

Biologia Quantitativa

Análises de Agrupamento

Depto de Zoologia
13 de agosto de 2024

Bibliografia de Referência

- Pielou, E. C. 1984. The Interpretation of Ecological Data. John Wiley & Sons, New York. Xi + 263 pp.
- Capítulo 2 do livro de Pielou. Figs do powerpoint são do livro.
- Ayres et al. 2000. Bioestat 2. Cap. 4 discute distancias euclidiana e outras distancias multivariadas (mahalonobis e penrose)
- Gotteli e Ellison. 2010. Princípios de Estatística em Ecologia. Editora Artmed

Análise de Classificação

- Parte de métodos multivariados que permitem organizar múltiplas amostras contendo múltiplas variáveis
- Podemos usar estes métodos para classificar ou ordenar amostras, ou para extrair e consolidar variáveis
- A análise de classificação é um método para gerar classes hierárquicas
- Portanto se aplica aos casos em que há razões para supor ou para testar a existência de classes nos conjuntos de dados
- A classificação hierárquica permite dar métrica ao grau de separação entre as classes

Análise de Classificação

- Objetivo: agrupar as amostras em classes hierárquicas
- Metodologia:
 - Definir a métrica de distância multidimensional entre as amostras. Cada variável corresponde a uma dimensão
 - Definir a fórmula de agregação sequencial das amostras na árvore. Cada método tem seu algoritmo
 - Cada tipo de análise de classificação é caracterizada pela métrica multidimensional usada e pelo algoritmo de agregação escolhido
- Objetivo final: usar as classes resultantes da análise para organizar os dados e elucidar hipóteses

Distância Multivariada

- Objeto: determinar agregações em conjuntos de objetos
- Metodologia:
- Definem-se as distâncias multidimensionais entre os objetos (ex. Distância euclidiana)
- Implementa-se algoritmo sequencial de agrupamento segundo regra hierárquica
- Portanto o tipo de análise (=cluster analysis) é nomeado de acordo com a medida de distância e o método de agrupamento.

Exemplos de Distâncias

- Podem ser métricas ou não métricas
- Métrica: propriedades geométricas desigualdade do triângulo (cada lado menor que a soma dos demais)
- Distância euclidiana - métrica
- Manhattan - métrica
- Percentagem dissimilaridade - não métrica
- Índice de similaridade Sorensen - não métrica
- Índice de Jaccard - métrico
- Dist. Marczewski-Steinhaus - métrica

Exemplos de Algoritmos

- Vizinho mais próximo
- Vizinho mais distante
- Centróide
- Distância média
- UPGMA
- Não hierárquicos - redução de dados

Escolha de Algoritmos

- Vizinho mais próximo ou mais distante -raro
- Variância mínima - útil para classes reais. Permite testar hipóteses sobre as classes
- Valor médio com ou sem peso. Úteis para controlar importância relativa das variáveis no resultado
- Centróide ou média. Centróide geometricamente lógico, mas pode dar reversões
- Portanto o tipo de análise (=cluster analysis) é nomeado de acordo com a medida de distância e o método de agrupamento.

