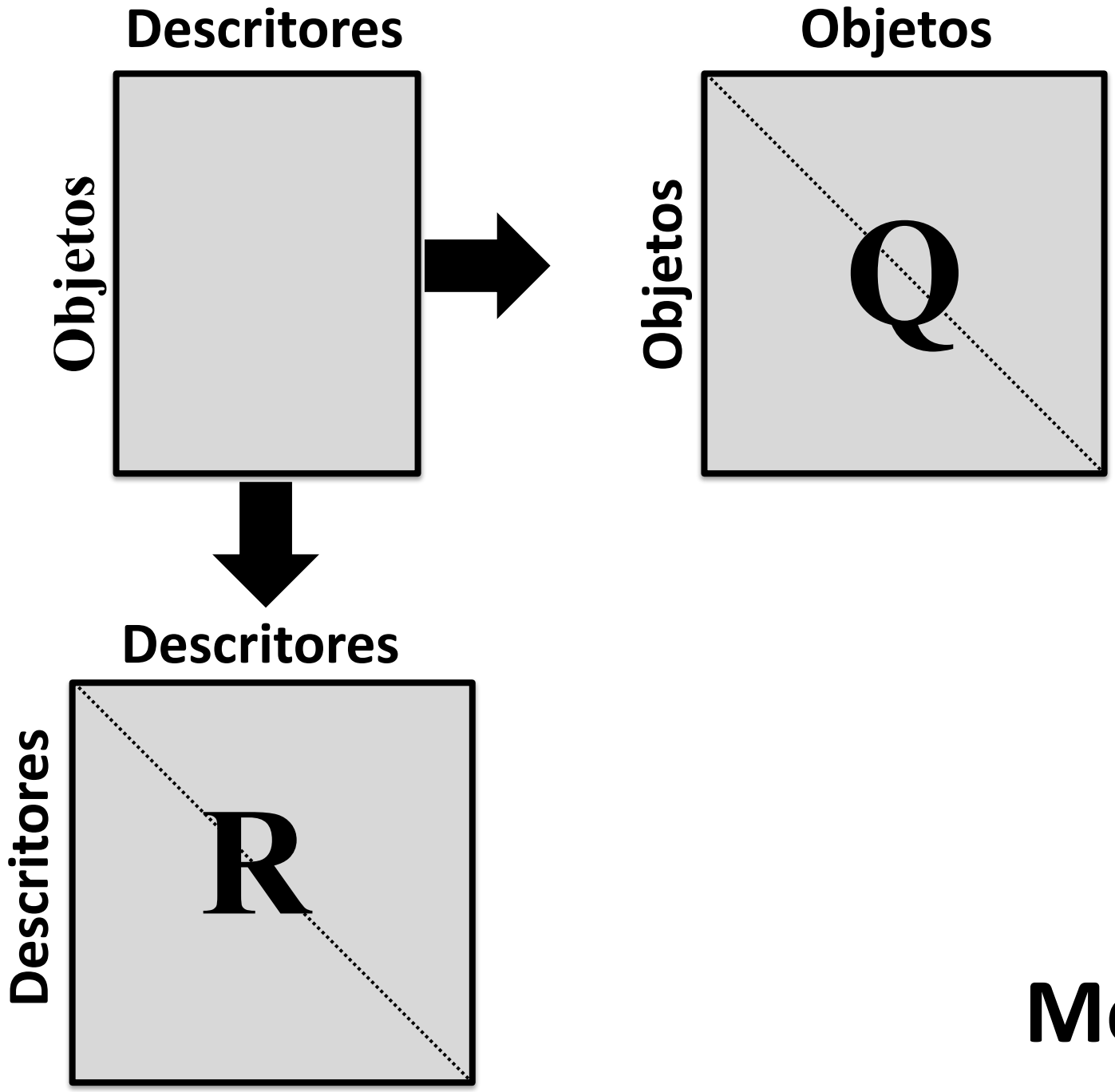

QUANDO ME PERGUNTAM
@unnieaegyo
~ AH CLARO
~tô amando
~ Vou até pular ali daquela ponte de felicidade
SE EU TÔ FELIZ PELAS AULAS TEREM VOLTADO

Aula 2: Coeficientes de associação

Capítulo 7 Legendre & Legendre

Ao fim da aula você deverá ter compreendido

- O que são e em que contexto são usados coeficientes de distância
- Diferença entre distância e similaridade
- Tipos de coeficientes
  - Métricos
  - Semi-métricos
  - Não-métricos
- Coeficientes para descritores e objetos
- Como escolher os coeficientes de acordo com o tipo de dado (item 7.6 L&L)



Modo R e Q

Slide T. Gonçalves-Souza

Medidas de associação

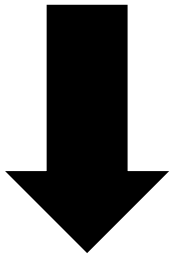
- **Q mode** => Entre objetos (espécies) => (dis)similaridade
  - No R, trabalhamos sempre com dissimilaridade

- **R mode** => Entre descritores => correlação ou covariância

Similaridade e Distância

- **Similaridades** são máximas (*S=1*) quando dois objetos são idênticos
- O comprimento da linha separando dois objetos é a sua **distância**
  - Não tem limite
- Distâncias são o contrário da similaridade
- D = 1 – S
- Existem outras transformações possíveis
  - $D = \sqrt{1-S}$ ou $D = \sqrt{1-S^2}$
  - Vamos ver adiante que elas podem ser úteis em alguns casos, e.g., para tornar uma matriz de distância euclidiana para ordenações (adespatial::is.euclid)

Se a similaridade (S) varia de ***0 a 1***



A distância (D) varia de ***1 a 0***

D = 1 - S

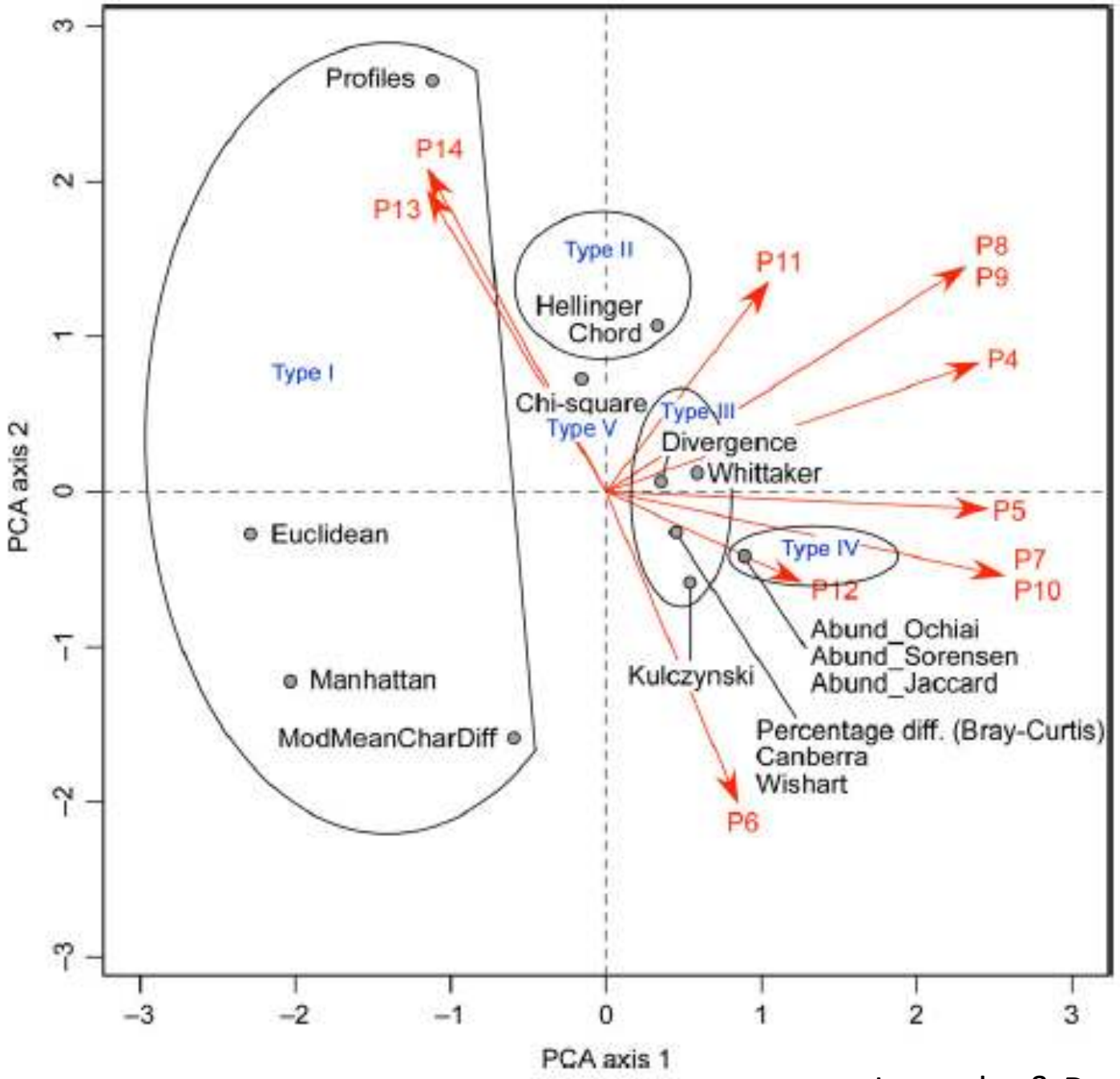
Slide T. Gonçalves-Souza

Table 7.2 Some properties of distance coefficients calculated from the similarity coefficients presented in Section 7.3. These properties (from Gower & Legendre, 1986), which will be used in Section 9.3, strictly apply when there are no missing data.

Similarity coefficient	$D = 1 - S$ metric, etc.	$D = 1 - S$ Euclidean	$D = \sqrt{1 - S}$ metric	$D = \sqrt{1 - S}$ Euclidean
$S_1 = \frac{a+d}{a+b+c+d}$ (simple matching; eq. 7.1)	metric	No	Yes	Yes
$S_2 = \frac{a+d}{a+2b+2c+d}$ (Rogers & Tanimoto; eq. 7.2)	metric	No	Yes	Yes
$S_3 = \frac{2a+2d}{2a+b+c+2d}$ (eq. 7.3)	semimetric	No	Yes	No
$S_4 = \frac{a+d}{b+c}$ (eq. 7.4)	nonmetric	No	No	No
$S_5 = \frac{1}{4}\left[\frac{a}{a+b} + \frac{a}{a+c} + \frac{d}{b+d} + \frac{d}{c+d}\right]$ (eq. 7.5)	semimetric	No	No	No
$S_6 = \frac{a}{\sqrt{(a+b)\,(a+c)}} \frac{d}{\sqrt{(b+d)\,(c+d)}}$ (eq. 7.6)	semimetric	No	Yes	Yes
$S_7 = \frac{a}{a+b+c}$ (Jaccard; eq. 7.10)	metric	No	Yes	Yes
$S_8 = \frac{2a}{2a+b+c}$ (Sørensen; eq. 7.11)	semimetric	No	Yes	Yes
$S_9 = \frac{3a}{3a+b+c}$ (eq. 7.12)	semimetric	No	No	No
$S_{10} = \frac{a}{a+2b+2c}$ (eq. 7.13)	metric	No	Yes	Yes
$S_{11} = \frac{a}{a+b+c+d}$ (Russell & Rao; eq. 7.14)	metric	No	Yes	Yes
$S_{12} = \frac{a}{b+c}$ (Kulczynski; eq. 7.15)	nonmetric	No	No	No

Table 7.2 Continued.

Similarity coefficient	$D = 1 - S$ metric, etc.	$D = 1 - S$ Euclidean	$D = \sqrt{1 - S}$ metric	$D = \sqrt{1 - S}$ Euclidean
$S_{13} = \frac{1}{2}\left[\frac{a}{a+b} + \frac{a}{a+c}\right]$ (eq. 7.16)	semimetric	No	No	No
$S_{14} = \frac{a}{\sqrt{(a+b)\,(a+c)}}$ (Ochiai; eq. 7.17)	semimetric	No	Yes	Yes
$S_{15} = \sum w_j s_j / \sum w_j$ (Gower; eq. 7.21)	metric	No	Yes	Likely* (S_1)
$S_{16} = \sum w_j s_j / \sum w_j$ (Estabrook & Rogers; eq. 7.22)	metric	No	Yes	Likely* (S_1)
$S_{17} = \frac{2W}{A+B}$ (Steinhaus; eq. 7.24)	semimetric	No	Likely* (S_8)	Likely* (S_8)
$S_{18} = \frac{1}{2}\left[\frac{W}{A} + \frac{W}{B}\right]$ (Kulczynski; eq. 7.25)	semimetric	No	No* (S_{13})	No* (S_{13})
$S_{19} = \sum w_j s_j / \sum w_j$ (Gower; eq. 7.26)	metric	No	Yes	Likely
$S_{20} = \sum w_j s_j / \sum w_j$ (Legendre & Chodorowski; 7.27)	metric	No	Yes	Likely* (S_7)
$S_{21} = 1 - \chi^2$ *metric* (eq. 7.28)	metric	Yes	Yes	Yes
$S_{22} = 2\,(\sum d)\,/\,n\,(n-1)$ (Goodall; eq. 7.29)	semimetric	No	–	–
$S_{23} = 1 - p(\chi^2)$ (Goodall; eq. 7.30)	semimetric	No	–	–
$S_{26} = (a + d/2)\,/\,p$ (Faith, 1983; eq. 7.18)	metric	No	Yes	Yes



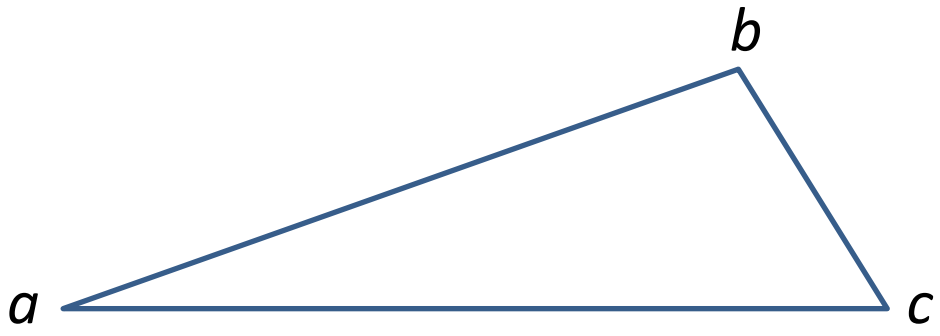
Legendre & De Cáceres 2013

Principais características dos índices de **similaridade**

- Coeficientes de **Similaridade** nunca são métricos
- Não podem ser usados para posicionar objetos num espaço métrico (Euclidiano)
  - Tem de ser convertidos em distâncias (adiante nessa aula)
- Simetria
  - Como lidam com duplo zero
  - Assimétricos => desconsideram duplo zero
- Tipos de dados
  - Qualitativos/quantitativos
    - Dados binários ou contínuos
  - Probabilísticos (e.g., Raup-Crick), não falaremos deles

Coeficientes de **distância**

- As quatro propriedades métricas
  1. Mínimo 0: se $a=b$, $\therefore D(a,b) = 0$
  2. Positividade: se $a \neq b$, $\therefore D(a,b) > 0$
  3. Simetria: $D(a,b)=D(b,a)$
  4. Inequalidade do triângulo: $D(a,b)+D(b,c) \geq D(a,c)$



3 Tipos de coeficientes de distância

- Distância => Dados quantitativos
- Métricos
  - Têm todas as 4 propriedades
- Semi-métricos
  - Violam a inequalidade do triângulo
- Não-métricos
  - Podem violar propriedades 1-3
  - Não são usados em ecologia

Coeficientes para objetos (Q mode)

Coeficientes métricos para dados quantitativos

- **Geralmente usados para variáveis ambientais, contínuas, medidas morfométricas, biomassa etc**

- **Distância Euclidiana**, Canberra, Mahalanobis, Manhattan, Chord, χ^2, Hellinger
- Hellinger não dá peso para espécies raras, diferentemente do Chi-quadrado

