

SUS-5020

**Técnicas de Modelagem para
Estudos Ambientais**

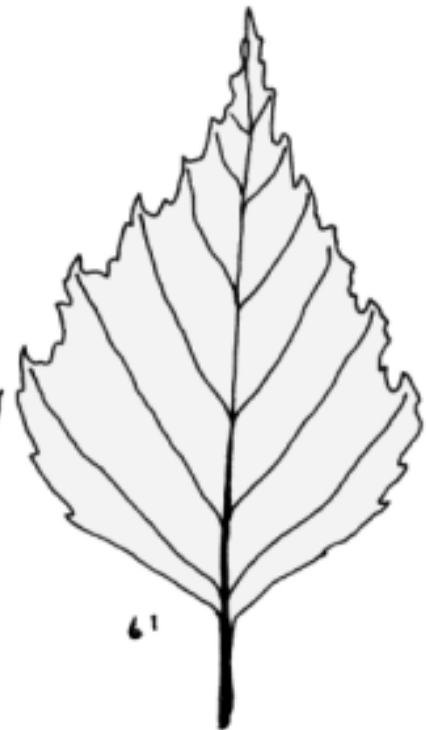
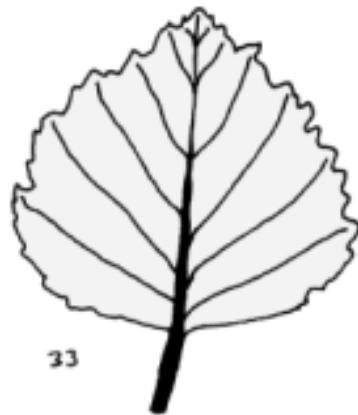
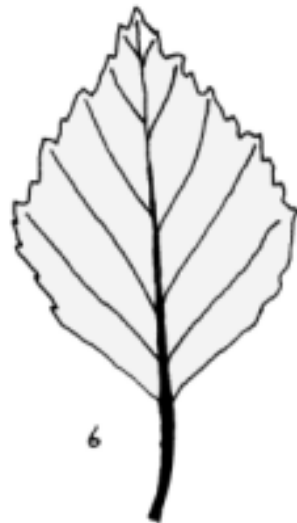
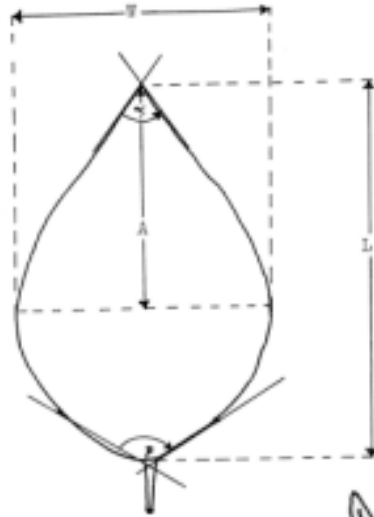
**Escolhendo
variáveis/descriptores/caracteres
Coeficientes - Distâncias
Transformação**

Fácil de observar ?

PIRATAS DO TIETÊ - Laerte



Objetivo ?



Homologia

How To Tell The Birds From The Flowers.

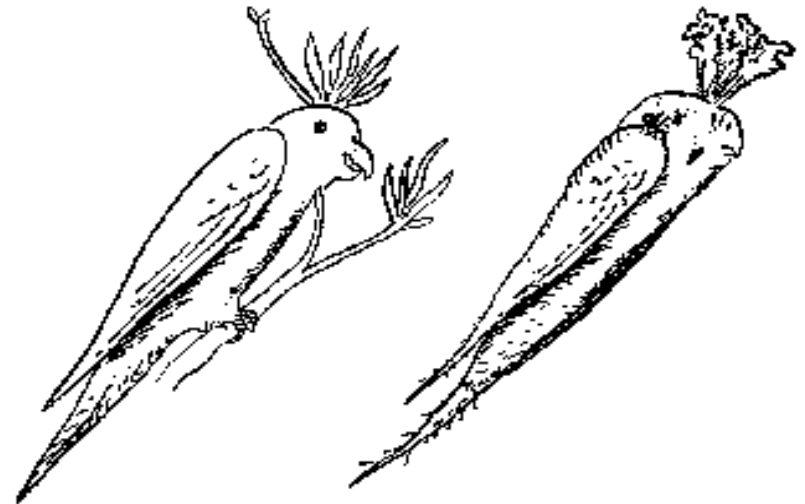
A Manual of Flornithology
for Beginners.



—• Verses and Illustrations •—

By Robert Williams Wood.

Published by Paul Elder and Company
San Francisco and New York.



The Parrot. The Carrot.

The Parrot and the Carrot we may
easily confound,
They're very much alike in looks
and similar in sound,
We recognize the Parrot by his
clear articulation,
For Carrots are unable to engage
in conversation.