Distância Euclidiana

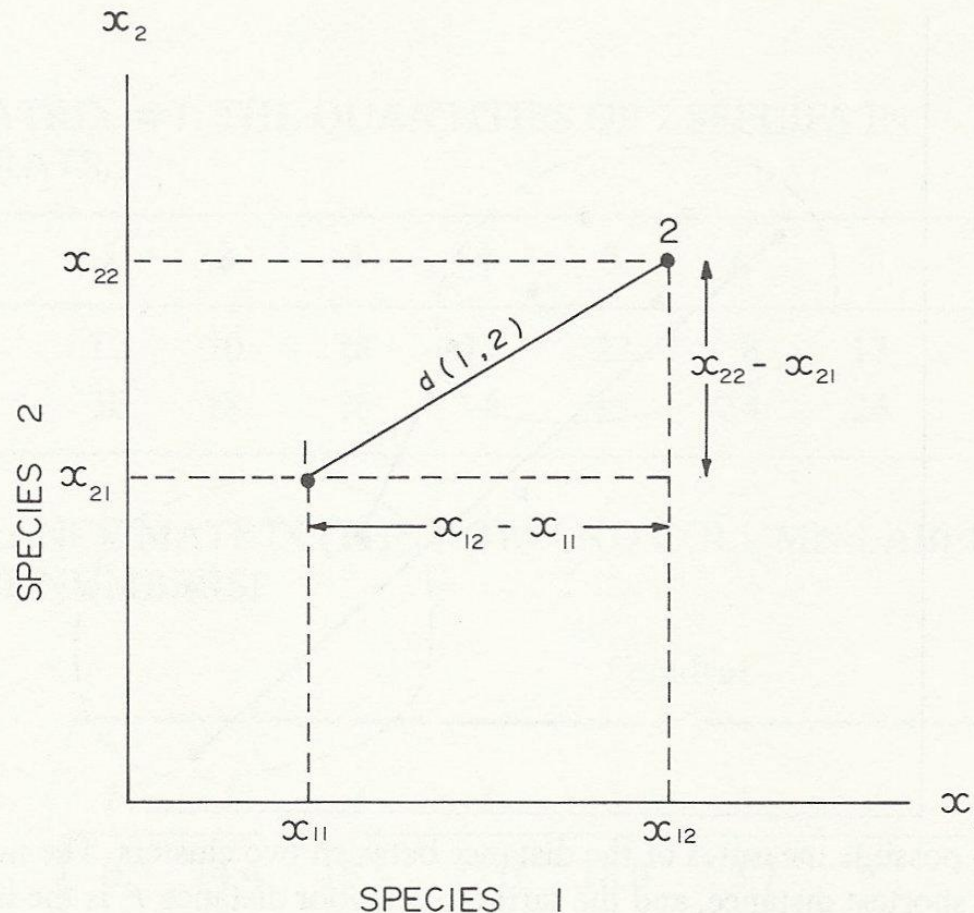


Figure 2.1. The distance between points 1 and 2, with coordinates (x_{11}, x_{21}) and (x_{12}, x_{22}) , respectively, is $d(1,2)$. From Pythagoras's theorem,

$$d(1,2) = \sqrt{\{(x_{12} - x_{11})^2 + (x_{22} - x_{21})^2\}}.$$

[illegible]

Métodos de Agregação

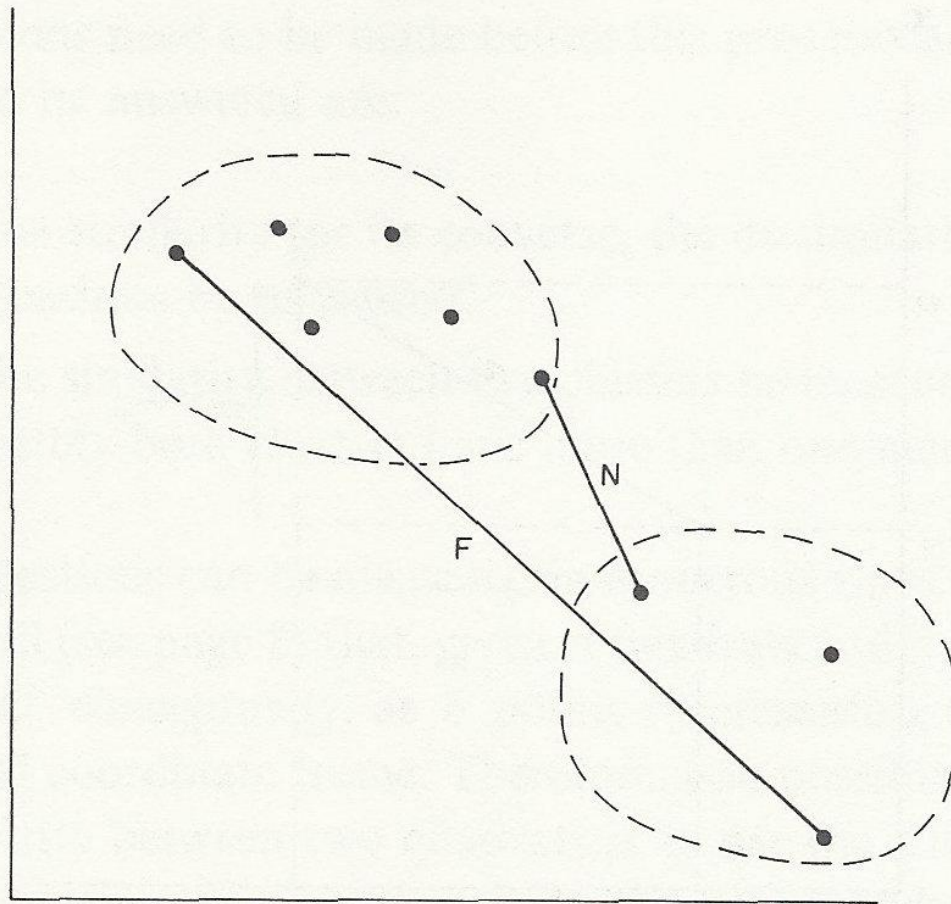


Figure 2.2. Two possible measures of the distance between two clusters. The nearest-neighbor distance N is the shortest distance, and the farthest-neighbor distance F is the longest distance between a member of one cluster and a member of the other.