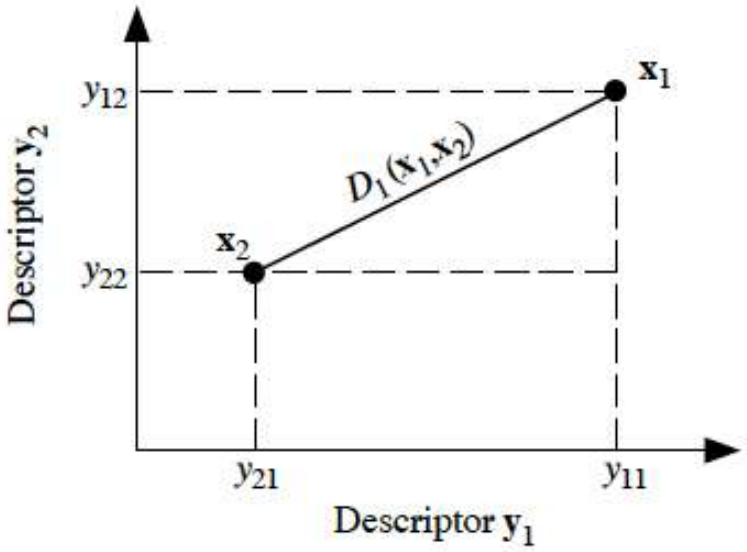
Exemplos

Euclidean distance

$$ED_{i,h} = \sqrt{\sum_{j=1}^{p} (a_{i,j} - a_{h,j})^2}$$

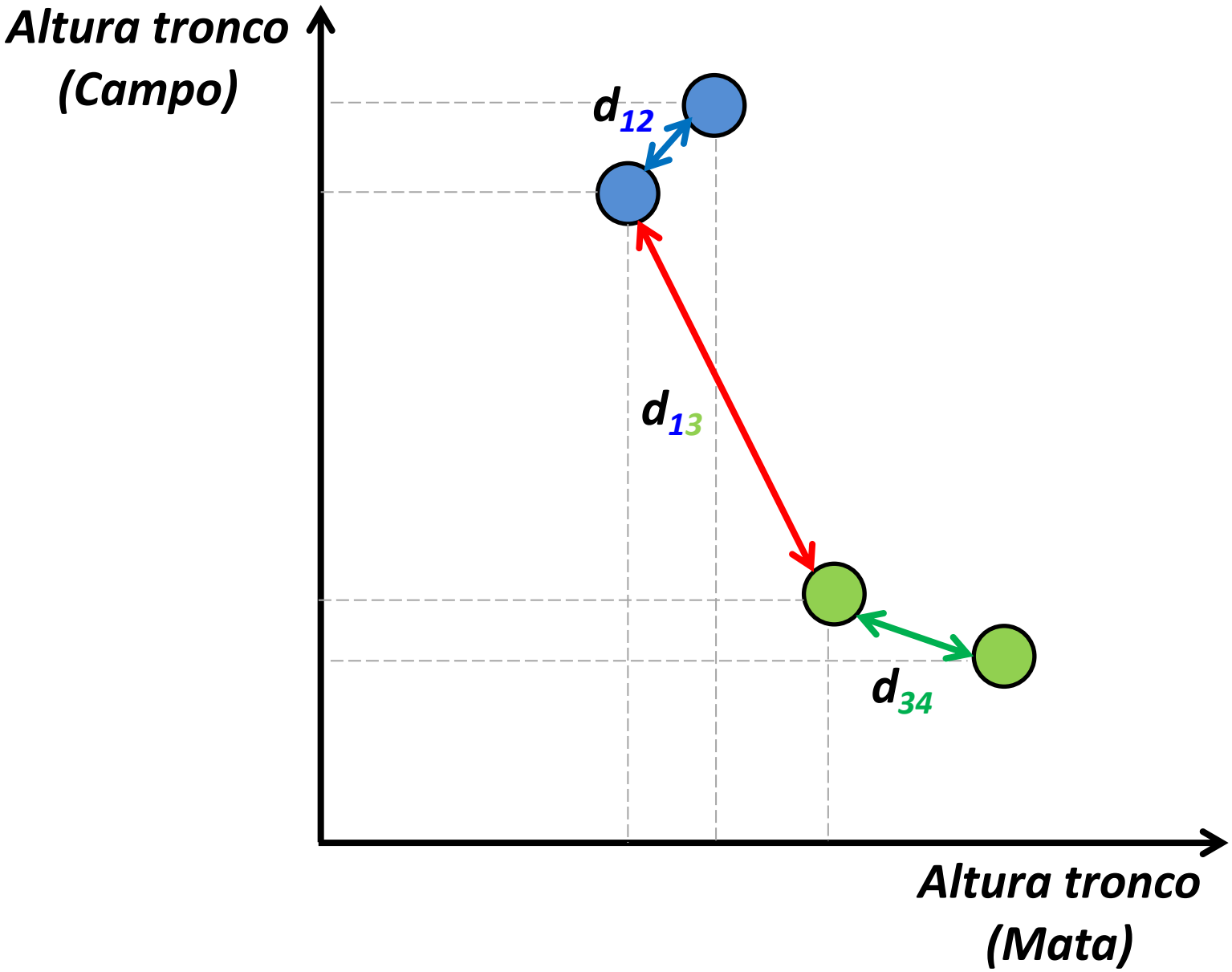
This formula is simply the Pythagorean theorem applied to p dimensions rather than the usual two dimensions (Fig. 6.2).

McCune & Grace 2002

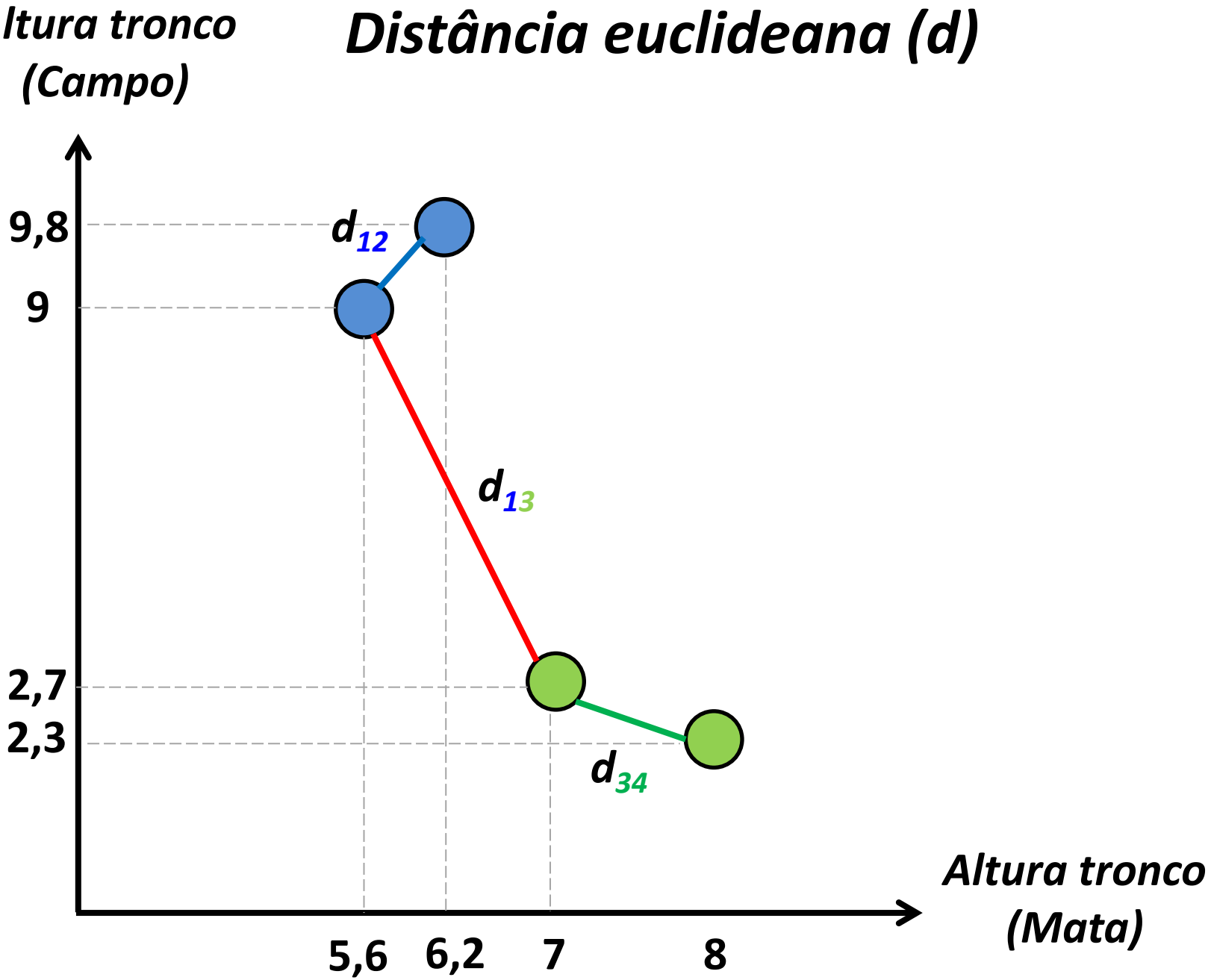


Legendre & Legendre 2012

Distância euclideana (d)

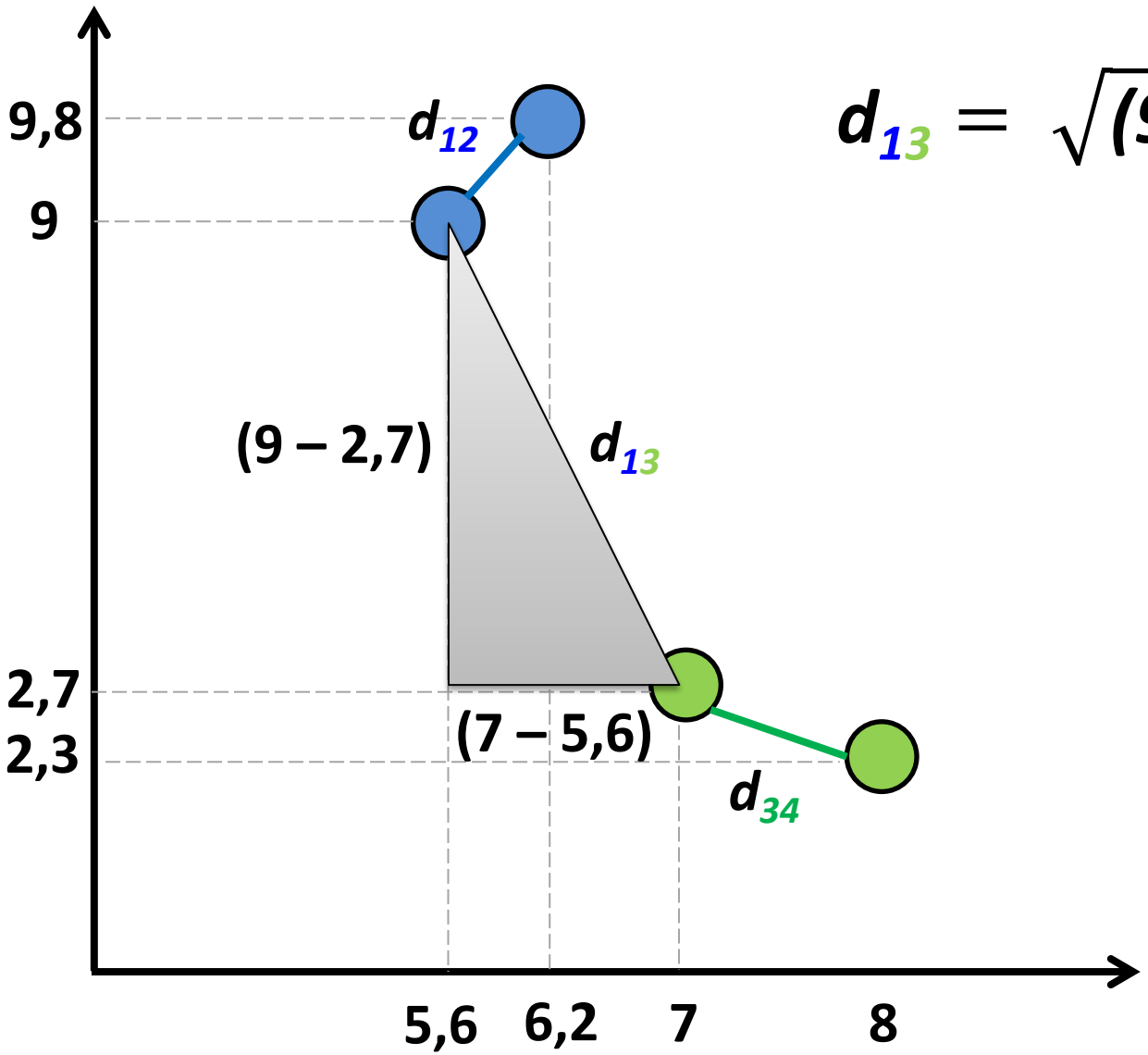


Distância euclideana (d)



Slide T. Gonçalves-Souza

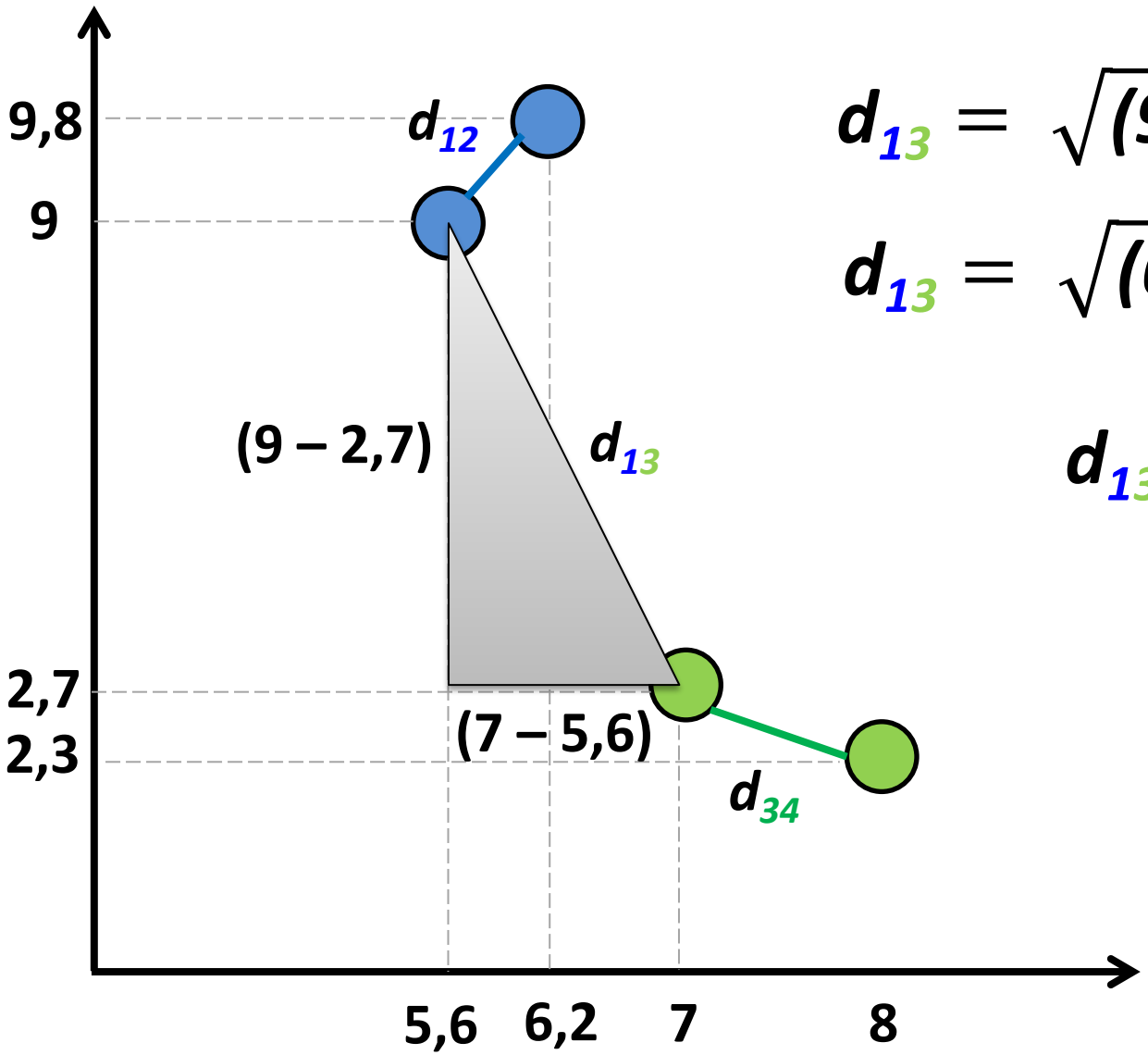
Distância euclideana (d)



$$d_{13} = \sqrt{(9-2,7)^2 + (7-5,6)^2}$$

Slide T. Gonçalves-Souza

Distância euclideana (d)



$$d_{13} = \sqrt{(9-2,7)^2 + (7-5,6)^2}$$

$$d_{13} = \sqrt{(6,3)^2 + (1,4)^2}$$

$$d_{13} = 6.454$$

Slide T. Gonçalves-Souza

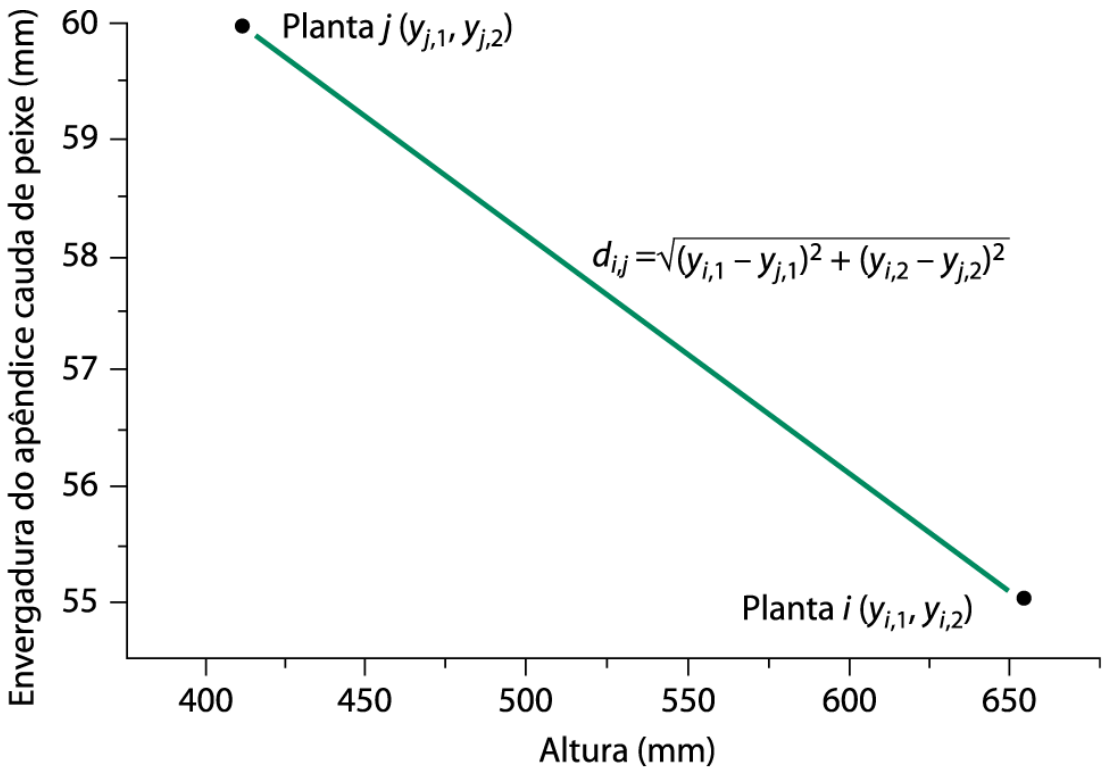


Figura 12.2 Em duas dimensões, a distância euclidiana $d_{i,j}$ é aquela em linha reta entre os pontos. Neste exemplo, as duas variáveis morfológicas que medimos são a altura da planta e a envergadura do apêndice da extremidade do lírio-cobra carnívoro, *Darlingtonia californica* (Tabela 12.1). Aqui, as medidas para duas plantas individuais são plotadas no espaço bivariado. A Equação 12.15 é usada para calcular a distância euclidiana entre elas.