H o m o i o g i a



Outras considerações

- Caracteres logicamente redundantes
- Caracteres invariantes
- Quantos caracteres ?
- Caracteres ponderados ?

modo "Q" e "R"

■ modo Q

- estamos interessados em comparações entre **objetos** (amostras)
- "espaço **A**"
- Jaccard
- Sorenson
- distância euclidiana

■ modo R

- estamos interessados em relações entre **descritores** (caracteres, espécies)
- "espaço **I**"
- correlação (Pearson)
- distância qui-quadrado

Coeficientes e distâncias

		UTO i		
		+	-	
UT O j	+	a	b	$a + b$
	-	c	d	$c + d$
		$a + c$	$b + d$	$p = a + b + c + d$

Cross-classification ~
classificação cruzada

$d = \text{zero duplo!}$

Coeficientes e distâncias

		UTO i	
		+	-
UTO j	+	a	b
	-	c	d

$$S_{SM} = \frac{(a + d)}{N}$$

(N = a + b + c + d)

■ Concordância Simples

Simétrico Binário –
Qualitativo - zero-duplo!
Modo Q

Coeficientes e distâncias

		UTO i	
		+	-
UTO j	+	a	b
	-	c	d

$$S_Y = \frac{ad - bc}{ad + bc}$$

■ Yule

$$\chi_{ij}^2 = \frac{N(ad - bc)^2}{(a + b)(c + d)(a + c)(b + d)}$$

■ Chi - quadrado

$$S_{RT} = \frac{(a + d)}{(a + d) + 2(b + c)}$$

■ Rogers-Tanimoto

Simétrico Binário –
Qualitativo - zero-duplo!
Modo Q

O problema do zero duplo!

Ecologia x Taxonomia

Percepção do universo amostral

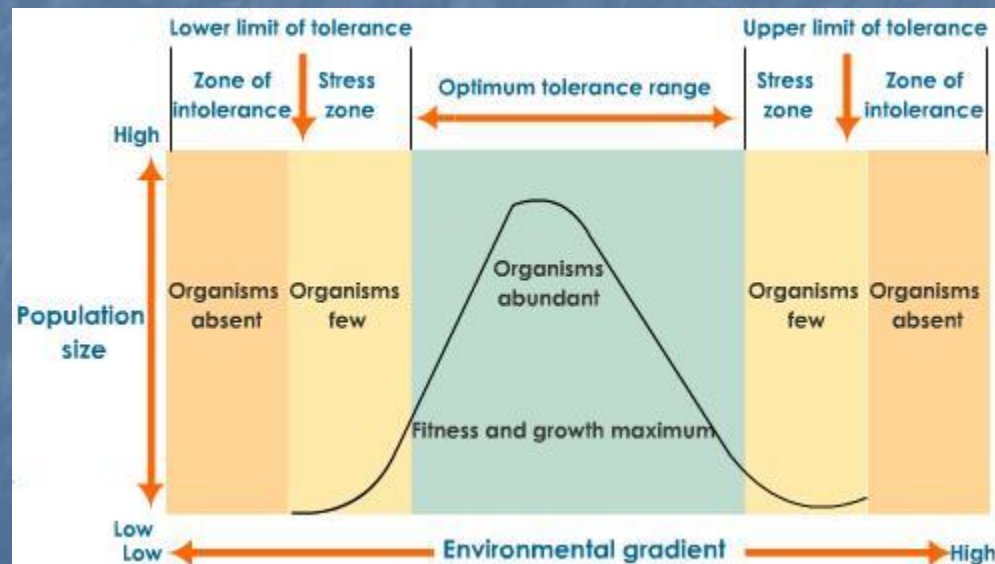


Figure 9.7 An environmental gradient, such as temperature, can determine the survival of an organism across a range of conditions. The organism will achieve its maximum population where it experiences an optimal environmental gradient. (Modified from C.B. Cox, I.N. Healey, and P.B. Moore, 1976.)

Coeficientes e distâncias

		UTO i	
		+	-
UTO j	+	a	b
	-	c	d

$$S_{JAC} = \frac{a}{(a + b + c)}$$

$$S_{SOR} = \frac{2a}{(2a + b + c)}$$

■ Jaccard

Assimétrico Binário –
Qualitativo - zero-duplo!
Modo Q

■ Dice/Sorenson/ Czekanowski

Outros coeficientes

■ Kulczynski

Não-métrico

$$S_{KUL} = \frac{1}{2} \left[\frac{a}{(a+b)} + \frac{a}{(a+c)} \right]$$

Média dos somatórios das abundâncias mínimas pelo total da abundância em cada site

■ Ochiai

Assimétrico qualitativo

$$S_{OCH} = \frac{a}{\sqrt{[(a+b)(a+c)]}}$$

Média geométrica das razões do total de espécies em cada site

■ Steinhaus

$$S_{St} = \frac{2 \sum_{k=1}^M \min(x_{ik}, x_{jk})}{\left(\sum_{k=1}^M x_{ik} + \sum_{k=1}^M x_{jk} \right)}$$

Compara os sites em termos da abundância mínima de cada espécie pelo somatório das abundâncias de todas as espécies em cada site

■ = S_{sor} para binários

Assimétricos quantitativos (Kulczynski e Steinhaus) – adequados para dados de abundância de espécies

Ex: Coeficiente de Steinhaus

	Abundância das Espécies	A	B	W
Site x ₁	7 3 0 5 0 1	16		
Site x ₂	2 4 7 6 0 3		22	
Mínimo	2 3 0 5 0 1			11

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

$$S_{17}(\mathbf{x}_1, \mathbf{x}_2) = \frac{2 \times 11}{16 + 22} = 0.579$$

Coeficientes para dados mistos

■ Gower = 

$$S_{Gow} = \frac{\sum_1^p S_k w_k}{\sum_1^p w_k}$$

- w_k é o peso para cada comparação.
Para cada variável, k , se X_{ik} ou X_{jk} está faltando, então $w_k = 0.0$.