Métodos de Agregação

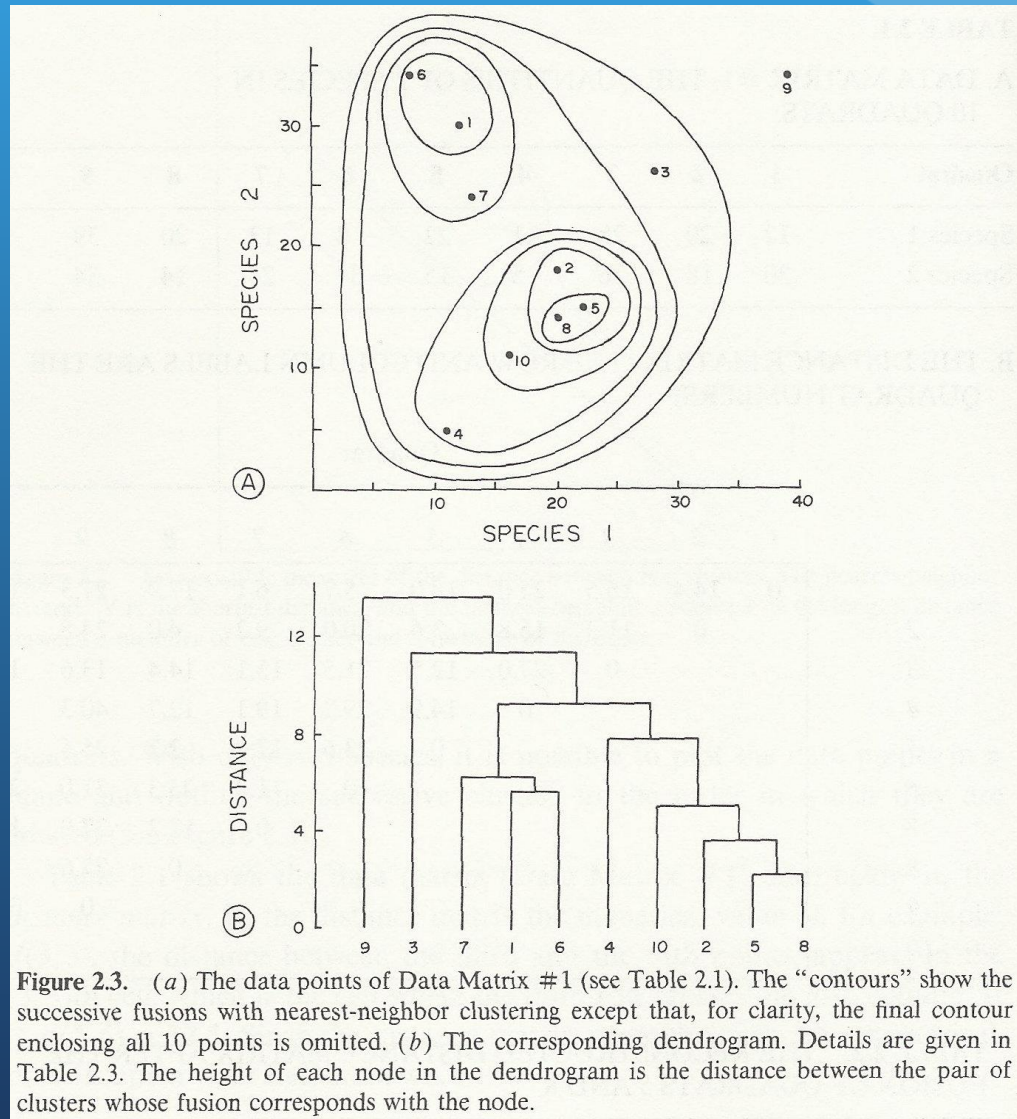
TABLE 2.3. STEPS IN THE NEAREST-NEIGHBOR CLUSTERING OF DATA MATRIX #1.

Step Number	Fusions ^a	Nearest Points ^b	Distance Between Clusters
1	5, 8	5, 8	2.2
2	[5, 8], 2	2, 5	3.6
3	[5, 8, 2], 10	8, 10	5.0
4	1, 6	1, 6	5.7
5	[1, 6], 7	1, 7	6.1
6	[5, 8, 2, 10], 4	4, 10	7.8
7	[5, 8, 2, 10, 4], [1, 6, 7]	2, 7	9.2
8	[5, 8, 2, 10, 4, 1, 6, 7], 3	2, 3	11.3
9	[5, 8, 2, 10, 4, 1, 6, 7, 3], 9	3, 9	13.6
10	All points are in one cluster		

^aUnbracketed numbers refer to individual quadrats. The numbers in square brackets are the quadrats in a cluster.

^bThe distance between these two points defines the distance (given in the last column) between the two clusters united at this step.

Análise de Agrupamento



Métodos de Agregação

TABLE 2.5. STEPS IN THE FARTHEST-NEIGHBOR CLUSTERING OF DATA MATRIX #1.

Step Number	Fusions	“Farthest” Points ^a	Distance between Clusters
1	5, 8	5, 8	2.2
2	[5, 8], 2	2, 8	4.0
3	1, 6	1, 6	5.7
4	4, 10	4, 10	7.8
5	[1, 6], 7	6, 7	11.2
6	3, 9	3, 9	13.6
7	[5, 8, 2], [4, 10]	2, 4	15.8
8	[1, 6, 7], [5, 8, 2, 4, 10]	4, 6	29.2
9	[1, 6, 7, 5, 8, 2, 4, 10], [3, 9]	4, 9	40.3
10	All points are in one cluster		

^aThese are the quadrats whose distance apart, shown in the last column, defines the distance between the two clusters.