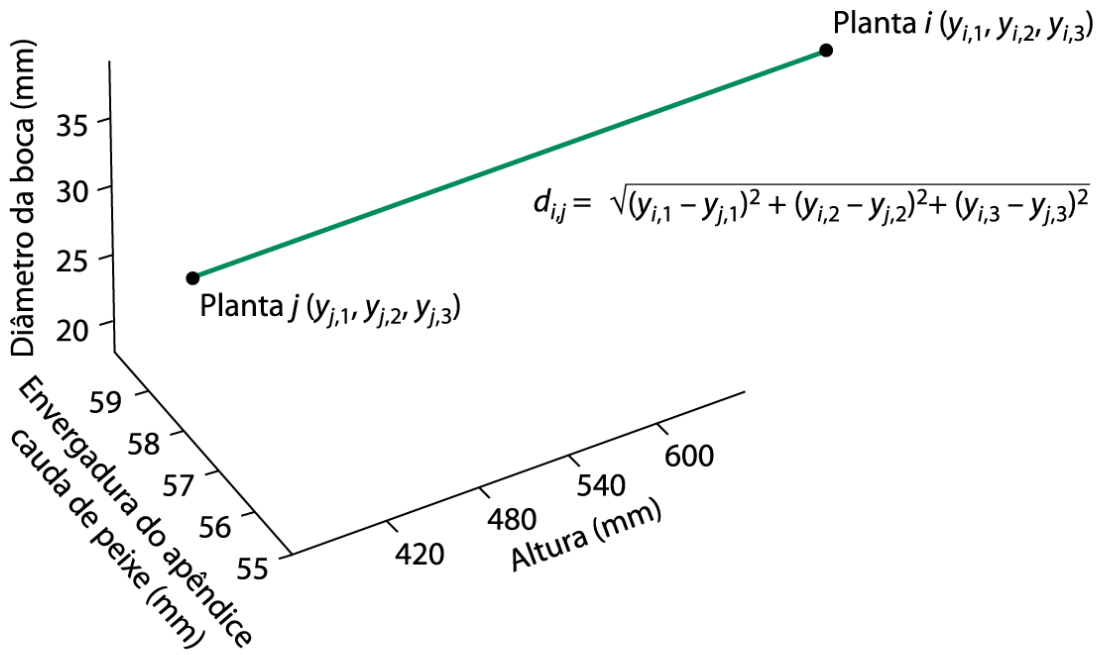
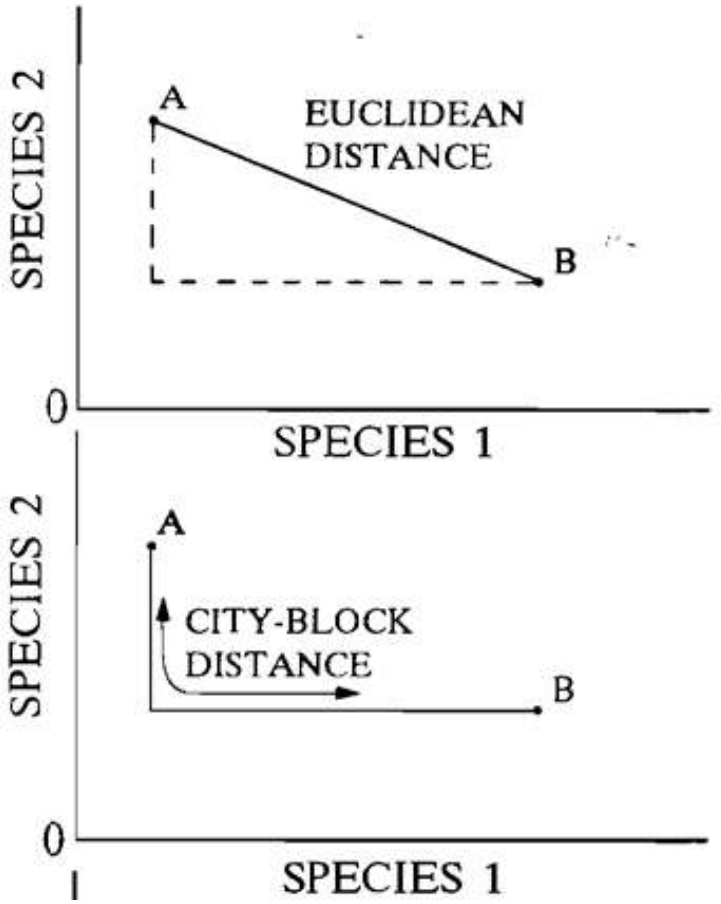


Figura 12.3 Medindo a distância euclidiana em um espaço tridimensional. O diâmetro da boca foi adicionado às duas variáveis morfológicas mostradas na Figura 12.2, e as três variáveis são plotadas no espaço tridimensional. A Equação 12.16 é usada para calcular a distância euclidiana, que é 241,58. Note que essa distância é virtualmente idêntica à euclidiana medida em duas dimensões (241,05, Figura 12.2). A razão é que a terceira variável (diâmetro da boca) tem média e variância muito menores que as duas primeiras variáveis, e então não afeta muito a medida de distância. Por essa razão, as variáveis devem ser padronizadas (usando a Equação 12.17) antes de calcular a distância entre indivíduos.

Distância de Manhattan

City-block distance (= Manhattan distance)

$$CB_{i,h} = \sum_{j=1}^{p} |a_{i,j} - a_{h,j}|$$



McCune & Grace 2002

Distância de Chord

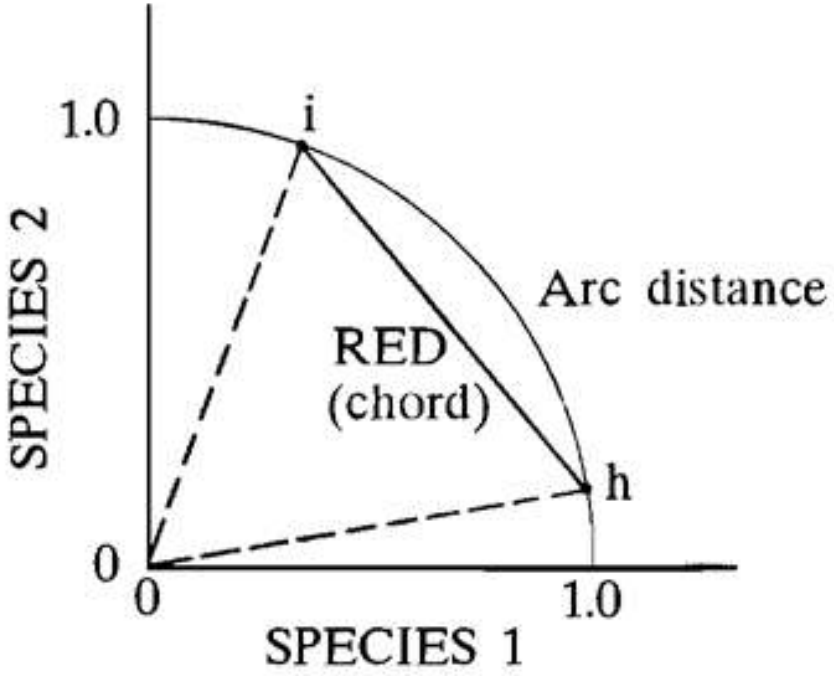


Figure 6.4. Relative Euclidean distance is the chord distance between two points on the surface of a unit hypersphere.

- Elimina diferenças entre abundância total de espécies
- Varia de 0 a √2

Distância de Mahalanobis

$$D_5^2(\mathbf{x}_1, \mathbf{x}_2) = \mathbf{d}_{12}\mathbf{V}^{-1}\mathbf{d}'_{12}$$

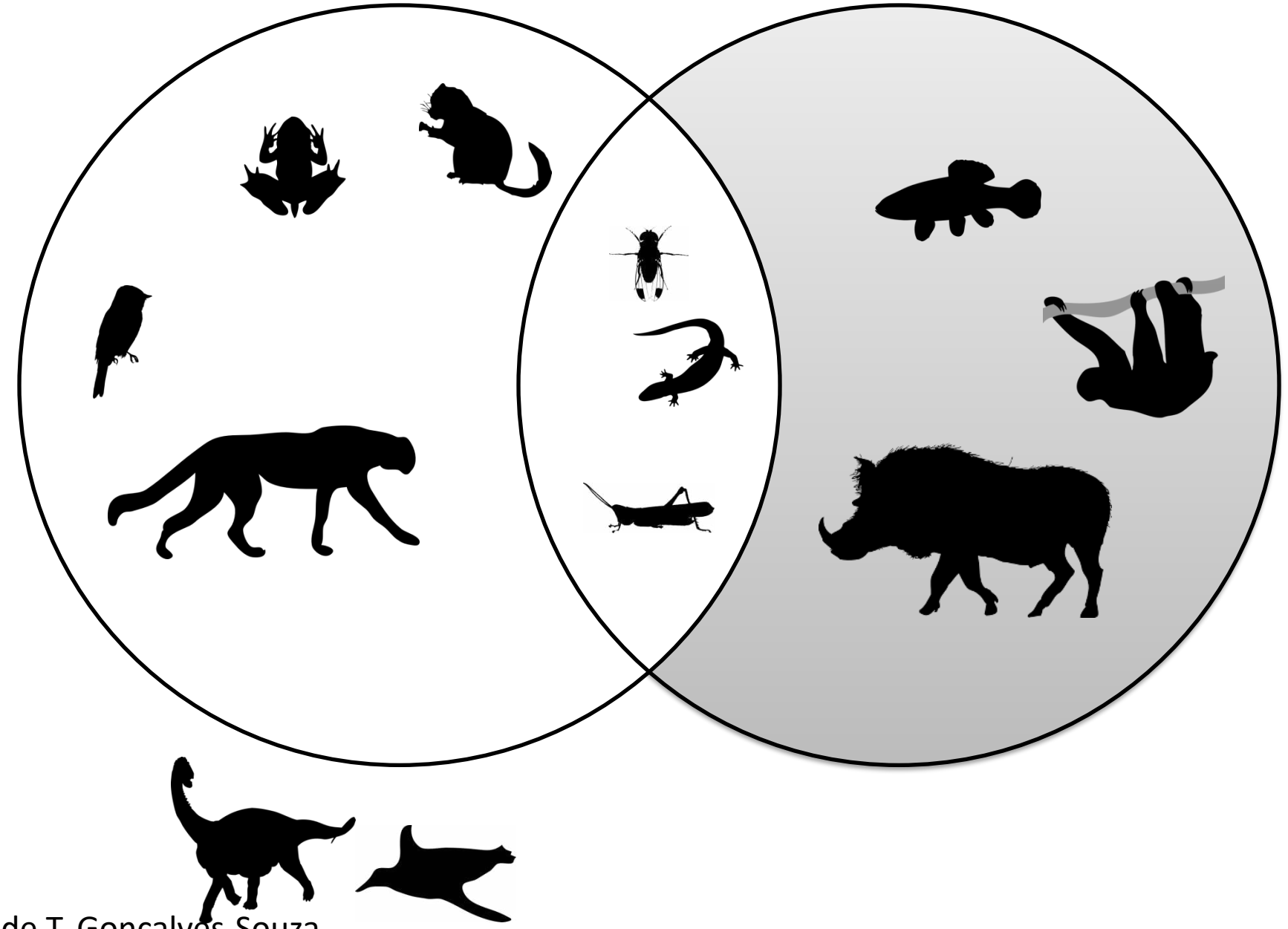
Computa a distância entre dois pontos num espaço não ortogonal. Leva em consideração a covariância entre descritores

Útil para comparar grupos de objetos, por isso é a distância preservada na Linear Discriminant analysis

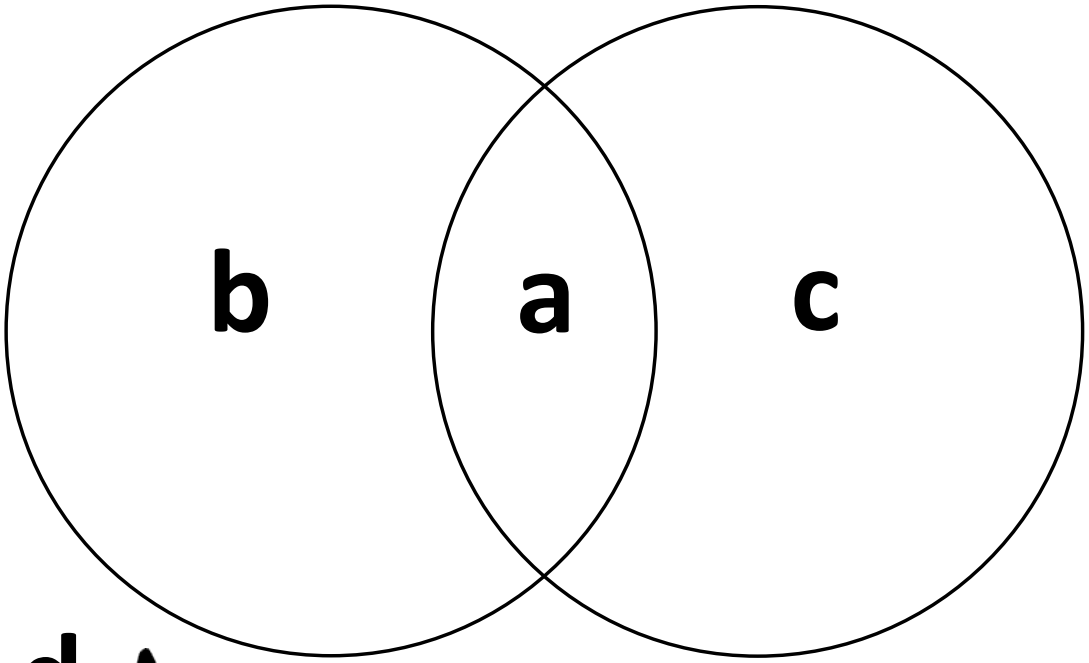
O problema do "duplo-zero"