- para variáveis quantitativas :

$$\rightarrow S_k = 1 - \frac{(X_{ik} - X_{jk})}{|X_{\max k} - X_{\min k}|} \quad w_k = 1.0$$

- para variáveis multiestado $S_k = 1.0$
se $X_{ik} = X_{jk}$, $w_k = 1.0$

- para variáveis binárias

$$S_k = 1.0 \text{ se } X_{ik} = X_{jk} = 1$$

$$w_k = 1.0 \text{ [ou 0 se ambos são 0]}$$

Coeficientes e distâncias

Distâncias métricas desenvolvidas para descritores quantitativos, ocasionalmente para semiquantitativos!

$$d_{ij}^2 = \sum_{k=1}^N (x_{ik} - x_{jk})^2$$

$$d_{ij} = \sqrt{d_{ij}^2}$$

■ Distância euclidiana

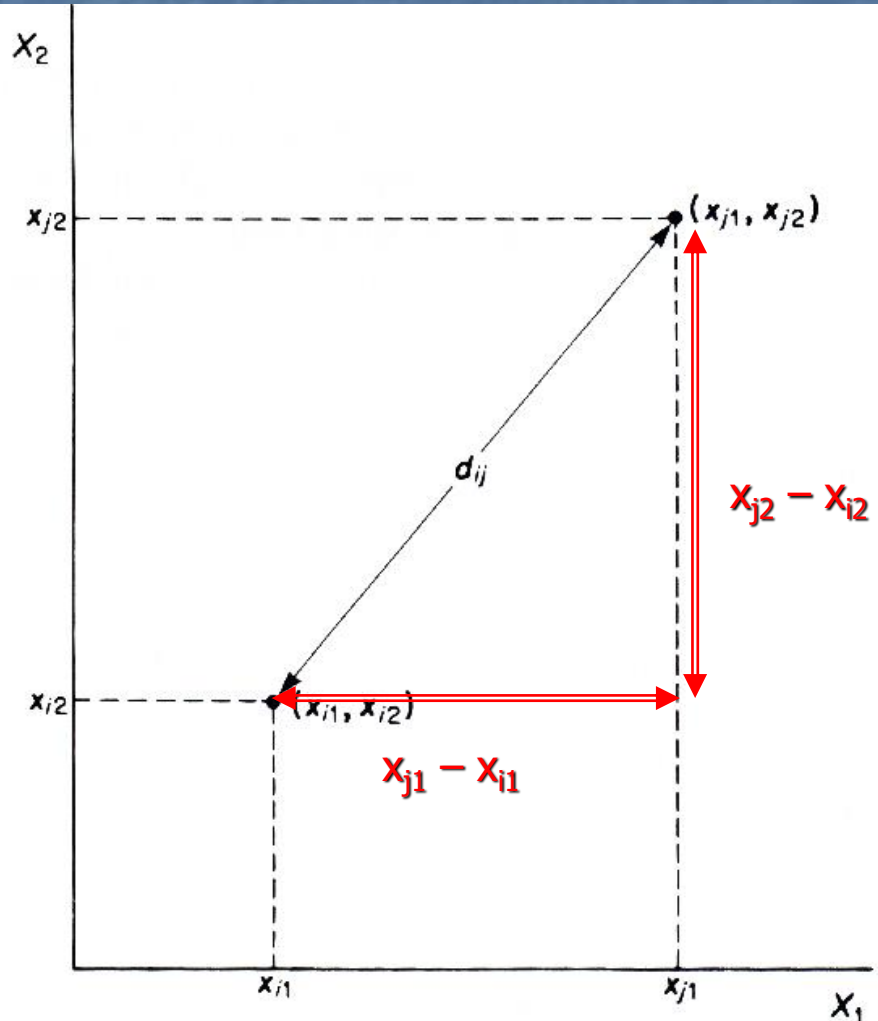


Figure 5.1 The Euclidean distance between objects i and j , with $p = 2$ variables.