Métodos de Agregação

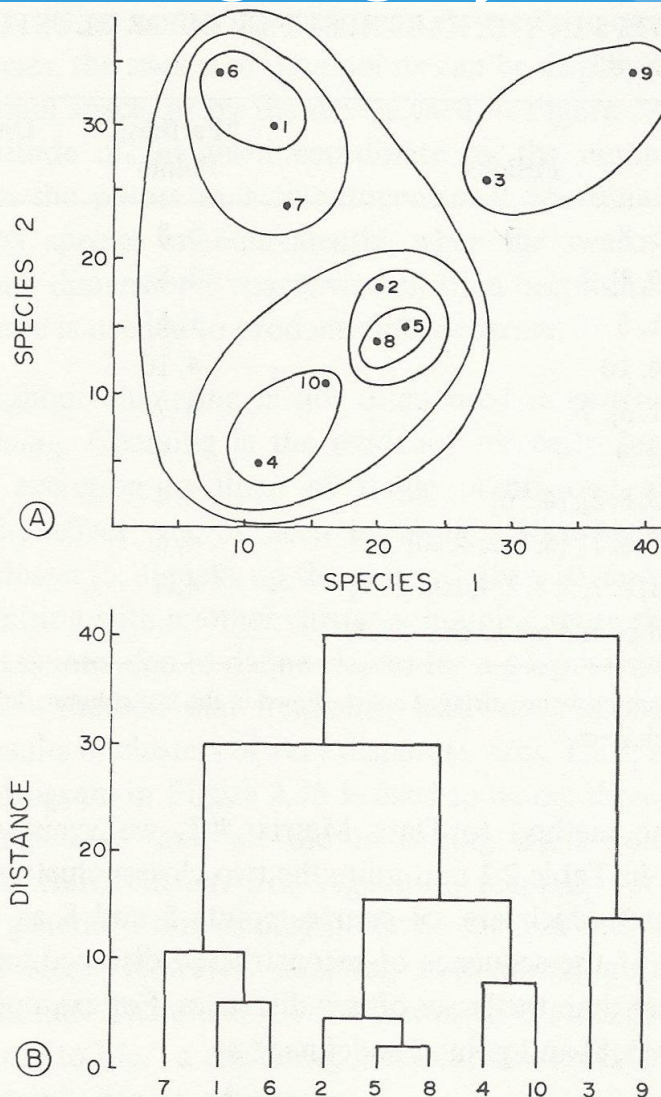


Figure 2.5. Farthest-neighbor clustering applied to Data Matrix #1. The dendrogram is based on Table 2.5.

Métodos de Agregação

TABLE 2.6. STEPS IN THE CENTROID CLUSTERING OF DATA MATRIX #1.

Step Number	Fusions	Square of Intercluster Distance
1	5, 8	5.00
2	2, [5, 8]	13.25
3	1, 6	32.00
4	10, [2, 5, 8]	43.56
5	7, [1, 6]	73.00
6	4, [2, 5, 8, 10]	162.50
7	3, 9	185.00
8	[7, 1, 6], [2, 5, 8, 10, 4]	326.25
9	[7, 1, 6, 2, 5, 8, 10, 4], [3, 9]	456.83
10	All points are in one cluster	