Comunidade 1

Comunidade 2



Slide T. Gonçalves-Souza



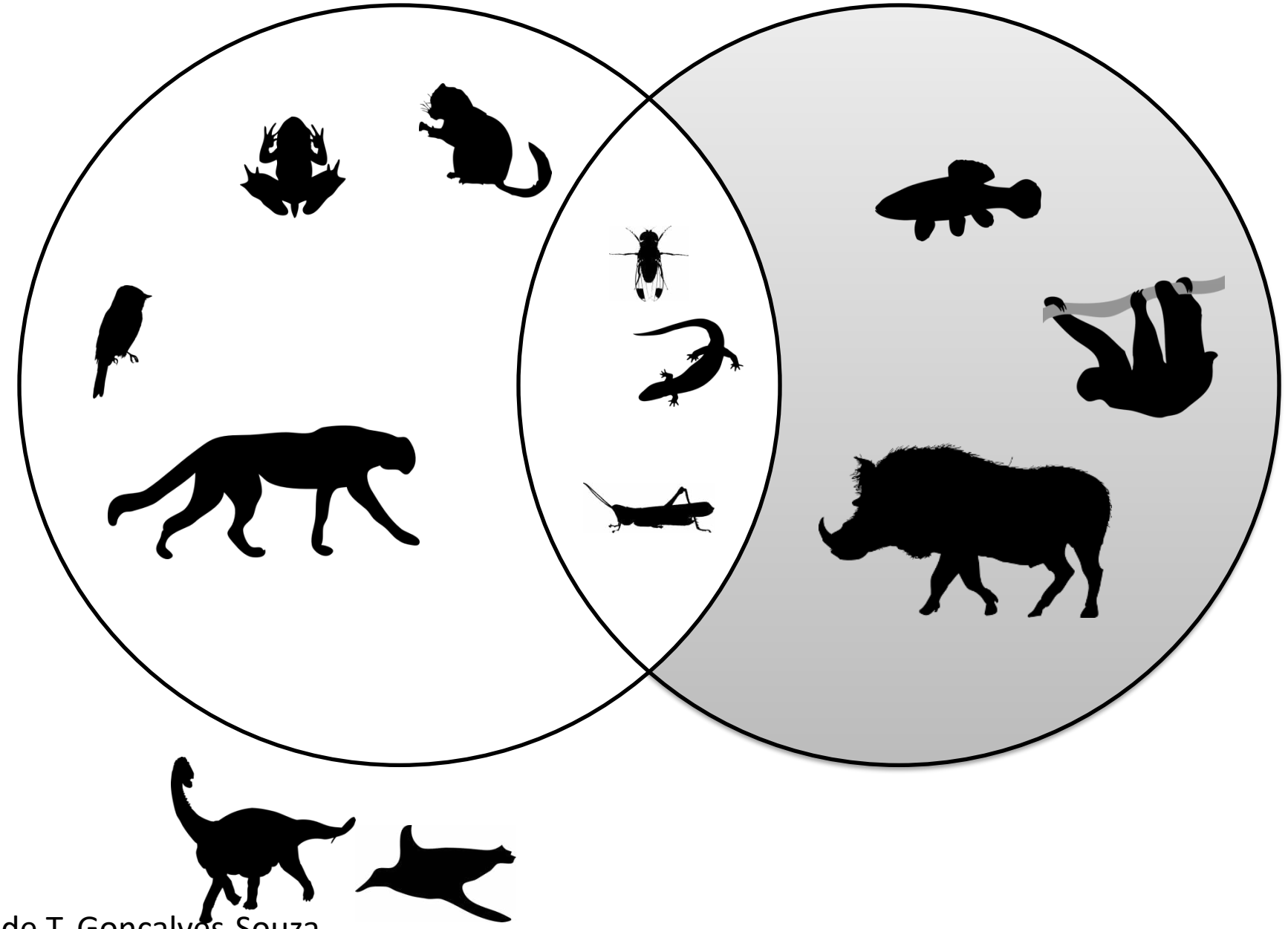
- **Problema com as métricas de distância**

- **Euclidiana – locais sem nenhuma espécie em comum pode possuir distância menor do que locais que compartilham espécies**

Slide T. Gonçalves-Souza

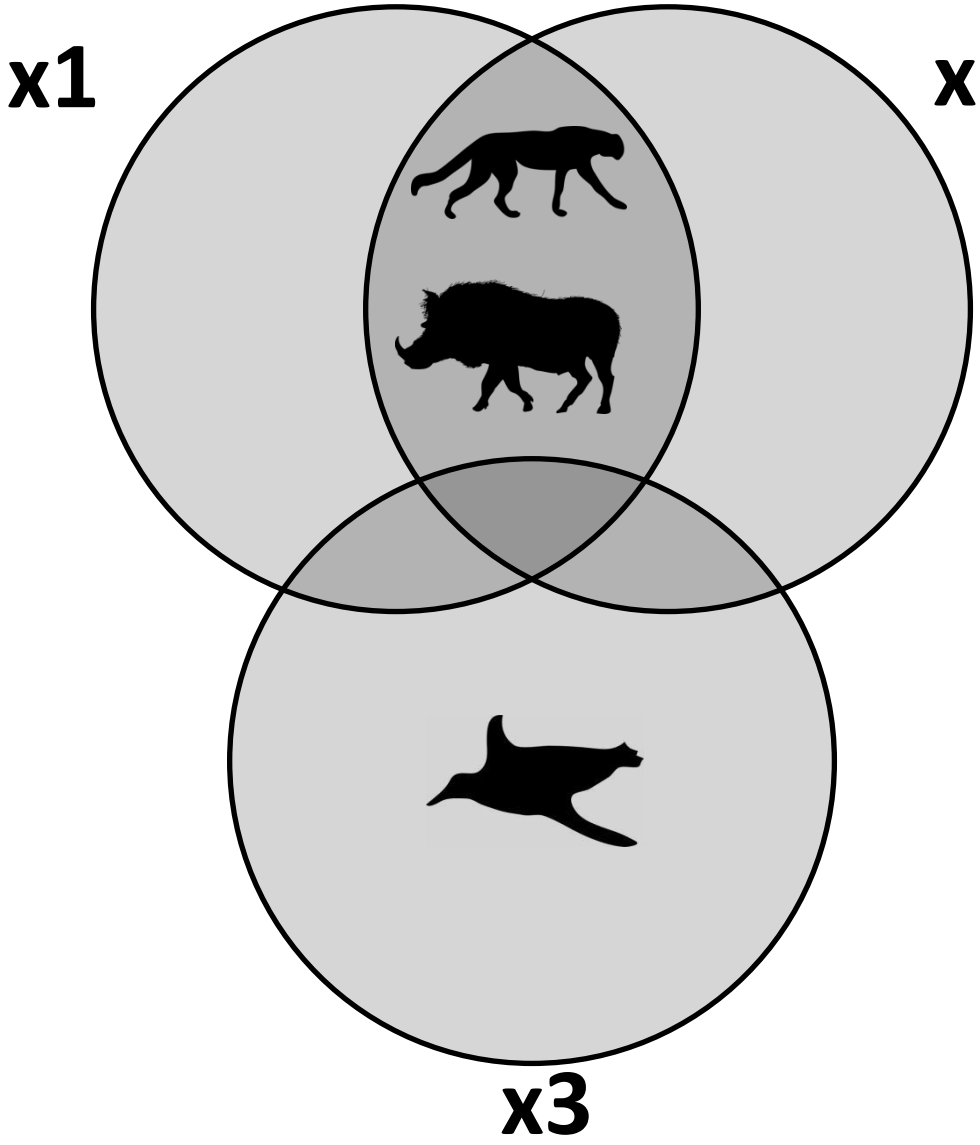
Comunidade 1

Comunidade 2



Slide T. Gonçalves-Souza

Exemplo com distância euclidiana

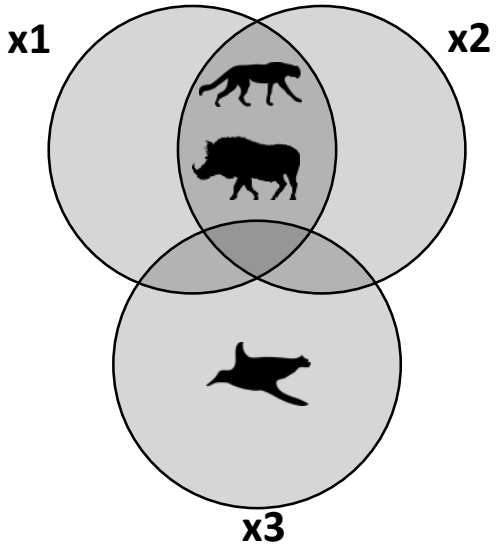


Legendre & Legendre 2012

Exemplo com distância euclidiana

Sites	Species		
	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$
$\mathbf{x}_1$	0	4	8
$\mathbf{x}_2$	0	1	1
$\mathbf{x}_3$	1	0	0

x1 e x2 compartilham todas espécies

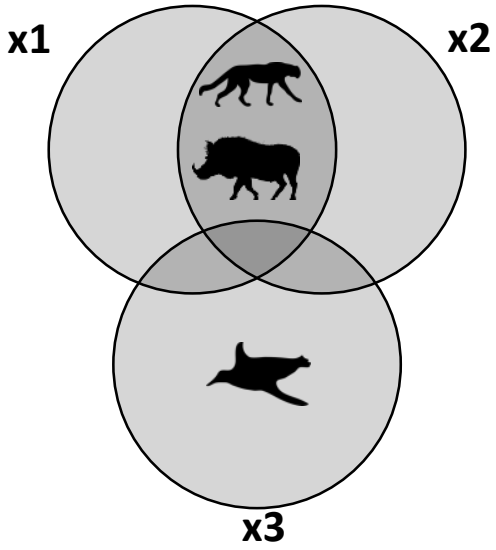


Legendre & Legendre 2012

Exemplo com distância euclidiana

Sites	Species		
	$\mathbf{y}_1$	$\mathbf{y}_2$	$\mathbf{y}_3$
$\mathbf{x}_1$	0	4	8
$\mathbf{x}_2$	0	1	1
$\mathbf{x}_3$	1	0	0

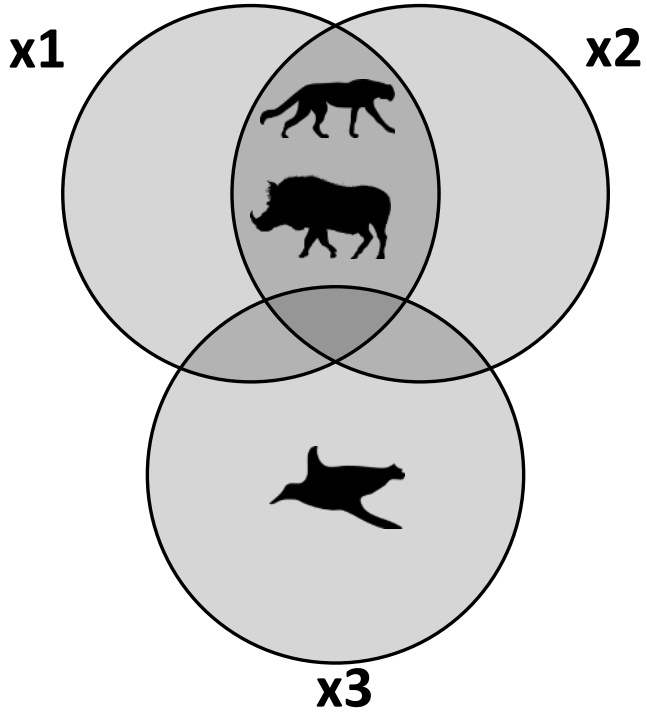
x2 e x3 não compartilham nenhuma espécie



Legendre & Legendre 2012

Exemplo com distância euclidiana

Sites	Species		
	y_1	y_2	y_3
x_1	0	4	8
x_2	0	1	1
x_3	1	0	0

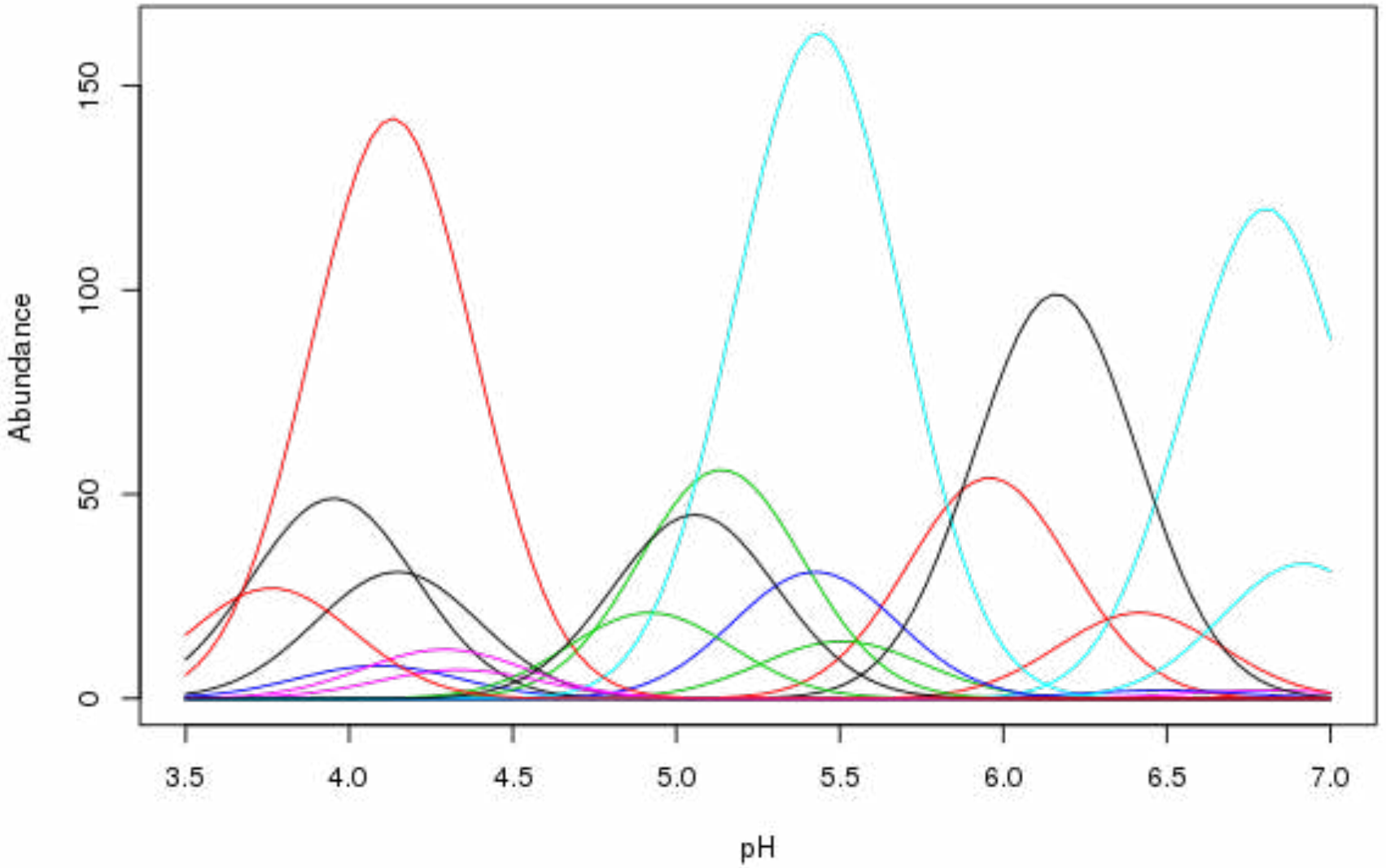


Sites	Sites		
	x_1	x_2	x_3
x_1	0	7.6158	9.0000
x_2	7.6158	0	1.7321
x_3	9.0000	1.7321	0

Legendre & Legendre 2012


poderia ser legal,e interessante...
mais não foi.
vk.com/yoda_advice

Teoria de Nicho ecológico



https://cran.r-project.org/web/packages/coenocliner/vignettes/coenocliner.pdf

Teoria de Nicho ecológico

- Amostramos indivíduos ao longo de gradientes ambientais
- Alguns sítios podem estar num extremo do gradiente e suas condições ambientais podem estar fora do espaço de nicho de uma(s) espécie(s)
- Isso resulta em vários zeros na matriz de composição

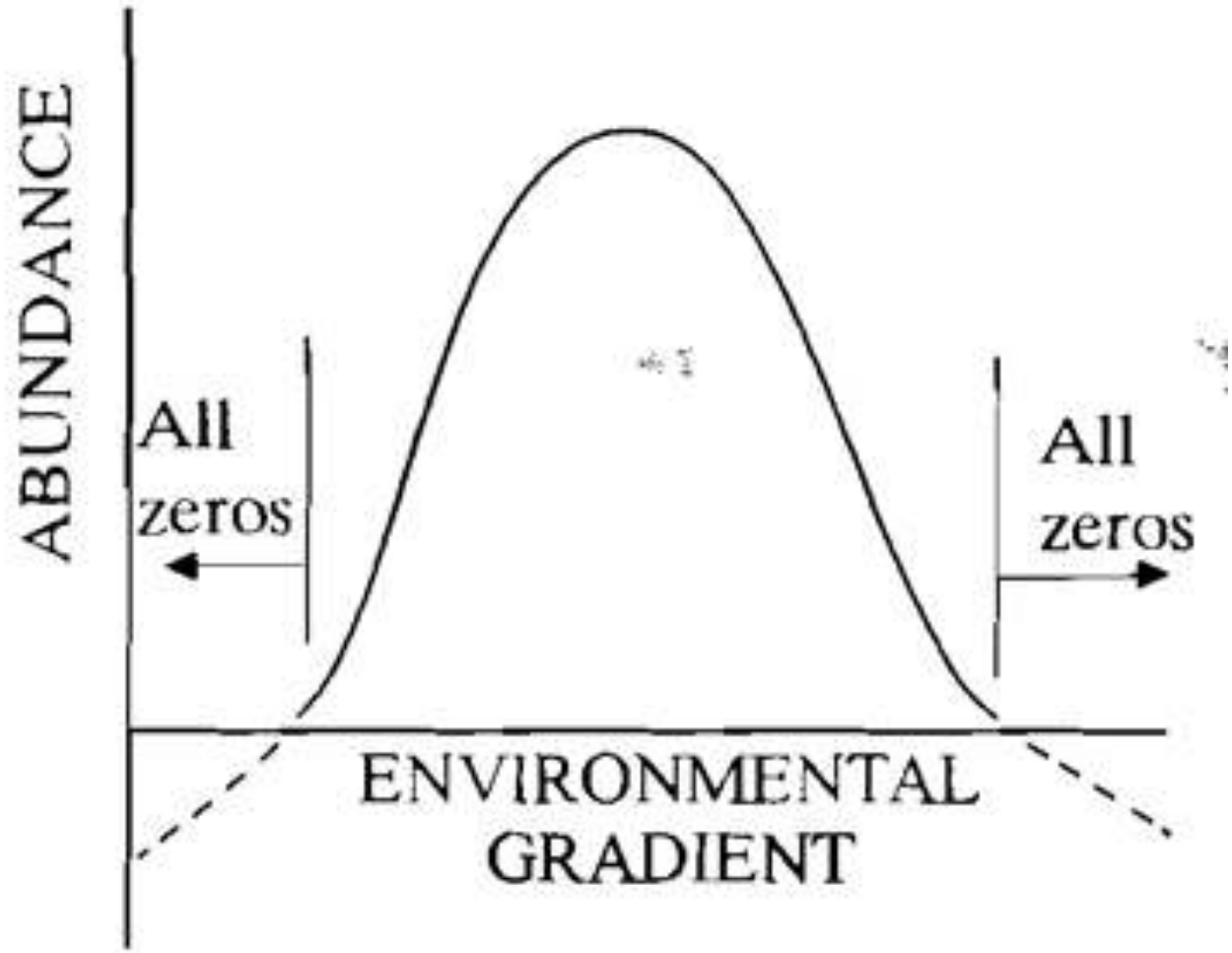


Figure 5.3. The zero truncation problem.