Tamanho e forma

Diferença de tamanho

$$d_{ij} = \frac{1}{N^2} \left(\sum_{k=1}^N x_{ik} - \sum_{k=1}^N x_{jk} \right)^2$$

ou

$$d_{ij} = \frac{(b-c)^2}{N^2}$$

para dados binários

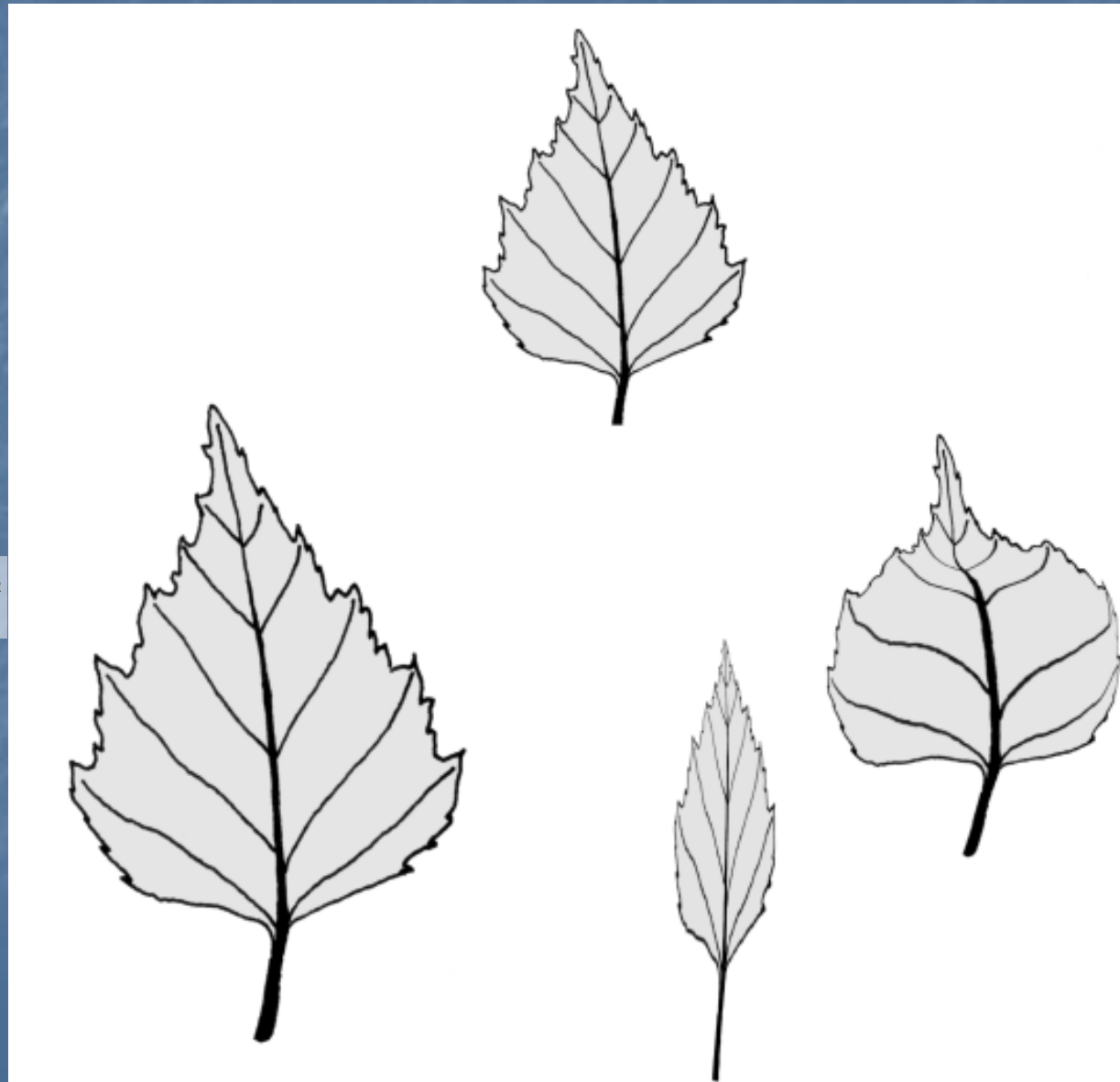
Diferença de forma

$$d_{ij} = \frac{1}{N} \sum_{k=1}^N (x_{ik} - x_{jk})^2 - \frac{1}{N^2} \left(\sum_{k=1}^N x_{ik} - \sum_{k=1}^N x_{jk} \right)^2$$

ou

$$d_{ij} = \frac{N(b+c) - (b-c)^2}{N^2}$$

para dados binários



Coeficientes e distâncias

Dados quantitativos - adequados para composição de comunidades

$$d_{Mij} = \sum_{k=1}^N |x_{ik} - x_{jk}|$$

■ Distância manhattan

A distância entre dois descritores de dois objetos é a soma da distância da abcissa do descritor y1 e da ordenada y2.

$$d_{BC} = \frac{\sum_{k=1}^N |x_{ik} - x_{jk}|}{\sum_{k=1}^N (x_{ik} + x_{jk})}$$

■ Distância Bray-Curtis

Os valores de similaridade não mudam com a inserção dos zeros-duplos, mas é uma distância semimétrica. Sua formulação utiliza W, A e B. Contribuição de espécies abundantes e raras similares

$$D = 1 - \frac{2W}{A + B}$$

$$d_{CANij} = \frac{1}{N} \sum_{k=1}^N \frac{|x_{ik} - x_{jk}|}{(x_{ik} + x_{jk})}$$

■ Distância Canberra

Variante da D. Manhattan que exclui os zeros duplos. Nela a diferença entre espécies abundantes contribui menos do que a abundância das espécies raras. É um coeficiente métrico.