Métodos de Agregação

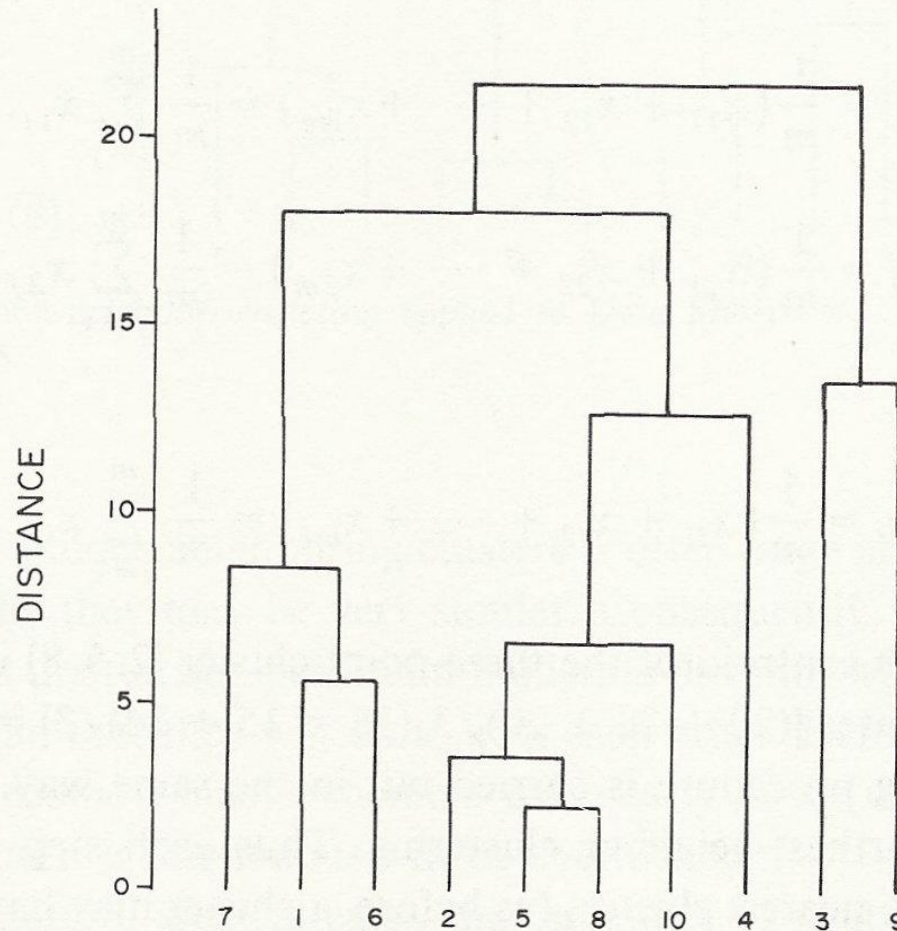


Figure 2.6. The dendrogram produced by applying centroid clustering to Data Matrix #1. It is based on Table 2.6.

Lagartos do Parque Nac GSV

Estudo de Ricarte e Nogueira

Lagartos PN Grande SV

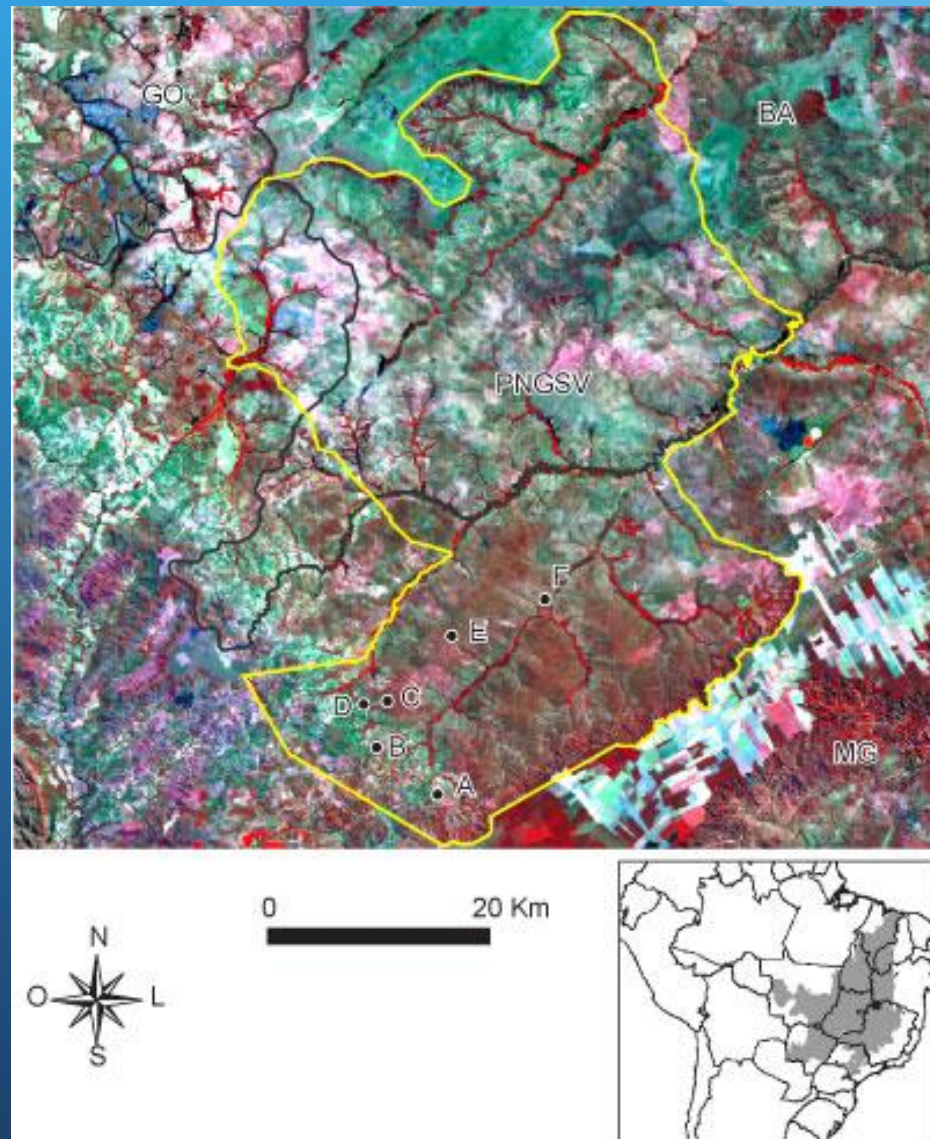


Figura 1. Área de estudo, com a localização das unidades amostrais e sede.

Figure 1. Study area, with location of sampling unites and base camp.