McCune & Grace 2002

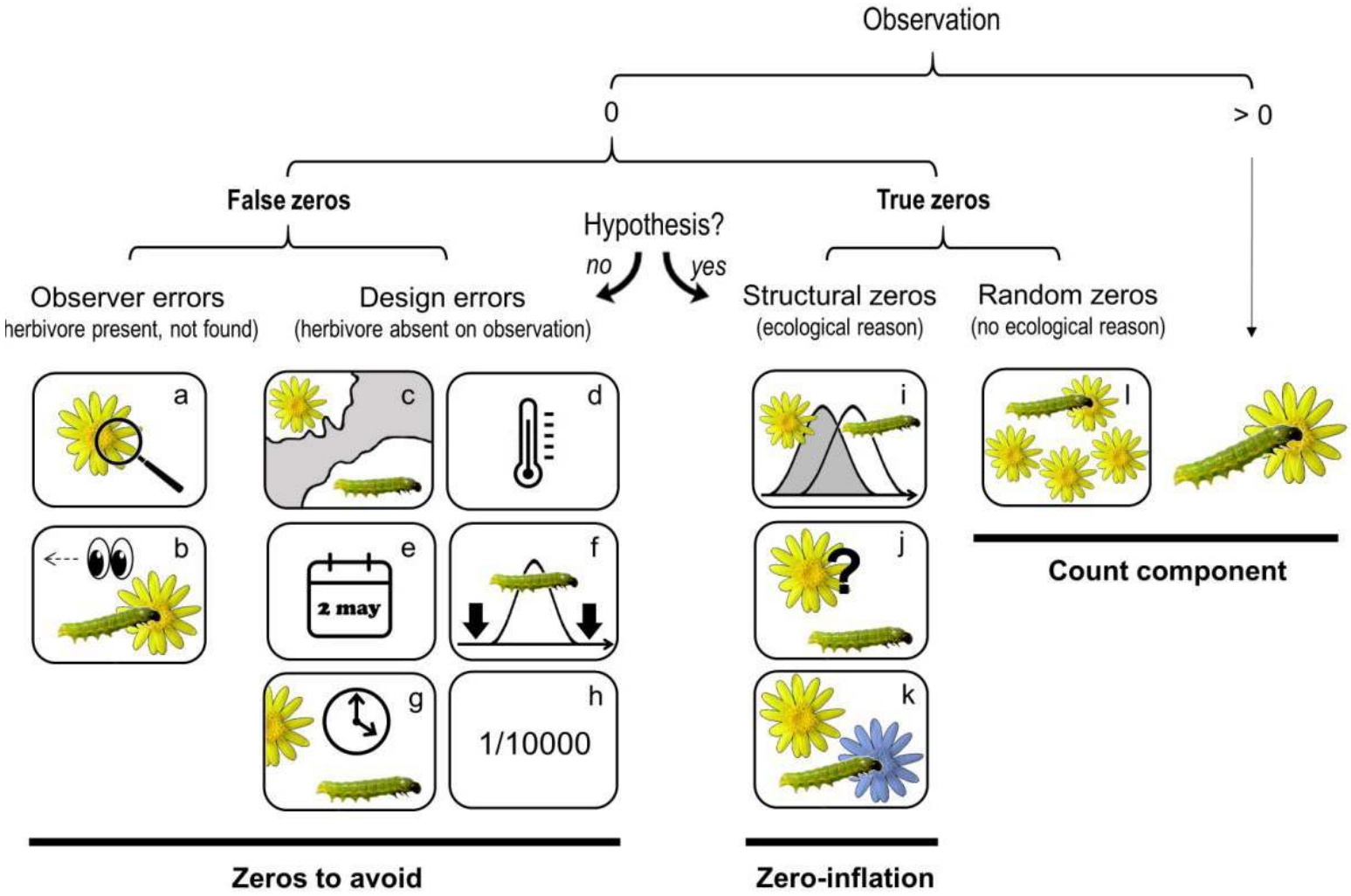
Methods in Ecology and Evolution
BRITISH ECOLOGICAL SOCIETY

Research Article | Full Access

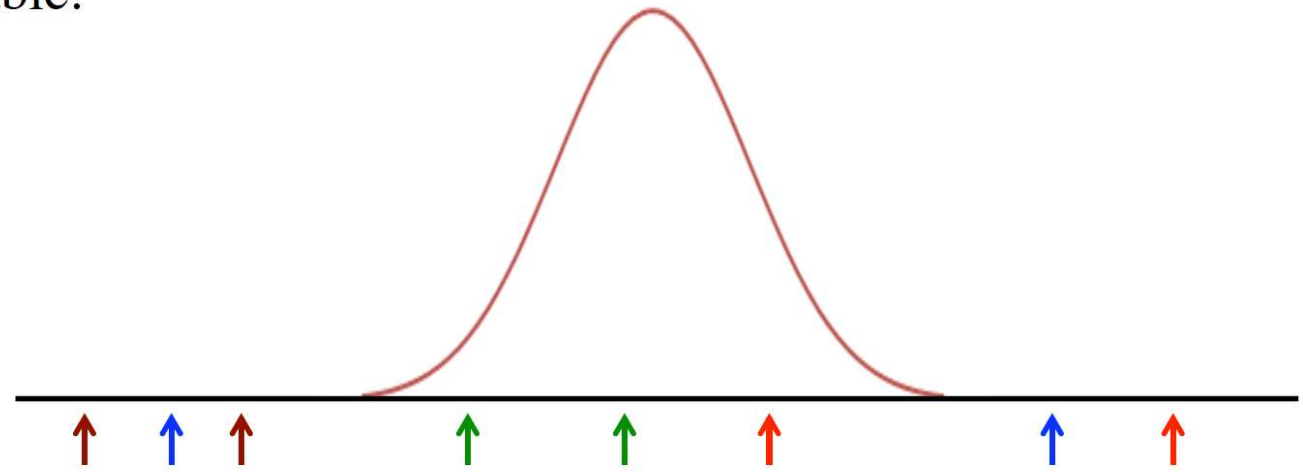
What does a zero mean? Understanding false, random and structural zeros in ecology

Anabel Blasco-Moreno, Marta Pérez-Casany, Pedro Puig, Maria Morante, Eva Castells

First published: 03 April 2019 | https://doi.org/10.1111/2041-210X.13185



Consider the distribution of a single species along that environmental variable:



For the presence or absence of that species, are the following pairs of observations an indication of …

	Similarity	Difference
Green arrows: 1, 1	✓	
Red arrows: 1, 0		✓
Brown arrows: 0, 0	Maybe	
Blue arrows: 0, 0		Maybe

Conclusion: double zeros do not have an unambiguous interpretation.

Slide P. Legendre

O problema dos duplos zeros

- Se uma espécie está presente em dois locais, isso é uma indicação de **similaridade** entre locais
- A presença no local 1 e ausência no local 2 indica diferença em condições ecológicas, apesar de erros de amostragem
- No entanto, se uma espécie está ausente em dois locais isso pode ser devido à:
  - Condições dos locais estão fora do nicho da sp, mas não dá pra saber se essas condições desses locais são parecidas ou muito diferentes
  - Erro de amostragem
  - Dinâmicas estocásticas (e.g., core-satellite) causando extinção local transitória
- Então, *não podemos dizer que dois locais são parecidos porque compartilham ausências de espécies*

Logo, precisamos de coeficientes que desconsiderem duplas ausências. Esses coeficientes são chamados de *assimétricos*


AH ENTENDI...
MAS NÃO COMPREENDI
memegenerator.es

Coeficientes para objetos (Q mode)

Coeficientes assimétricos para dados binários

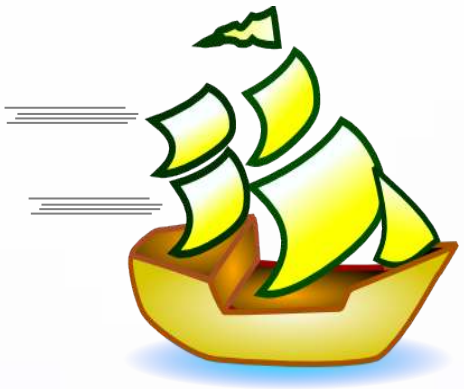
- Utilizado com dados de presença-ausência
- Não levam em conta duplo zeros (duplas ausências)
- Exemplos:
  - Jaccard
  - Sørensen
  - Ochiai

Navigating the multiple meanings of β diversity: a roadmap for the practicing ecologist

ECOLOGY LETTERS

Ecology Letters, (2011) **14**: 19–28

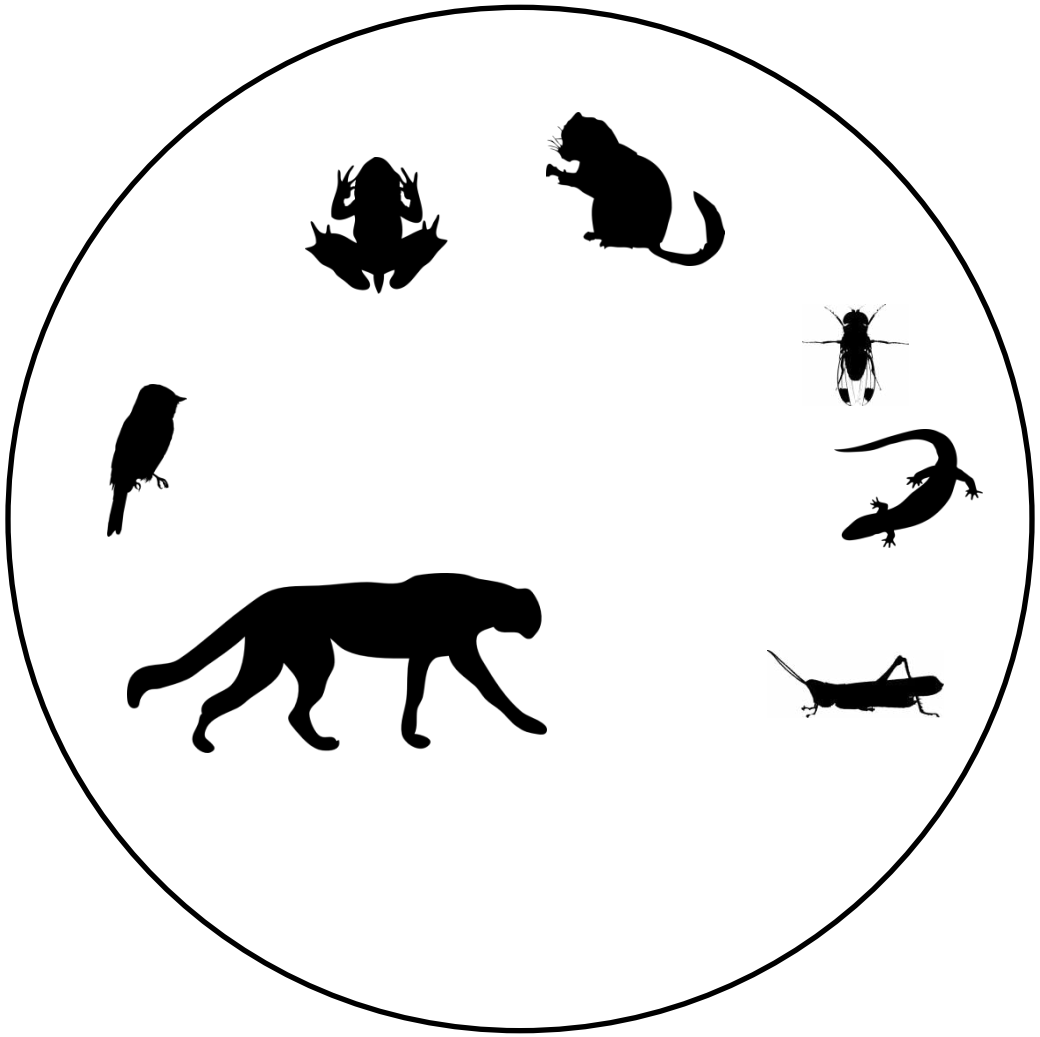
Marti Anderson et al. 2011



O que faz com que a comunidade de espécies seja mais ou menos semelhante uma das outras em diferentes locais e tempos?