■ Distância corda

- . Abordagem que destaca dados de composição de espécies
- . Calcula distância euclidiana depois que os sites são vetorizados para ter comprimento equivalente a 1
- . Destaca maior similaridade entre sites com o mesmo conjunto e proporcionalidade de abundância entre as espécies
- . Zero é o valor quando o conjunto de espécies e as proporções são as mesmas e $\sqrt{2}$ para sites que não compartilham nenhuma espécie

variáveis			
in	V1	V2	V3
I1	1,0	1,0	1,0
I2	5,0	5,0	5,0
I3	1,0	2,0	2,0

■ distância euclidiana simples

■	1	I1		
■	2	6,9282	I2	
■	3	1,4142	5,8310	I3
■		1	2	3

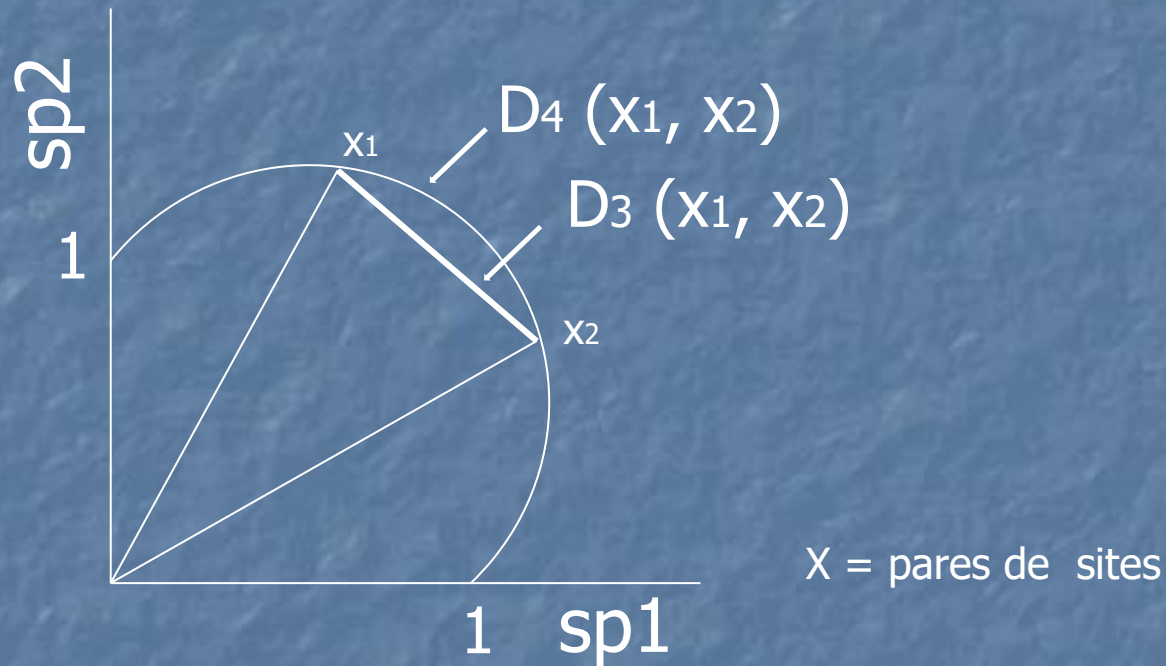
■ distância corda

■	1	I1		
■	2	0,0000	I2	
■	3	0,2748	0,2748	I3
■		1	2	3

Não muda com
zeros duplos

A distância corda é máxima quando espécies encontradas em dois sites são completamente diferentes

Dados centralizados



D_3 = distância corda com distância euclidina

D_4 = distância corda com métrica geodésica

Coeficientes e distâncias

- Correlação

$$r_{ij} = \frac{\sum (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)}{\sqrt{\sum (x_{ik} - \bar{x}_i)^2 \sum (x_{jk} - \bar{x}_j)^2}}$$

Propriedades de coeficientes e distâncias

Uma distância é métrica se :

1. É simétrica $d(x,y) = d(y,x)$

Uma distância é métrica se :

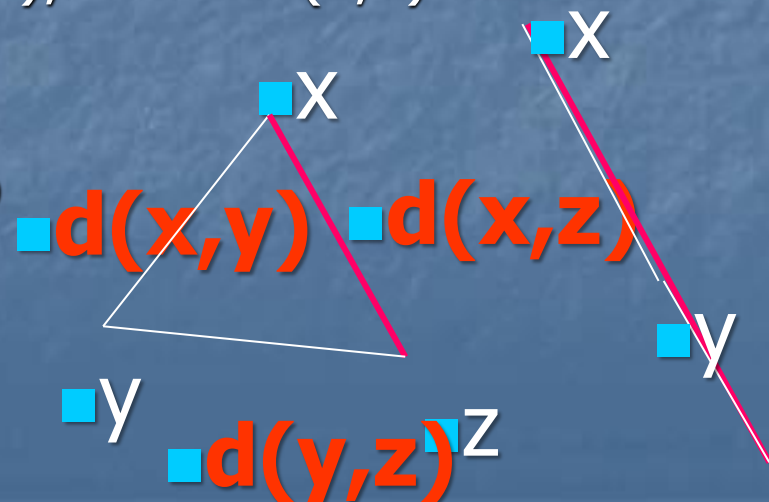
1. É simétrica $d(x,y) = d(y,x)$
2. Indivíduos não idênticos podem ser distinguidos – se $d(x,y) \neq 0$ então $x \neq y$