Lagartos PN Grande SV

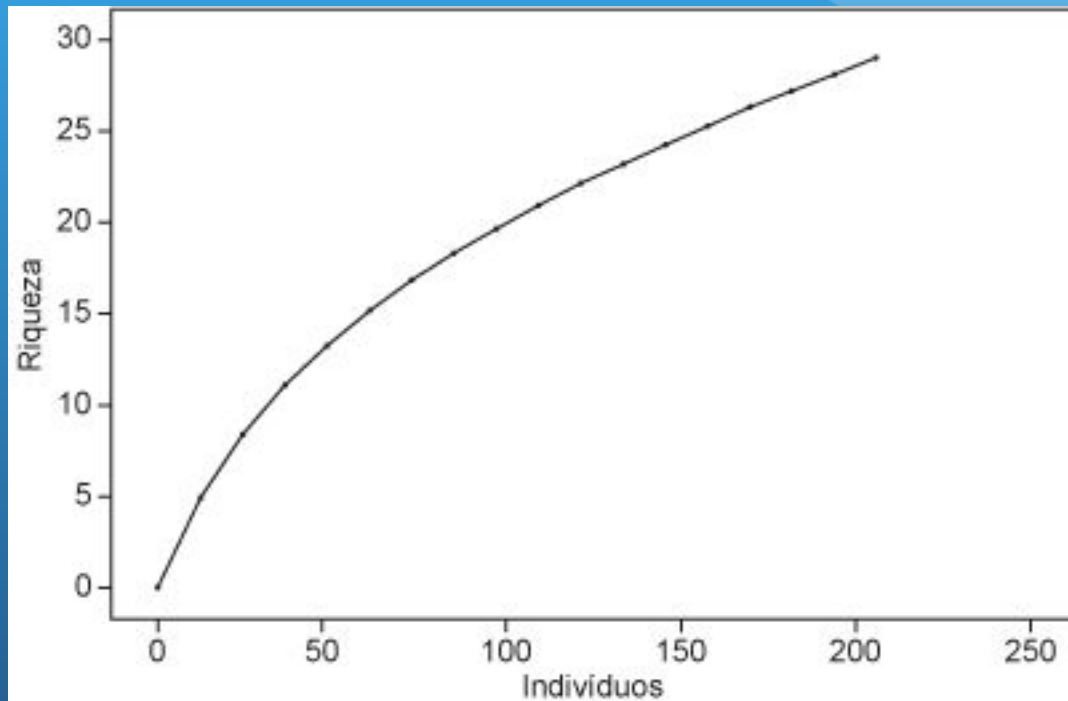


Figura 2. Curva de rarefação das espécies amostradas de Squamata no PNGSV para 206 espécimes registrados em 17 dias de coleta em campo (armadilhas + manual).

Figure 2. Rarefaction curve of the sampled species of squamates in GSVNP for 206 specimens registered in 17 days of field work (pitfall traps + manual sampling).