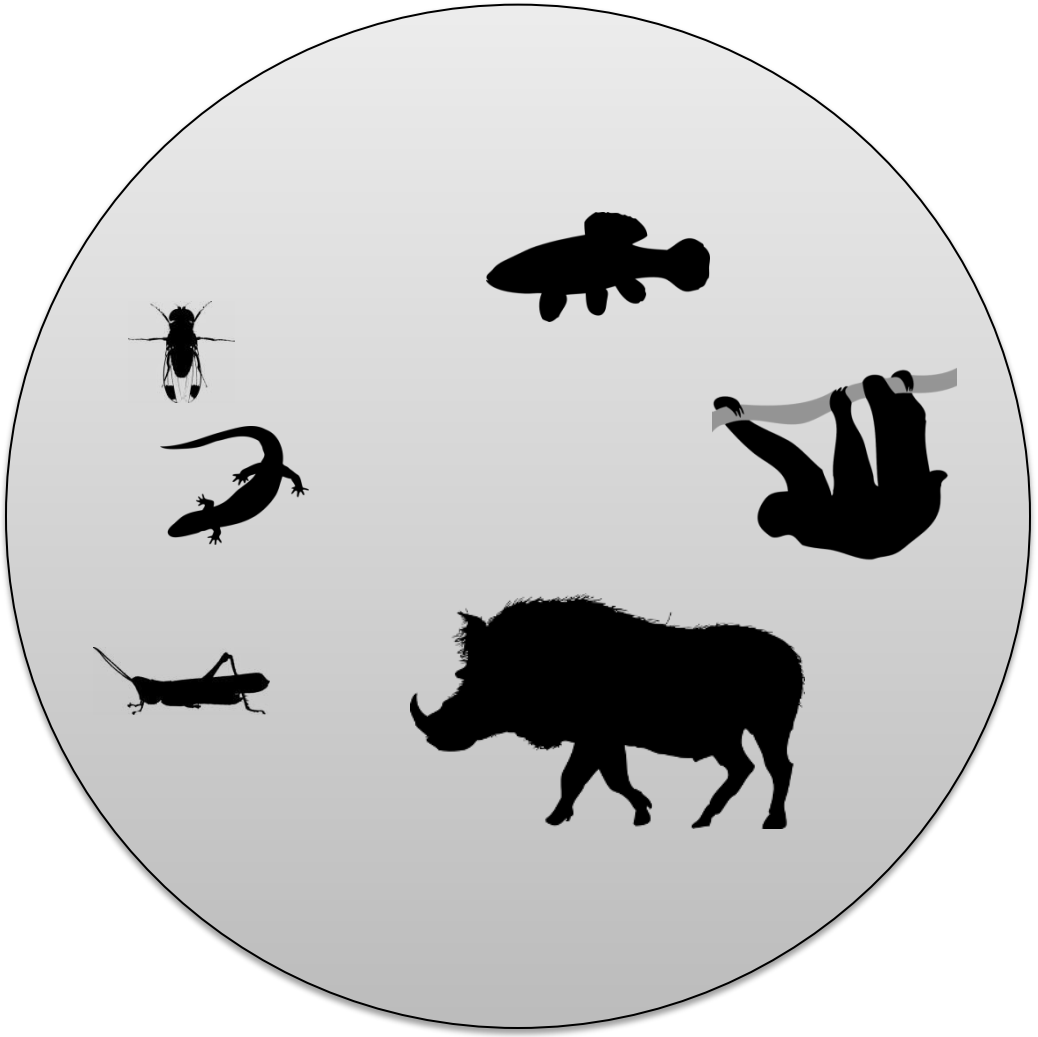
Slide T. Gonçalves-Souza

Comunidade 1

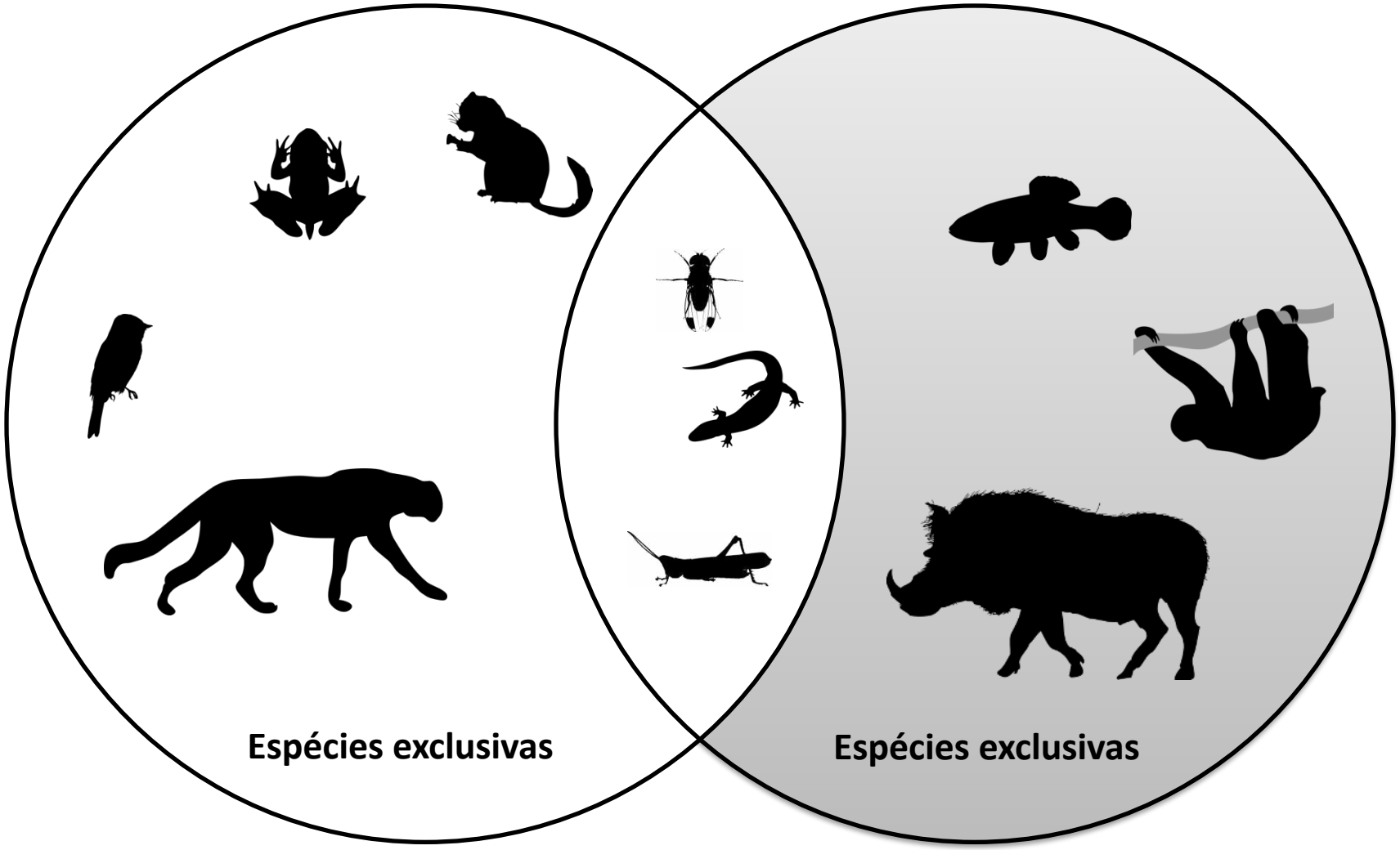


Slide T. Gonçalves-Souza

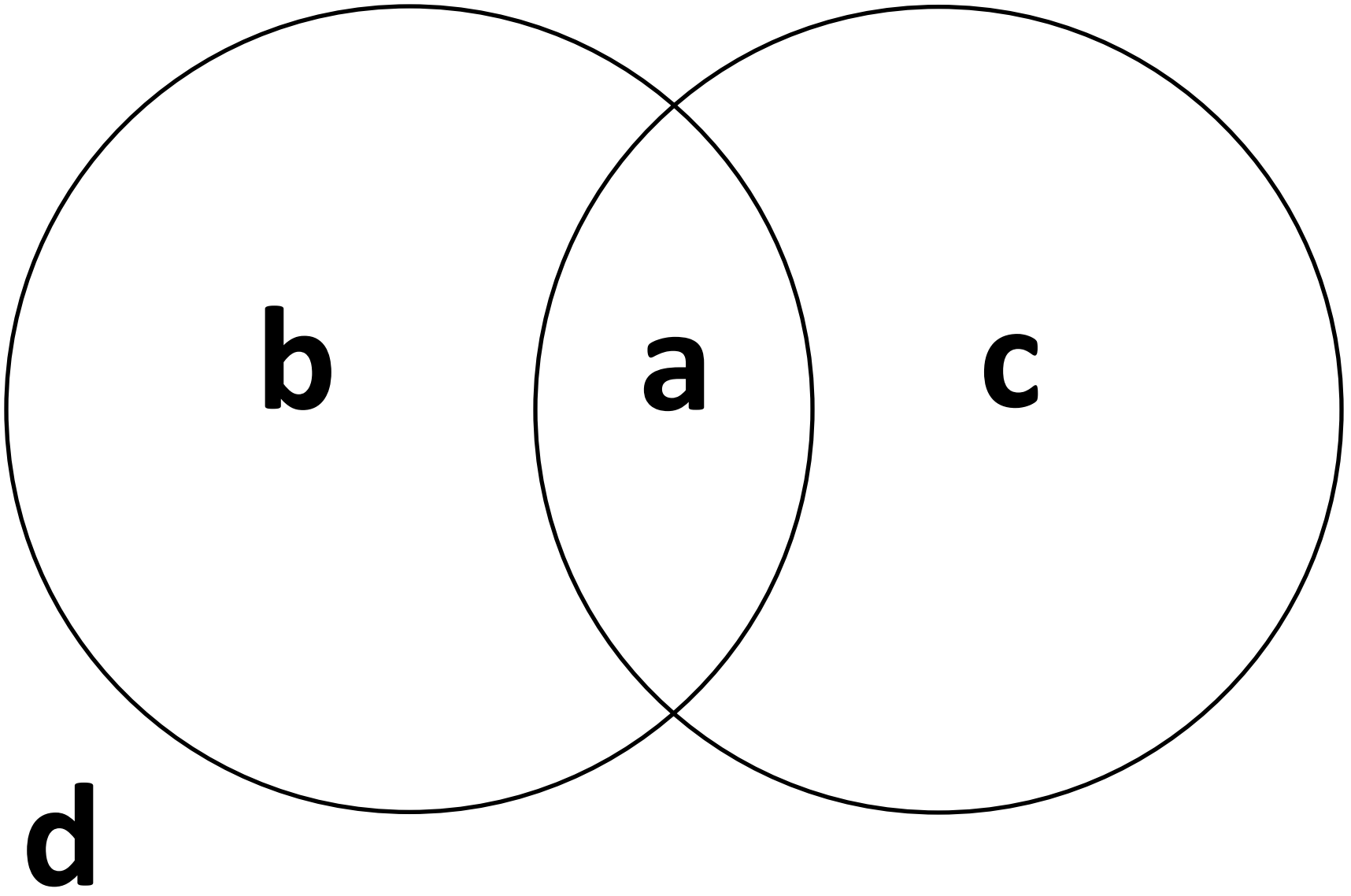
Comunidade 2



Slide T. Gonçalves-Souza



Slide T. Gonçalves-Souza



Slide T. Gonçalves-Souza

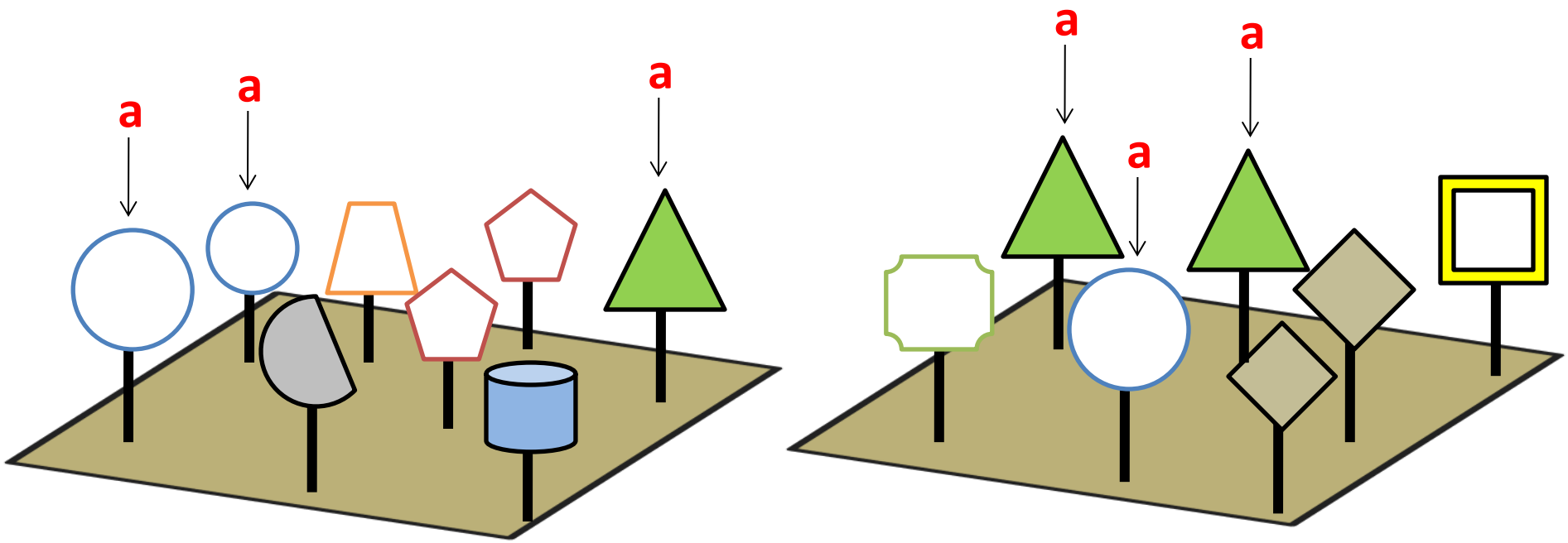
Comunidade 1

Comunidade 2

	1	0
1	a	b
0	c	d

Slide T. Gonçalves-Souza

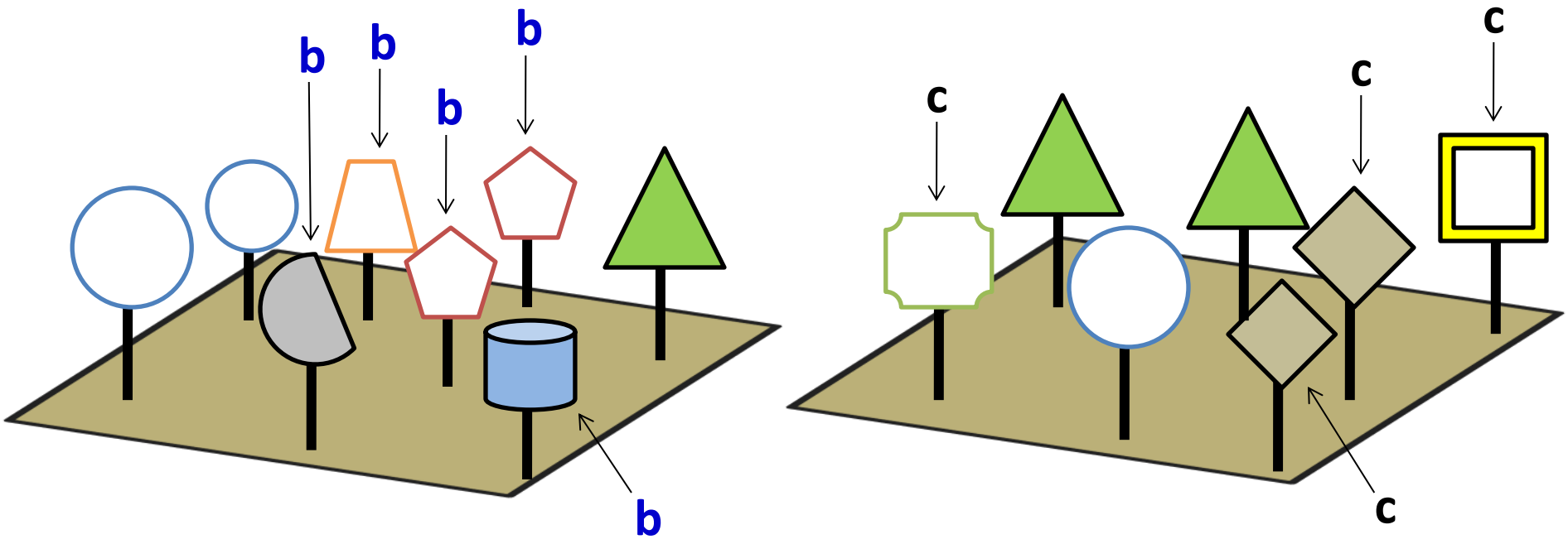
Índices binários de espécies (Q mode)