Uma distância é métrica se :

1. É simétrica $d(x,y) = d(y,x)$
2. Indivíduos não idênticos podem ser distinguidos – se $d(x,y) \neq 0$ então $x \neq y$
3. Indivíduos idênticos não podem ser distinguidos – se temos elementos idênticos (x,x) , então $d(x,x) = 0$

Uma distância é métrica se :

1. É simétrica $d(x,y) = d(y,x)$
2. Indivíduos não idênticos podem ser distinguidos – se $d(x,y) \neq 0$ então $x \neq y$
3. Indivíduos idênticos não podem ser distinguidos – se temos elementos idênticos (x,x) , então $d(x,x) = 0$
4. Desigualdade triangular:
dado 3 entidades (x,y,z) então
 $d(x,z) \leq d(x,y) + d(y,z)$



coeficientes- métrico/não-métrico

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 9.2, only apply when there are no missing data.

Similarity	$D = 1 - S$ 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

escolhendo coeficientes – Legendre & Legendre (1998)

Table 7.3

Choice of an association measure among objects (Q mode), to be used with species descriptors (asymmetrical coefficients). For explanation of levels 5, 7 and 8, 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: <i>coefficient of community</i> (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) Data: raw abundances | see 4 |
| 4) Coefficients without associated probability levels | see 5 |
| 5) No standardization by object; the same difference for either abundant or rare species, contributes equally to the similarity between sites: <i>coefficients of Steinhaus</i> (S_{17}) and <i>Kulczynski</i> (S_{18}) | |
| 5) 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 dist. (D_{16}), <i>Hellinger dist.</i> (D_{17}) | |
| 4) Probabilistic coefficient: probabilistic χ^2 similarity (S_{22}) | |
| 3) Data: normalized abundances (or, at least, distributions not skewed) or classes on a scale of relative abundances (e.g. 0 to 5, 0 to 7). [Normalization, Subsection 1.5.6, is useful when abundances cover several orders of magnitude.] | see 6 |

Efeitos de “dados faltando”