Lagartos PN Grande SV

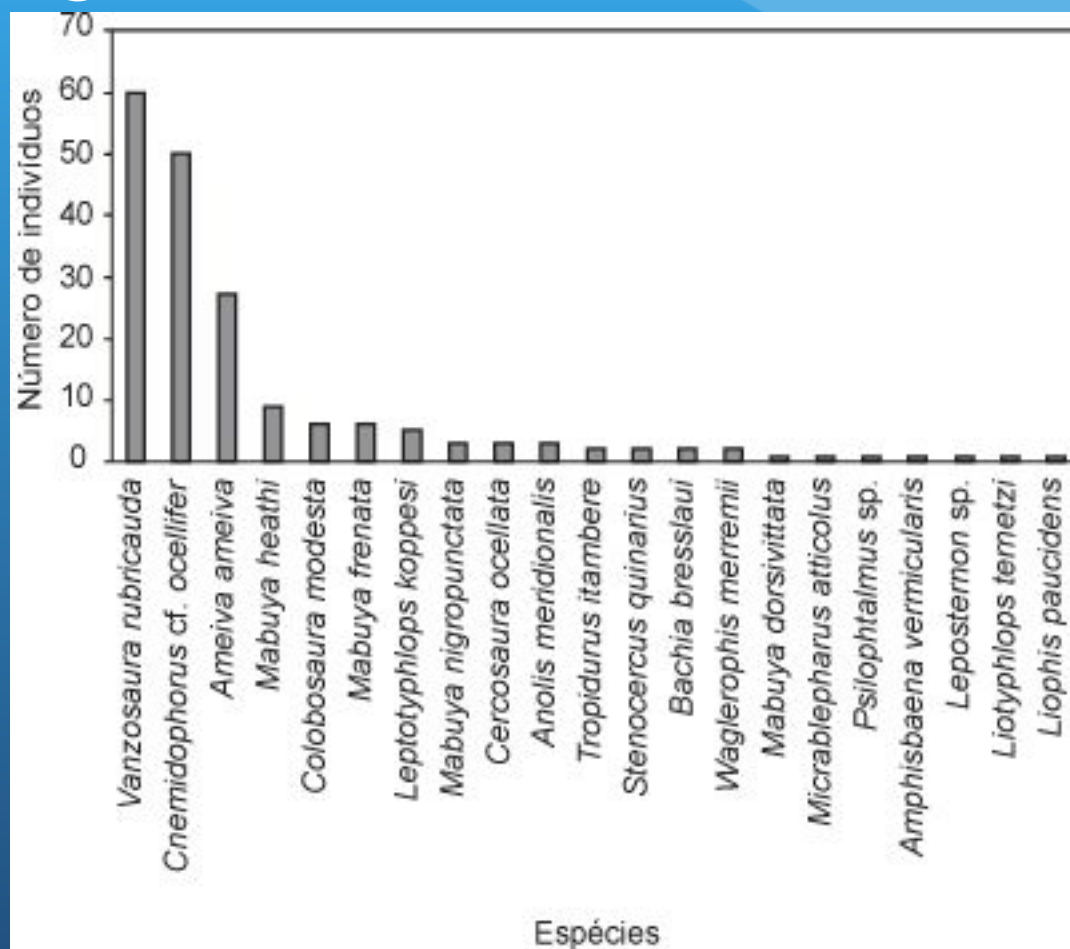


Figura 3. Abundância relativa das espécies de répteis Squamata coletadas em armadilhas de queda no PNGSV.

Figure 3. Relative abundance of squamate species, collected in pitfall traps in GSVNP.

Lagartos PN Grande SV

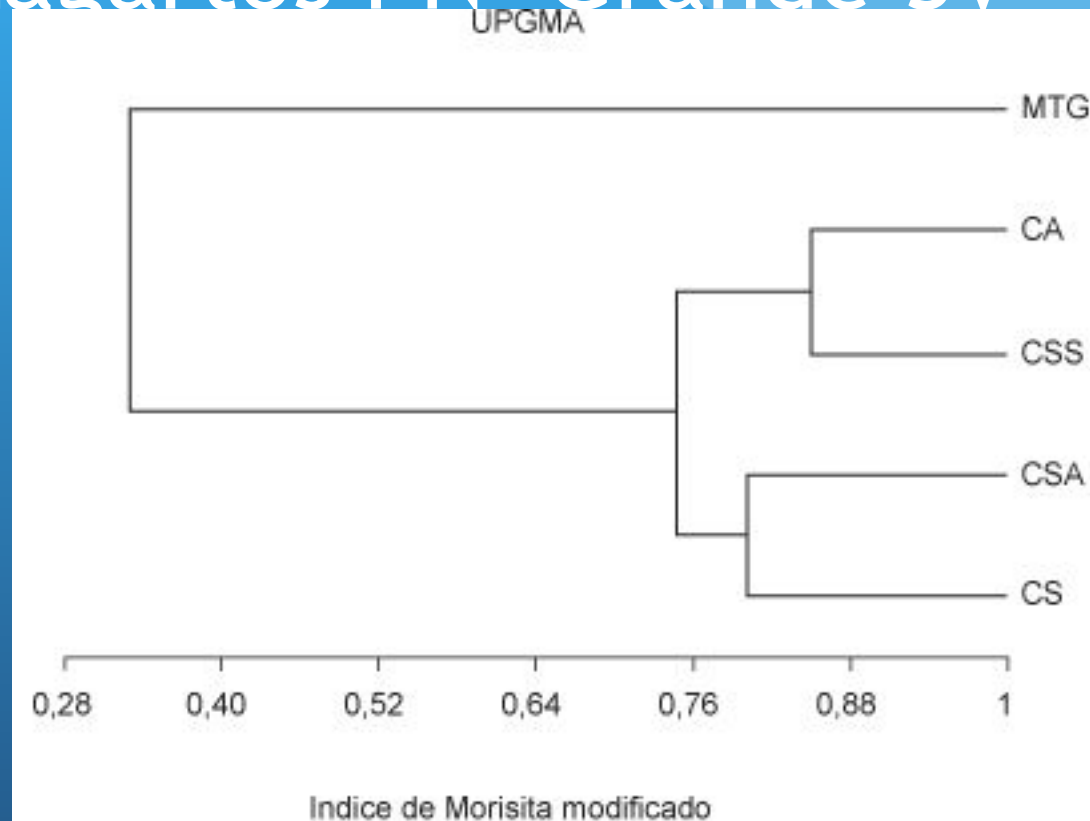


Figura 5. Dendrograma da similaridade entre os habitats amostrados no PNGSV baseado na riqueza e abundância de espécies amostradas. mtg, mata de galeria; csa, campo sujo arenoso; cs, campo sujo; ca, carrasco; e css, cerrado sensu stricto.