Jaccard (β_j) = a / a+b+c

Slide T. Gonçalves-Souza

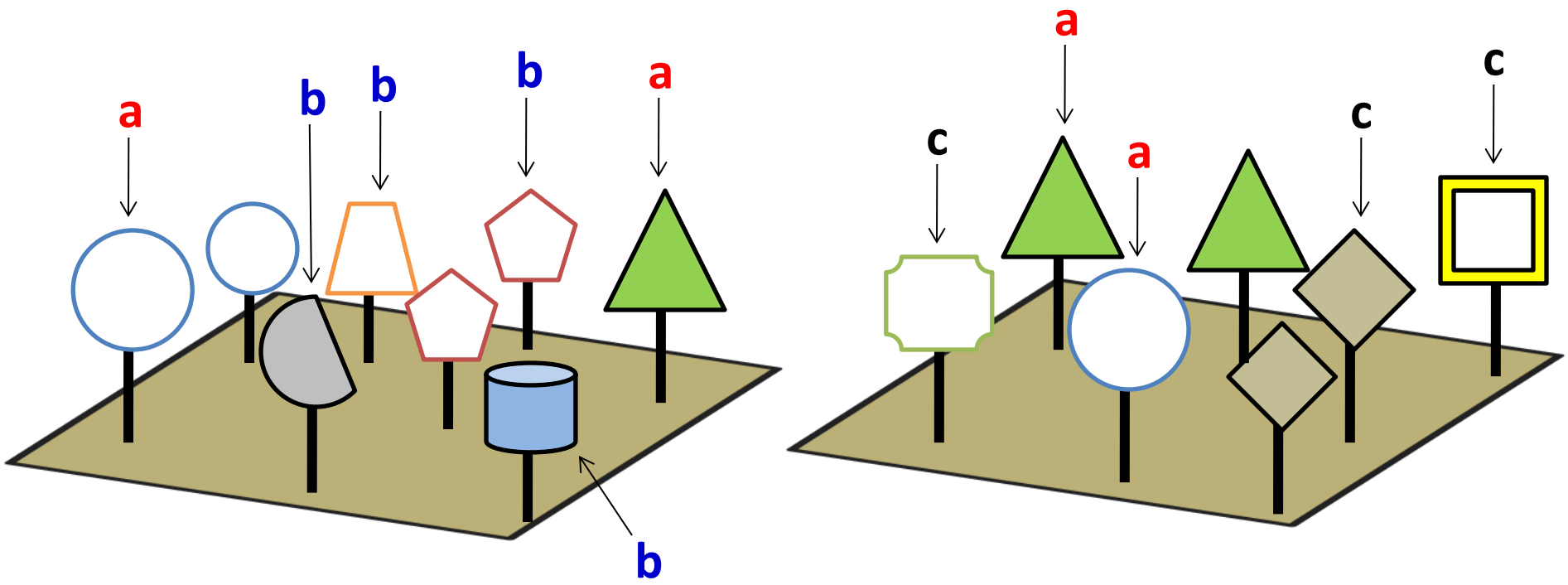
Índices binários



Jaccard (β_j) = a / a+b+c

Slide T. Gonçalves-Souza

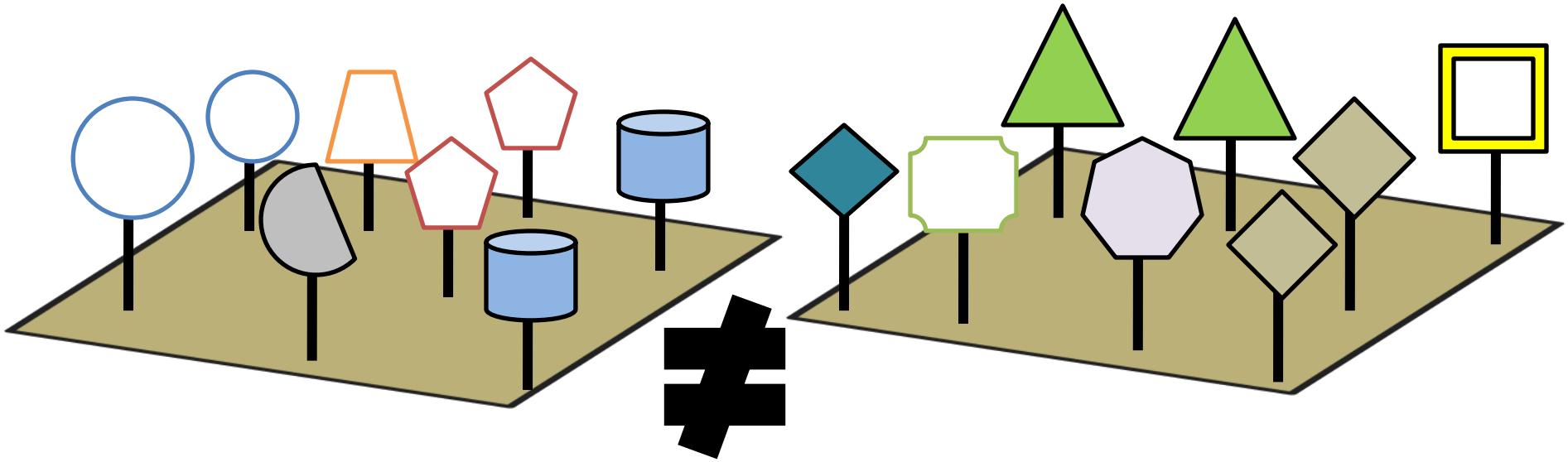
Índices binários



$$\beta_j = a / a+b+c = 2 / 2 + 3 + 3 = 2/9 = 0{,}222$$

Slide T. Gonçalves-Souza

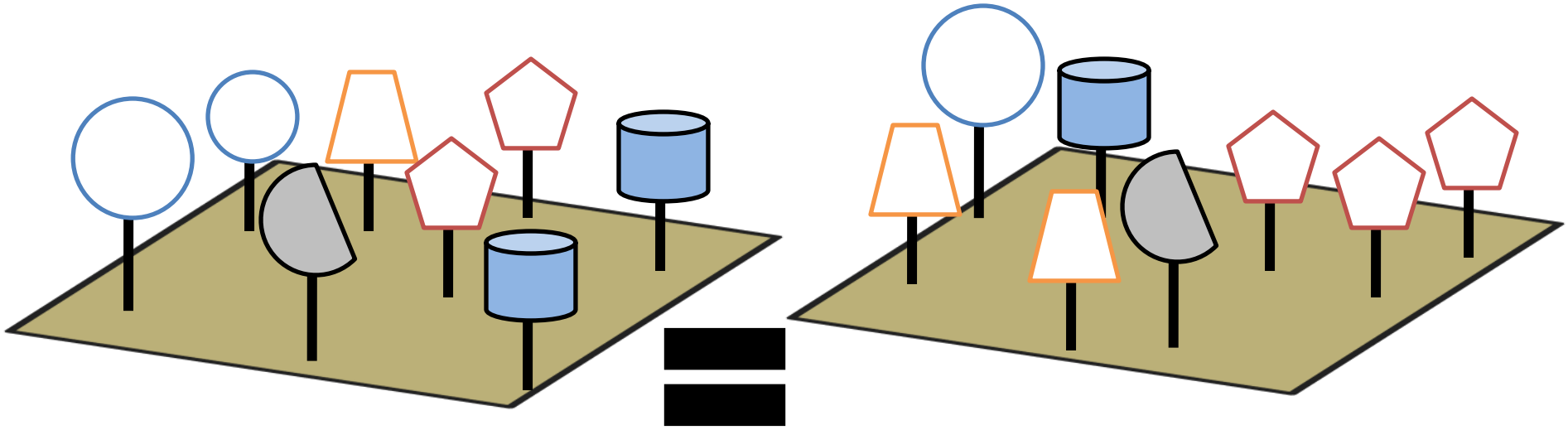
Comunidades sem espécies compartilhadas



β_j = a / a+b+c = 0 / 0 + 5 + 6 = 0/11 = 0

Slide T. Gonçalves-Souza

Comunidades compartilham todas espécies



β_j = a / a+b+c = 5 / 5 + 0 + 0 = 5/5 = 1

Slide T. Gonçalves-Souza

Sørensen

$$S_8(\mathbf{x}_1, \mathbf{x}_2) = \frac{2a}{2a + b + c}$$

Bray-Curtis

$$S_{17}(\mathbf{x}_1, \mathbf{x}_2) = \frac{W}{(A + B)/2} = \frac{2W}{(A + B)}$$

Bray-Curtis é a versão quantitativa do Sørensen

Dados binários de descritores (R mode)

- Presença-ausência de características físicas em locais

- Geralmente o coeficiente de Sokal & Michener (S_1 de L&L) é a melhor escolha

Coeficientes quantitativos para objetos (Q mode)

Coeficientes semi-métricos assimétricos

- Usados para dados quantitativos, e.g., abundância, % cobertura, biomassa etc

- Bray-Curtis (percentage difference), Chord, log-Chord, Hellinger, chi-quadrado, Morisita-Horn

Table 6.2. Reasonable and acceptable domains of input data, x, and ranges of distance measures, $d = f(x)$.

Name (synonyms)	Domain of x	Range of $d = f(x)$	Comments
Sørensen (Bray & Curtis; Czekanowski)	$x \geq 0$	$0 \leq d \leq 1$ (or $0 \leq x \leq 100\%$)	proportion coefficient in city-block space; semimetric
Relative Sørensen (Kulczynski; Quantitative Symmetric)	$x \geq 0$	$0 \leq d \leq 1$ (or $0 \leq x \leq 100\%$)	proportion coefficient in city-block space; same as Sørensen but data points relativized by sample unit totals; semimetric
Jaccard	$x \geq 0$	$0 \leq d \leq 1$ (or $0 \leq d \leq 100\%$)	proportion coefficient in city-block space; metric
Euclidean (Pythagorean)	all	non-negative	metric
Relative Euclidean (Chord distance; standardized Euclidean)	all	$0 \leq d \leq \sqrt{2}$ for quarter hypersphere; $0 \leq d \leq 2$ for full hypersphere	Euclidean distance between points on unit hypersphere; metric
Correlation distance	all	$0 \leq d \leq 1$	converted from correlation to distance; proportional to arc distance between points on unit hypersphere; cosine of angle from centroid to points; metric
Chi-square	$x \geq 0$	$d \geq 0$	Euclidean but doubly weighted by variable and sample unit totals; metric
Squared Euclidean	all	$d \geq 0$	metric
Mahalanobis	all	$d \geq 0$	distance between groups weighted by within-group dispersion; metric

McCune & Grace 2002

Todos implementados no vegan::vegdist, ade4::dist.ktab ou adespatial::dist.ldc

Coeficientes para descritores (R mode) que incluem mistura de tipos de dados

Coeficiente de Gower (1971)

- Pode lidar com misturas de dados (descritores)
  - Com diferentes unidades de medidas
  - Contínuo, binário, multiestado etc
  - E até dados faltantes!
- Pavoine et al. (2009) Oikos expandiram para mais tipos de dados
  - Dados circulares (fenológicos), fuzzy (=pertencimento difuso), e **ordinal**

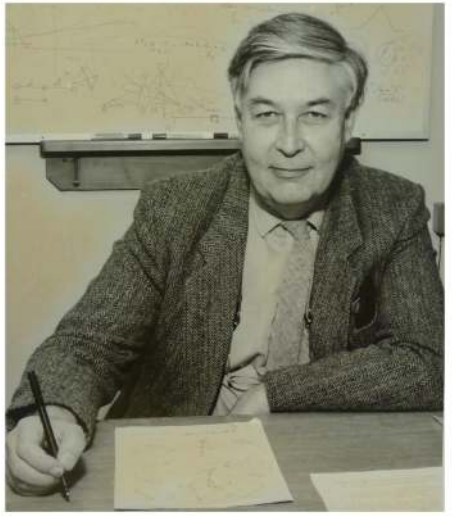


Figure 1. *John Gower, circa 1985.*

The general form of the coefficient is the following:

$$D_{Gower}(x_1, x_2) = 1 - \left(\frac{1}{p} \sum_{j=1}^{p} s_j(x_1, x_2) \right)$$

Where $s_j(x_1, x_2)$ is a partial similarity function computed separately for each descriptor.

- For quantitative descriptors, $s_j(x_1, x_2)$ is computed as follows:

$$s_j(x_1, x_2) = 1 - \frac{|y_{1j} - y_{2j}|}{R_j}$$

$|y_{1j} - y_{2j}|$: absolute difference between two values for a single variable

R_j: maximum difference value (range)

- For qualitative (factors) or binary descriptors:

$s_j(x_1, x_2)$ is 1 if the two objects have the same state; otherwise 0.

Slide P. Legendre

Padronizações e transformações

Padronização

- Quando temos de usar?
- Descritores são medidos em escalas diferentes
- Exemplo: temperatura (ºC), distância da margem (m), área (m^2)
- Z-score

$$z_i = \frac{y_i - \bar{y}}{s_y}$$

- Todas as variáveis passam a ter distribuição *Z*, com média 0 e variância 1

Cap. 1 L&L

Padronização das variáveis ambientais

Matriz não-padronizada

Local	pH	T	Umid	Var *m*
Local 1	8.2	27	90	m1
Local 2	6.3	32	93	m2
Local 3	5.3	36	98	m3
Local 4	5.6	31	89	m4
Local 5	6.6	19	96	m5
Local 6	6.1	15	75	m6
Média	**6.4**	**26.7**	**90.2**	
DP	**1**	**8.1**	**8.2**	

pH: 0 ~ 14

T: 0 ~ 42 ºC

Umid: 0 ~ 100

Padronização das variáveis ambientais

Matriz não-padronizada

Local	pH	T	Umid	Var *m*
Local 1	8.2	27	90	m1
Local 2	6.3	32	93	m2
Local 3	5.3	36	98	m3
Local 4	5.6	31	89	m4
Local 5	6.6	19	96	m5
Local 6	6.1	15	75	m6
Média	**6.4**	**26.7**	**90.2**	
DP	**1**	**8.1**	**8.2**	

Matriz padronizada

Local	pH
Local 1	8.2 - 6.4 / 1
Local 2	
Local 3	
Local 4	
Local 5	
Local 6	
Média	
DP	

Padronização das variáveis ambientais

Matriz não-padronizada

Local	pH	T	Umid	Var *m*
Local 1	8.2	27	90	m1
Local 2	6.3	32	93	m2
Local 3	5.3	36	98	m3
Local 4	5.6	31	89	m4
Local 5	6.6	19	96	m5
Local 6	6.1	15	75	m6
Média	**6.4**	**26.7**	**90.2**	
DP	**1**	**8.1**	**8.2**	

Matriz padronizada

Local	pH
Local 1	8.2 - 6.4 / 1
Local 2	6.3 – 6.4 / 1
Local 3	5.3 – 6.4 / 1
Local 4	5.6 – 6.4 / 1
Local 5	6.6 – 6.4 / 1
Local 6	6.1 – 6.4 / 1
Média	
DP	

Legendre et al. 2011: "menor probabilidade de cometer erro do tipo I utilizando matrizes padronizadas".

Table 9.4. Suggested procedure for data adjustments of quantitative variables in environmental data matrices.

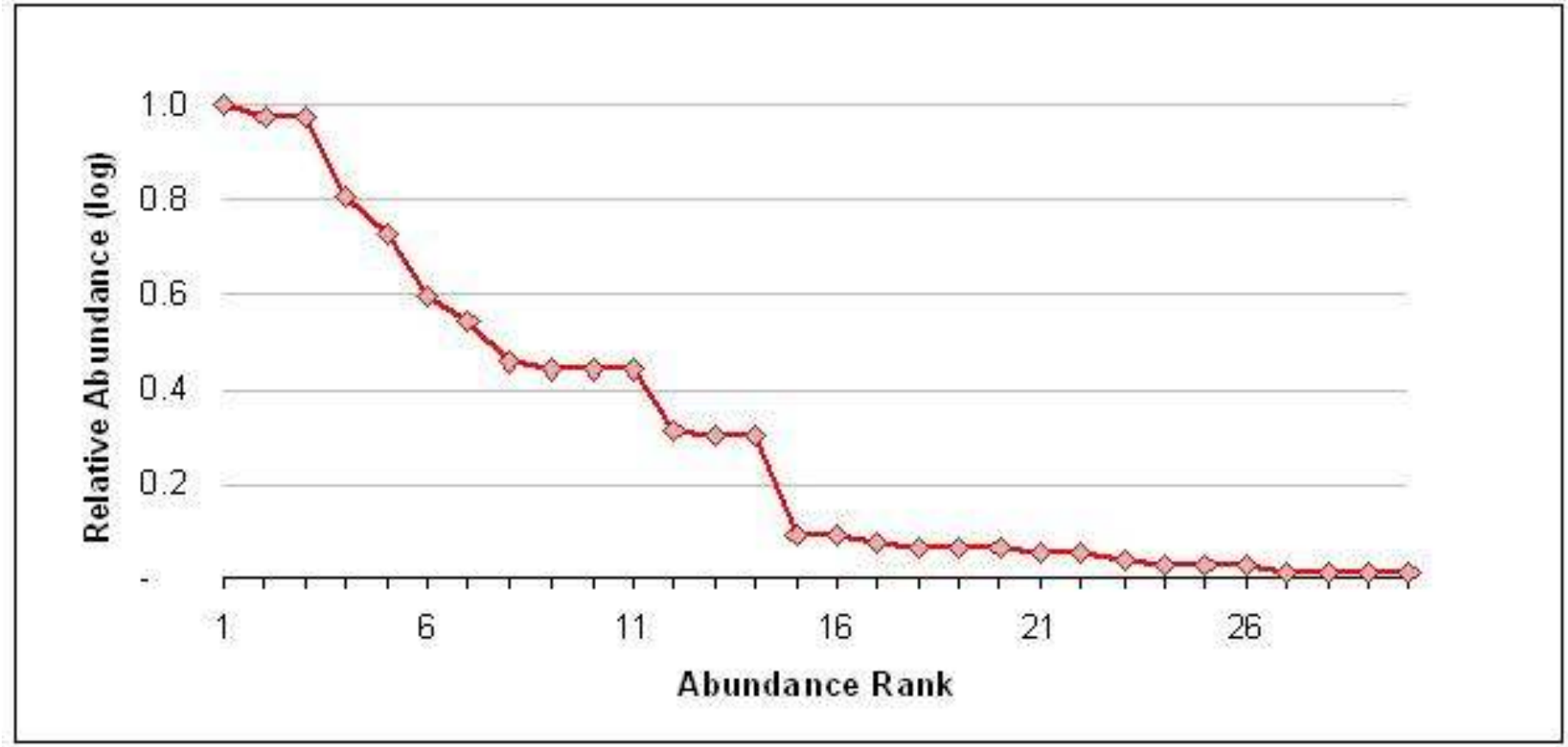
Action to be considered	Criteria
1. Calculate descriptive statistics for quantitative variables. Repeat this after each step below. (In PC-ORD run *Row & column summary*)	Always
Skewness and range for each variable (column)	
2. Monotonic transformation (applied to individual variables, depending on need)	Consider log or square root transformation for variables with skewness > 1 or ranging over several orders of magnitude.
	Consider arcsine squareroot transformation for proportion data.
3. Column relativizations	Consider column relativization (by norm or standard deviates) if environmental variables are to be used in a distance-based analysis that does not automatically relativize the variables (for example, using MRPP to answer the question: do groups of sample units defined by species differ in environmental space?). Column relativization is not necessary for analyses that use the variables one at a time (e.g., ordination overlays) or for analyses with built-in standardization (e.g., PCA of a correlation matrix).
4. Check for univariate outliers and take corrective steps if necessary.	Examine scatterplots or frequency distributions or relativize by standard deviates ("*z* scores") and check for high absolute values.

McCune & Grace 2002

Transformações para matriz de espécies

- Transformação de Hellinger ou Chord para abundância de espécies

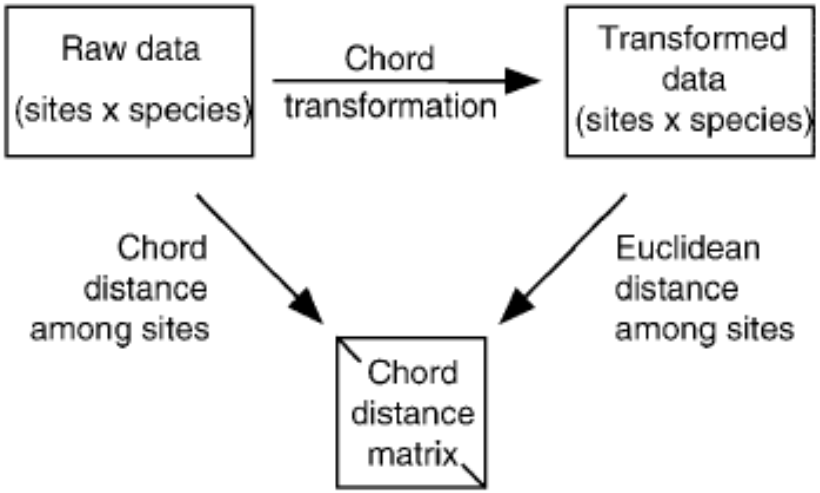
- Problema em ordenações com dados de abundância é que existem muitos zeros e diferenças muito grandes de abundância entre espécies



Oecologia (2001) 129:271–280
DOI 10.1007/s004420100716

Pierre Legendre · Eugene D. Gallagher

Ecologically meaningful transformations for ordination of species data



A idéia é que se faça uma transformação dos dados pela distância de Chord ou Hellinger e depois se faça uma ordenação (e.g., PCA, PCoA, RDA) de maneira a satisfazer os critérios da metricidade

Voltaremos a falar sobre isso na aula de ordenações restritas

ECOGRAPHY

Research

Box–Cox-chord transformations for community composition data prior to beta diversity analysis

Pierre Legendre and Daniel Borcard

P. Legendre (http://orcid.org/0000-0002-3838-3305) (pierre.legendre@umontreal.ca) and D. Borcard, Dépt de sciences biologiques, Univ. de Montréal, Montréal, QC, Canada.