■ Distância euclidiana

A

1	0,0		
2	2,24	0,0	
3	2,24	0,0	0,0
	1	2	3

$$d_{12} = d_{13} + d_{23}$$

1	2	3
0	1	1
0	1	1
1	1	1
0	0	0
0	1	1
1	0	0
1	0	0
1	1	1
0	0	0
0	0	0

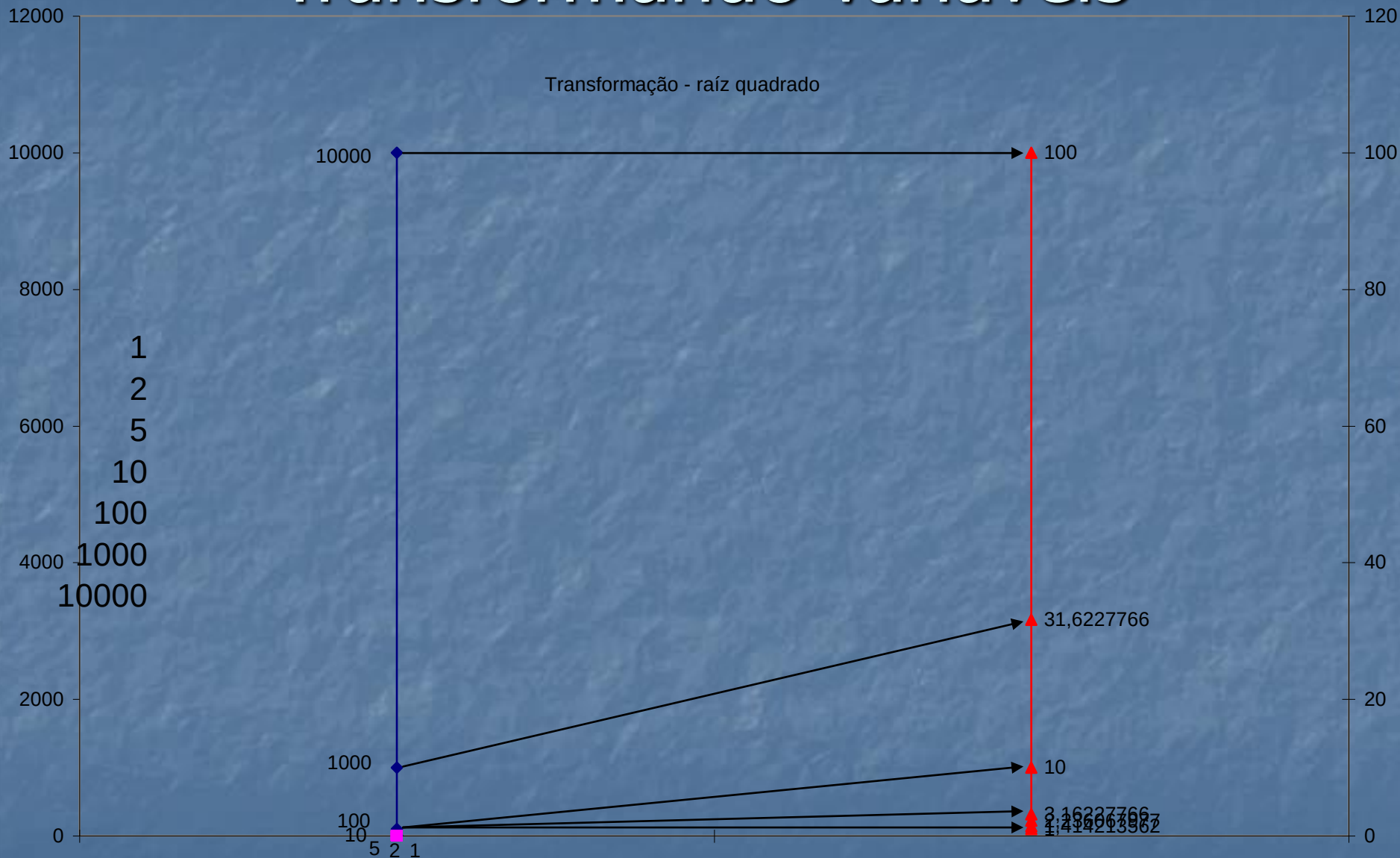
B

1	2	3
0	?	1
0	1	?
1	1	1
0	0	0
0	1	1
1	0	?
1	0	0
1	1	1
0	0	0
0	0	0

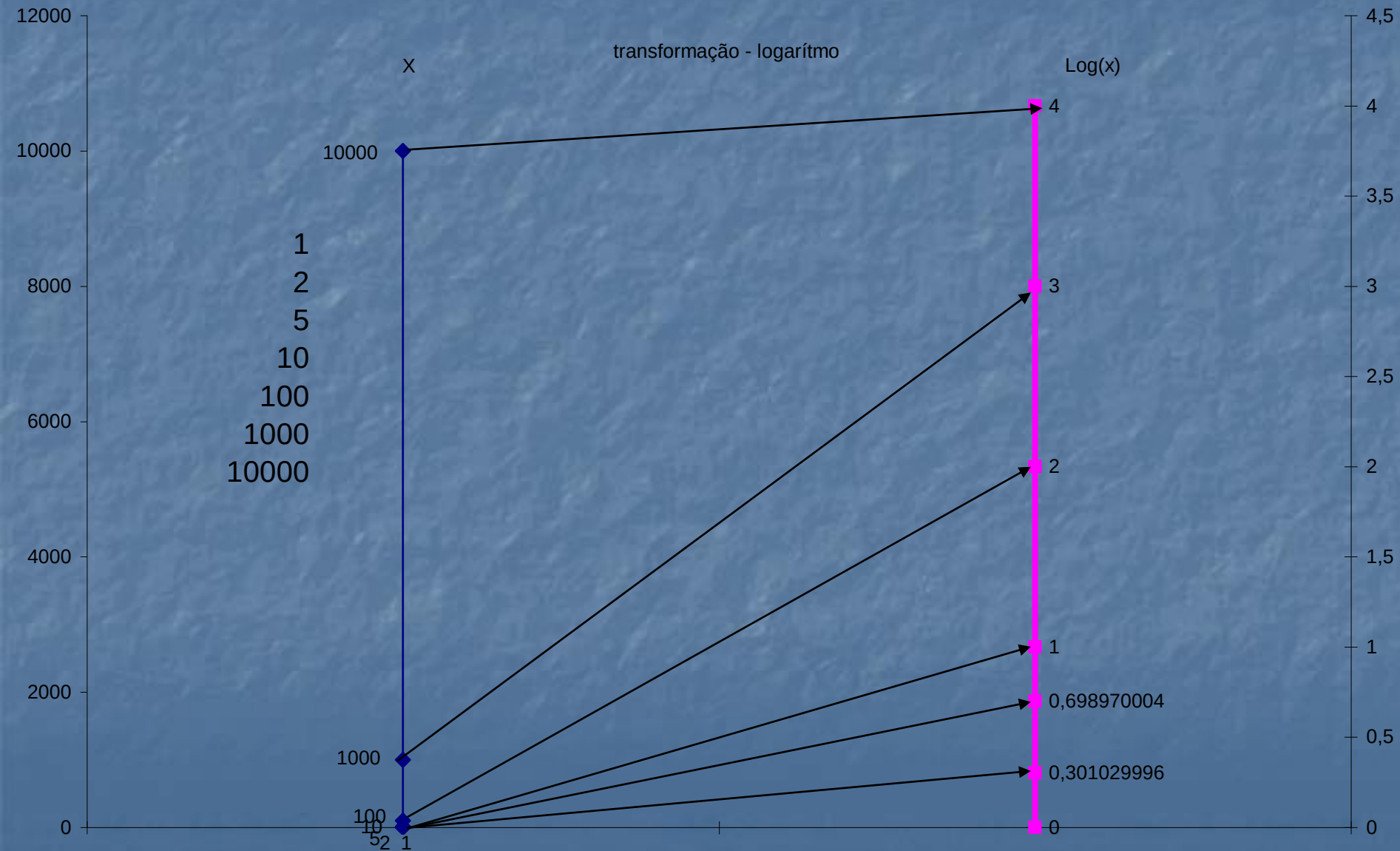
1	0,0		
2	2,0	0,0	
3	1,73	0,0	0,0
	1	2	3