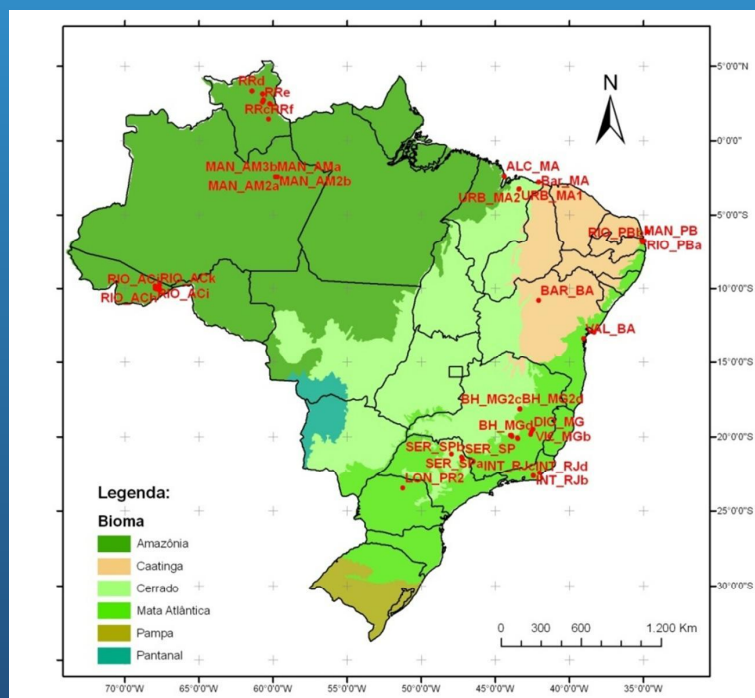
Figure 5. Dendrogram of similarity between sampled habitats in GSVNP based in the abundance of collected species. mtg, gallery forest; csa, “campo sujo” on sandy soil; cs, “campo sujo”; ca, “carrasco”; e css, cerrado sensu stricto

Artigo: Abelhas Euglossinas

- Nas últimas aulas discutimos o uso de análises estatísticas, com ênfase no tema de caracterização de distribuições e testes de hipóteses sobre tais características
- Primeiro devemos selecionar os dados que nos interessam como indicadores de parâmetros biológicos.
- Em seguida usamos a estatística para descrever as distribuições de dados e testar hipóteses a respeito das distribuições.
- Finalmente usamos os resultados dos testes estatísticos para fazer inferências sobre os processos biológicos dos quais os dados originais foram gerados

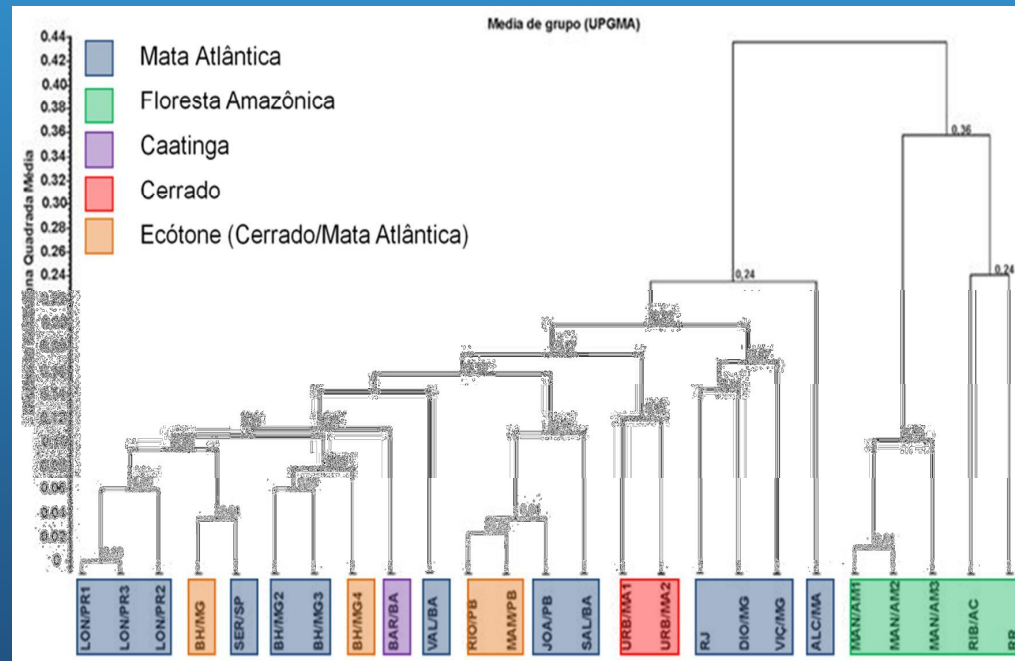
Áreas de Amostragem

com distribuição do biomas nacionais. As siglas estão discriminadas na Figura 2.



Cluster

levantamento.



Siglas: LON/PR: Londrina, Paraná; BH/MG: Belo Horizonte, Minas Gerais; SER/SP: Sertãozinho, SP; BAR/BA: Barras, BA; VAL/BA: Valença, BA; RIO/PB: Rio Tinto, PB; MAN/PB: Mangape, PB; JOA/PB: João Pessoa, PB; SAL/BA: Salvador, BA; URB/MA: Urbano de Santos, MA; RJ: Interior do estado do Rio de Janeiro; DIO/MG: Dionísio, MG; VIÇ/MG: Viçosa, MG; ALC/MA: Alcântara, MA; MAN/AM: Manaus, AM; RIO/AC: Rio Branco, AC; RR: Várias cidades do estado de Roraima.