Ecography
41: 1820–1824, 2018
doi: 10.1111/ecog.03498
Subject Editor: Luis Mauricio Bini

In studies of spatial or temporal beta diversity, community composition data, often containing many zeros, must be transformed in some way before they are analysed by multivariate methods of data analysis. Data are transformed to reduce the skewness of species distributions and make dissimilarities double-zero asymmetrical. Criteria have recently been proposed to determine which dissimilarity functions (or the corresponding data

- As distâncias de chord, Hellinger, e log-chord são parte de uma série de transformações normalizantes equivalentes à de Box-Cox
- ƛ = 1 => Chord
- ƛ = 0.5 => Hellinger
- ƛ = 0 => log-chord
- Permite a normalização de frequências de distribuição cada vez mais assimétricas num único arcabouço matemático
- Autores fornecem código no R

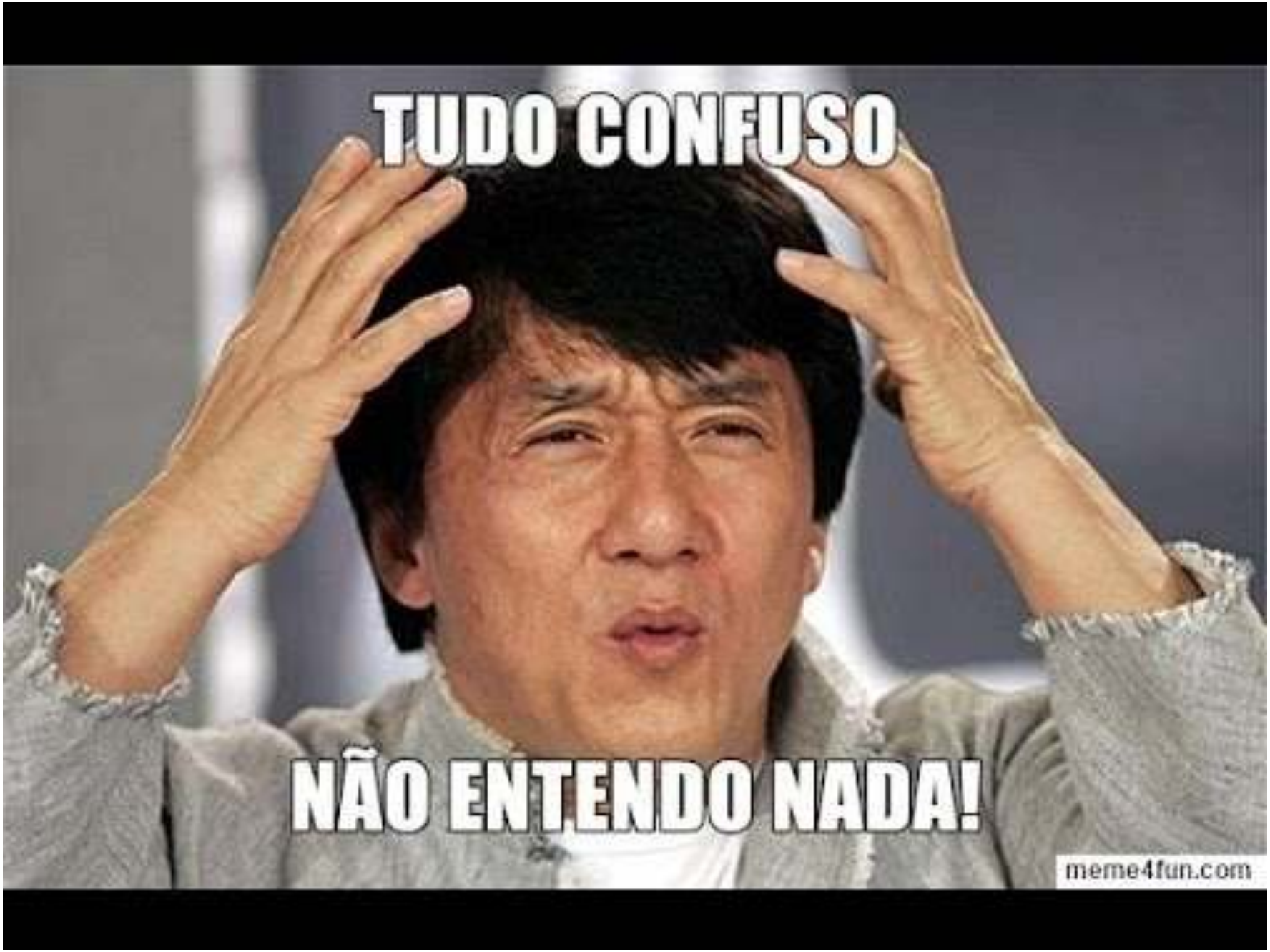

TUDO CONFUSO
NÃO ENTENDO NADA!
meme4fun.com

Table 9.3. Suggested procedure for data adjustments of species data matrices.

Action to be considered	Criteria
1. Calculate descriptive statistics. Repeat this after each step below. (In PC-ORD run *Row & column summary*) Beta diversity (community data sets) Average skewness of columns Coefficient of variation (CV, %) CV of row totals CV of column totals	Always
2. Delete rare species (< 5% of sample units)	Usually applied to community data sets, unless contrary to study goals
3. Monotonic transformation (if applied to species, then usually applied uniformly to all of them, so that all are scaled the same)	A. Average skewness of columns (species) B. Data range over how many orders of magnitude? (Count and biomass data often are extreme.) C. Beta diversity. (Consider presence/absence transformation for community data when β is high.)
4. Row or column relativizations	What is the question? Are units for all variables the same? Is relativization built into the subsequent analysis? CV of row totals CV of column totals What distance measure do you intend to use? Note: regardless of your decision to relativize or not, you should state your decision and justify it briefly on biological grounds.
5. Check for outliers based on the average distance of each point from all other points. Calculate standard deviation of these average distances. Describe outliers and take steps to reduce influence, if necessary	standard deviation / degree of problem < 2 — no problem 2 - 2.3 — weak outlier 2.3 - 3 — moderate outlier >3 — strong outlier

McCune & Grace 2002

Efeito de diferentes coeficientes na representação de objetos no espaço reduzido

The following numerical example, from Orlóci (1978: 59), shows that D_{14} does not obey the triangle inequality theorem and is thus not a metric distance:

3 sítios com ≠ abund de sp

Quadrats	Species				
	y_1	y_2	y_3	y_4	y_5
x_1	2	5	2	5	3
x_2	3	5	2	4	3
x_3	9	1	1	1	1

The distances between the three pairs of sites are:

$$D_{14}(x_1, x_2) = \frac{1+0+0+1+0}{17+17} = 0.059$$

$$D_{14}(x_1, x_3) = \frac{7+4+1+4+2}{17+13} = 0.600$$

$$D_{14}(x_2, x_3) = \frac{6+4+1+3+2}{17+13} = 0.533$$

hence $0.059 + 0.533 < 0.600$, which violates the triangle inequality theorem. Coefficient D_{14} is thus not a metric distance. Table 7.3 shows that D_{14}, which is equal

Cap. 7 Legendre & Legendre 2012

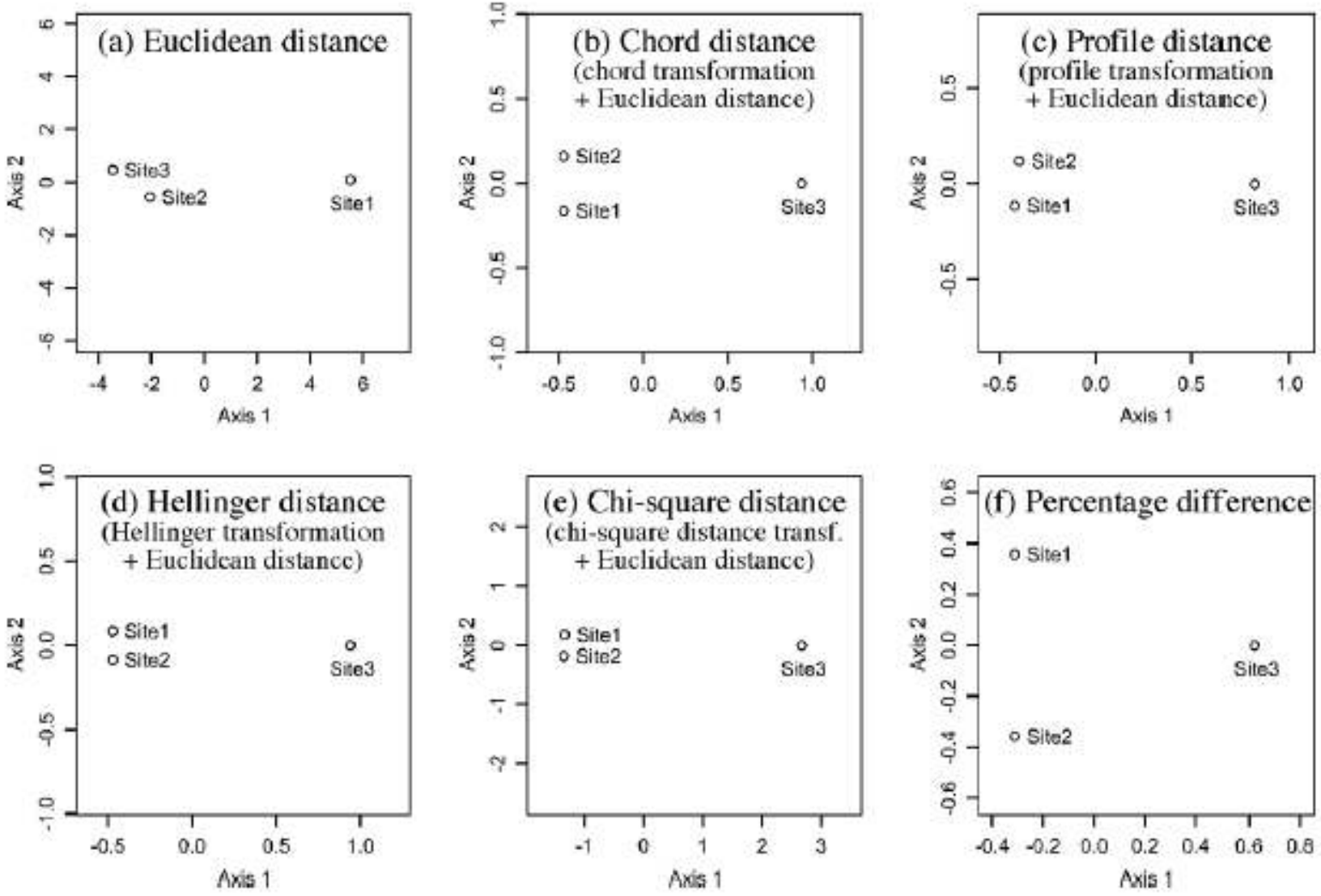


Figure 7.9 Principal coordinate ordination plots (PCoA, Section 9.3) of the distance matrices computed in Fig. 7.8: (a) D_1, (b) D_3, (c) D_{18}, (d) D_{17}, (e) D_{16}, and (f) a PCoA plot of the percentage difference (Steinhaus/Odum/Bray-Curtis) distance matrix (D_{14}) computed for the same data.

Uma chave dicotômica para escolher o coeficiente? Sim, ela existe!

Veja também item 7.6 L&L

Table 7.4 Choice of an association measure among objects (Q mode), to be used with species descriptors (asymmetrical coefficients). For explanation of levels 4 and 6, see the accompanying text.

1) Descriptors: presence-absence or ordered classes on a scale of relative abundances (no partial similarities computed between classes) **see 2**

   2) Metric coefficients: *coefficient of community* (S_7) and variants (S_{10}, S_{11})

   2) Semimetric coefficients: variants of the coef. community (S_8, S_9, S_{13}, S_{14})

   2) Nonmetric coefficient: Kulczynski (S_{12}) (non-linear: not recommended)

   2) Probabilistic coefficient: S_{27}

1) Descriptors: quantitative or semiquantitative (states defined in such a way that partial similarities can be computed between them) **see 3**

   3) Coefficients for raw or normalized abundance data **see 4**

      4) No standardization by object; the same difference for either abundant or rare species, contributes equally to the similarity between sites: *coefficients of Steinhaus* (S_{17}) and *Kulczynski* (S_{18}), *percentage difference* (D_{14}), $\sqrt{D_{14}}$

      4) Standardization by object-vector; if objects are of equal importance*, same contributions for abundant or rare species to the similarity or distance between sites: *chord distance* (D_3), *geodesic metric* (D_4), *index of association* (D_9), *Hellinger dist.* (D_{17}), *dist. between profiles* (D_{18})

      4) Standardization by object-vector*; differences for abundant species (in the whole data set) contribute more than differences between rare species to the similarity (less to the distance) between sites: χ^2 *similarity* (S_{21}), χ^2 *metric* (D_{15}), χ^2 *distance* (D_{16})

   3) Limited to normalized abundances (species distributions not strongly skewed). [Normalization of species abundance data: Sections 1.5.6 and 7.7] **see 5**

      5) Coefficients without associated probability levels **see 6**

         6) Differences for abundant species (for two sites under consideration) contribute more than differences between rare species to the similarity (less to the distance) between sites: *Canberra metric* (D_{10}), *coefficient of divergence* (D_{11}). Both have low resolution: not recommended for clustering

         6) Differences for abundant species (in the whole data set) contribute more than differences between rare species to the similarity (less to the distance) between sites: *asymmetrical Gower coefficient* (S_{19}), *coefficient of Legendre & Chodorowski* (S_{20})

         6) Differences for abundant and rare species contribute the same to the distance between sites: *modified mean character difference* or *modified Gower dissimilarity* (D_{19})

      5) Probabilistic coefficient: *Goodall coefficient* (S_{23})

L&L 2012

Table 7.5 Choice of an association measure among objects (Q mode), to be used with chemical, geological physical, etc. descriptors (symmetrical coefficients, using double-zeros).

1) Association measured between individual objects **see 2**

2) Descriptors: presence-absence or multistate (no partial similarities computed between states) **see 3**

3) Metric coefficients: *simple matching* (S_1) and derived coefficients (S_2, S_6)

3) Semimetric coefficients: S_3, S_5

3) Nonmetric coefficient: S_4

2) Descriptors: multistate (states defined in such a way that partial similarities can be computed between them) **see 4**

4) Descriptors: quantitative and dimensionally homogeneous **see 5**

5) Differences enhanced by squaring: *Euclidean distance* (D_1) and *average distance* (D_2)

5) Differences mitigated: *Manhattan metric* (D_7), *mean character difference* (D_8)

4) Descriptors: not dimensionally homogeneous; weights (equal or not, according to values w_j used) given to each descriptor in the computation of association measures **see 6**

6) Descriptors are qualitative (no partial similarities computed between states) and quantitative (partial similarities based on the range of variation of each descriptor): *symmetrical Gower coefficient* (S_{15})

6) Descriptors are qualitative (possibility of using matrices of partial similarities between states) and semiquantitative or quantitative (partial similarity function for each descriptor): *coefficient of Estabrook & Rogers* (S_{16})

1) Association measured between groups of objects

7) Removing the effect of correlations among descriptors: *Mahalanobis generalized distance* (D_5)

7) Not removing the effect of correlations among descriptors: *coefficient of racial likeness* (D_{12})

Table 7.6 Choice of a dependence measure among descriptors (R mode).

1) Descriptors: species abundances **see 2**

  2) Descriptors: presence-absence **see 3**

    3) Coefficients without associated probability levels: S_7, S_8, S_{14}, S_{24}

    3) Probabilistic coefficient: S_{25}

  2) Descriptors: multistate

    4) Data are raw abundances: χ^2 *similarity* (S_{21}), χ^2 *metric* (D_{15}), χ^2 *distance* (D_{16}), *Whittaker's SC* (D_{20}) **see 4**

    4) Data are abundances in linear or monotonic relationships **see 5**

      5) Coefficients without associated probabilities: *covariance*, *Pearson r*, *Spearman r*, Pearson or Spearman correlations among chord-transformed or Hellinger-transformed data

      5) Probabilistic coefficients: *probabilities associated to Pearson r* or *Spearman r*, *Goodall coefficient* (S_{23})

1) Descriptors: chemical, geological, physical, etc. **see 6**

  6) Coefficients without associated probability levels **see 7**

    7) Descriptors are quantitative and linearly related: *covariance*, *Pearson r*

    7) Descriptors are ordered and monotonically related: *Spearman r*, *Kendall* τ

    7) Descriptors are qualitative or ordered but not monotonically related: χ^2, *reciprocal information coefficient*, *symmetric uncertainty coefficient*

  6) Probabilistic coefficients **see 8**

    8) Descriptors are quantitative and linearly related: *probabilities associated to Pearson r*

    8) Descriptors are ordered and monotonically related: *probabilities associated to Spearman r* or *Kendall* τ

    8) Descriptors are qualitative or ordered but not monotonically related: *probabilities associated to* χ^2