$$d_{12} > d_{13} + d_{23} \quad !$$

Transformando variáveis

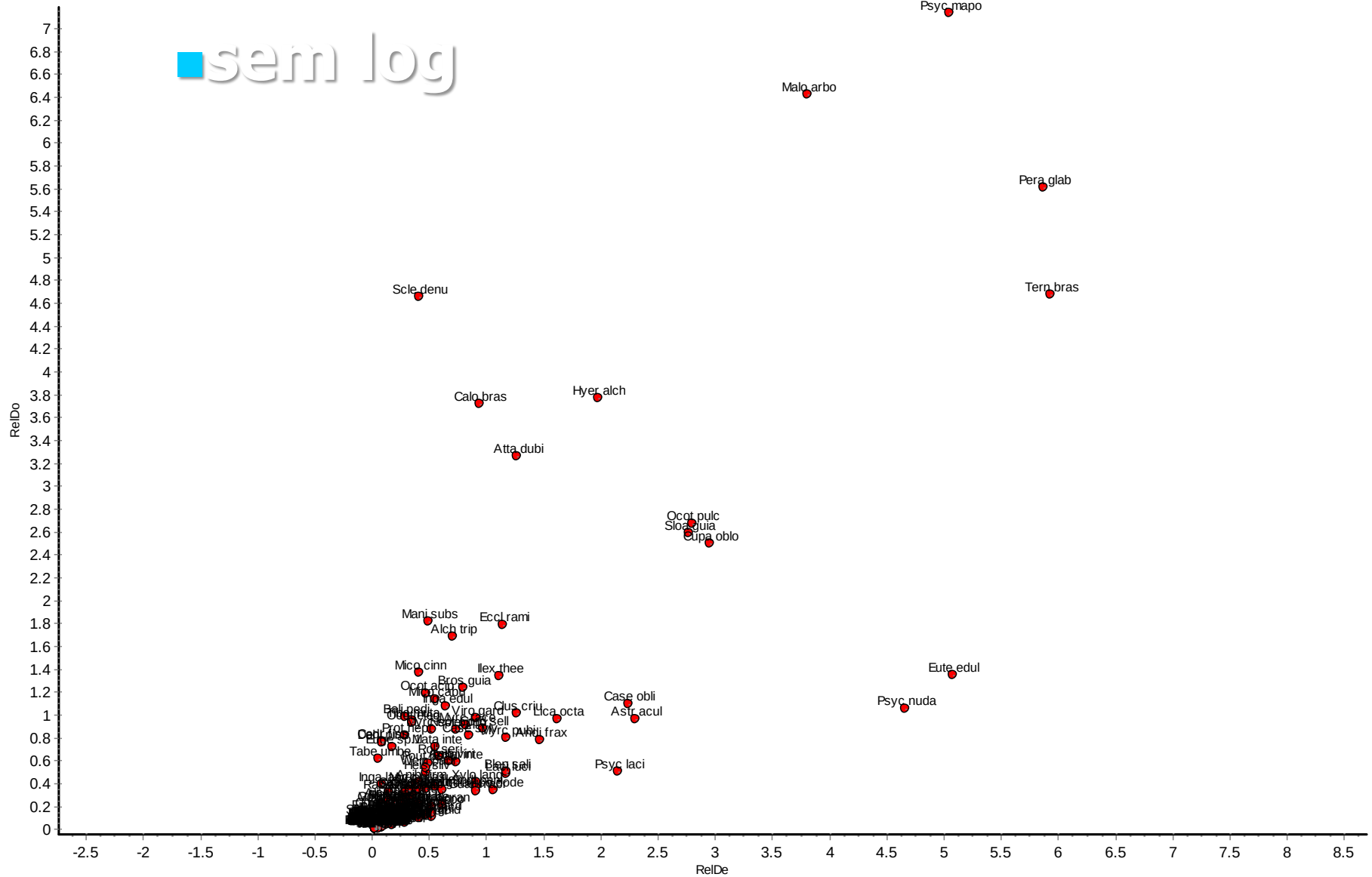


■ Transformando variáveis



Parâmetros para Espécies : Teste

ReIDe x ReIDo



com log

Parâmetros para Espécies : Teste

ReIDe x ReIDo

