Aula 23 Goodies*

* Goodies related to animals, plants and numbers…

Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003

Today at the Techonomy conference in Lake Tahoe, CA, the first panel featured Google CEO Eric Schmidt. As moderator David Kirkpatrick was introducing him, he rattled off a massive stat… Every two days now we create as much information as we did from the dawn of civilization up until 2003, according to Schmidt. That's something like five exabytes of data, he says. "*The real issue is user-generated content*," Schmidt said. He noted that pictures, instant messages, and tweets all add to this. Naturally, all of this information helps Google. But he cautioned that just because companies like his can do all sorts of things with this information, the more pressing question now is if they *should*. Schmidt noted that while technology is neutral, he doesn't believe people are ready for what's coming. "*I spend most of my time assuming the world is not ready for the technology revolution that will be happening to them soon*," Schmidt said.

https://techcrunch.com/2010/08/04/schmidt-data/?guccounter=1

SO WHAT IS A PETABYTE ANYWAY?

Source – www.mozy.com

WHAT IS A PETABYTE?

TO UNDERSTAND A PETABYTE WE MUST FIRST UNDERSTAND A GIGABYTE.

- 1 GIGABYTE = 7 MINUTES OF HD-TV VIDEO
- 2 GIGABYTES = 20 YARDS OF BOOKS ON A SHELF
- 4.7 GIGABYTES = SIZE OF A STANDARD DVD-R

THERE ARE A MILLION GIGABYTES IN A PETABYTE

"Let me repeat that: we create as much information in two days now as we did from the dawn of man through 2003." (That's something like 5 Exabytes of Data). - Eric Schmidt – Google 8/10

A PETABYTE IS A LOT OF DATA

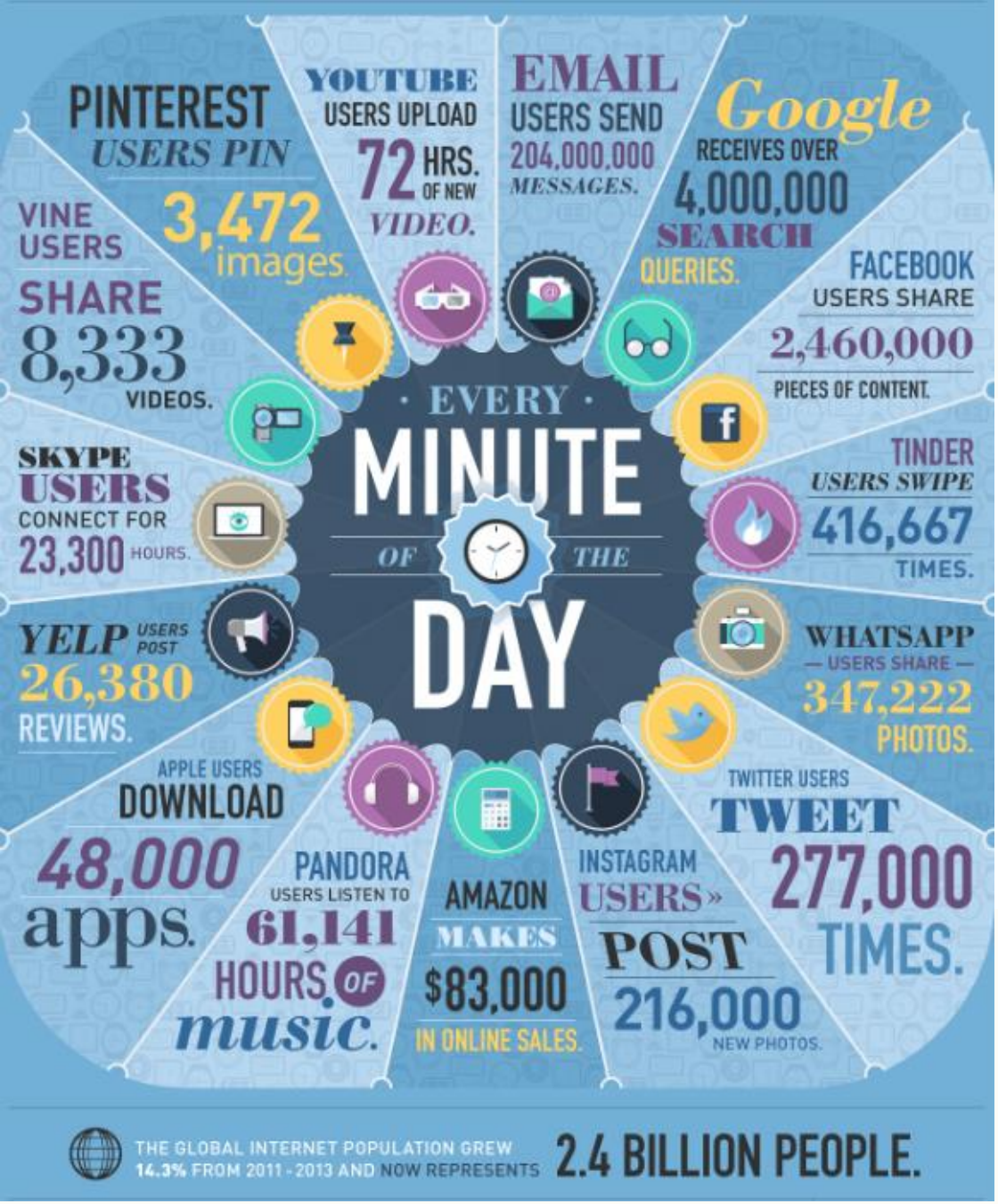
- 1 PETABYTE = 20 MILLION FOUR-DRAWER FILING CABINETS FILLED WITH TEXT
- 1 PETABYTE = 13.3 YEARS OF HD-TV VIDEO
- 1.5 PETABYTES = SIZE OF THE 10 BILLION PHOTOS ON → FACEBOOK
- 15+ PETABYTES = INTERNET USER'S DATA BACKED UP ON MOZY.COM
- 20 PETABYTES = THE AMOUNT OF DATA PROCESSED BY GOOGLE PER DAY
- 20 PETABYTES = TOTAL HARD DRIVE SPACE MANUFACTURED IN 1995
- 50 PETABYTES = THE ENTIRE WRITTEN WORKS OF MANKIND, FROM THE BEGINNING OF RECORDED HISTORY, IN ALL LANGUAGES

Twitter:
Over 7TB a Day in Tweets.

A ZETABYTE IS ONE MILLION PETABYTES!

Facebook:
More that 750 Million Users.
Average user creates 90 Pieces of content each month.
More than 30B pieces of content shared each month.

https://imgur.com/gallery/gpXRWoq



BIG DATA

https://aci.info/2014/07/12/the-data-explosion-in-2014-minute-by-minute-infographic/

Ecologia Numérica - Avaliação

- 50% exame teórico escrito (perguntas de desenvolvimento)
- 50% exame teórico-prático com recurso a computador e ao software R
- Duração de cerca de 2 horas
- Precisam de ter pelo menos 8 a ambas as partes para serem aprovados

Datas:

	1a Época	2a Época	Época Especial
Teórico	21 Jan 2020 – 16:30-19:00	07Fev 2020 – 16:30-19:30	tbc
Teórico-Prático	22 Jan 2020 – 8:00-18:00	08 Fev 2019 – 9:00-15:00	tbc

Para ser possível a aprovação da disciplina, os alunos terão que garantir presença em pelo menos 10 das aulas teórico-práticas, não sendo exigida presença obrigatória nas aulas teóricas.

EXAMES

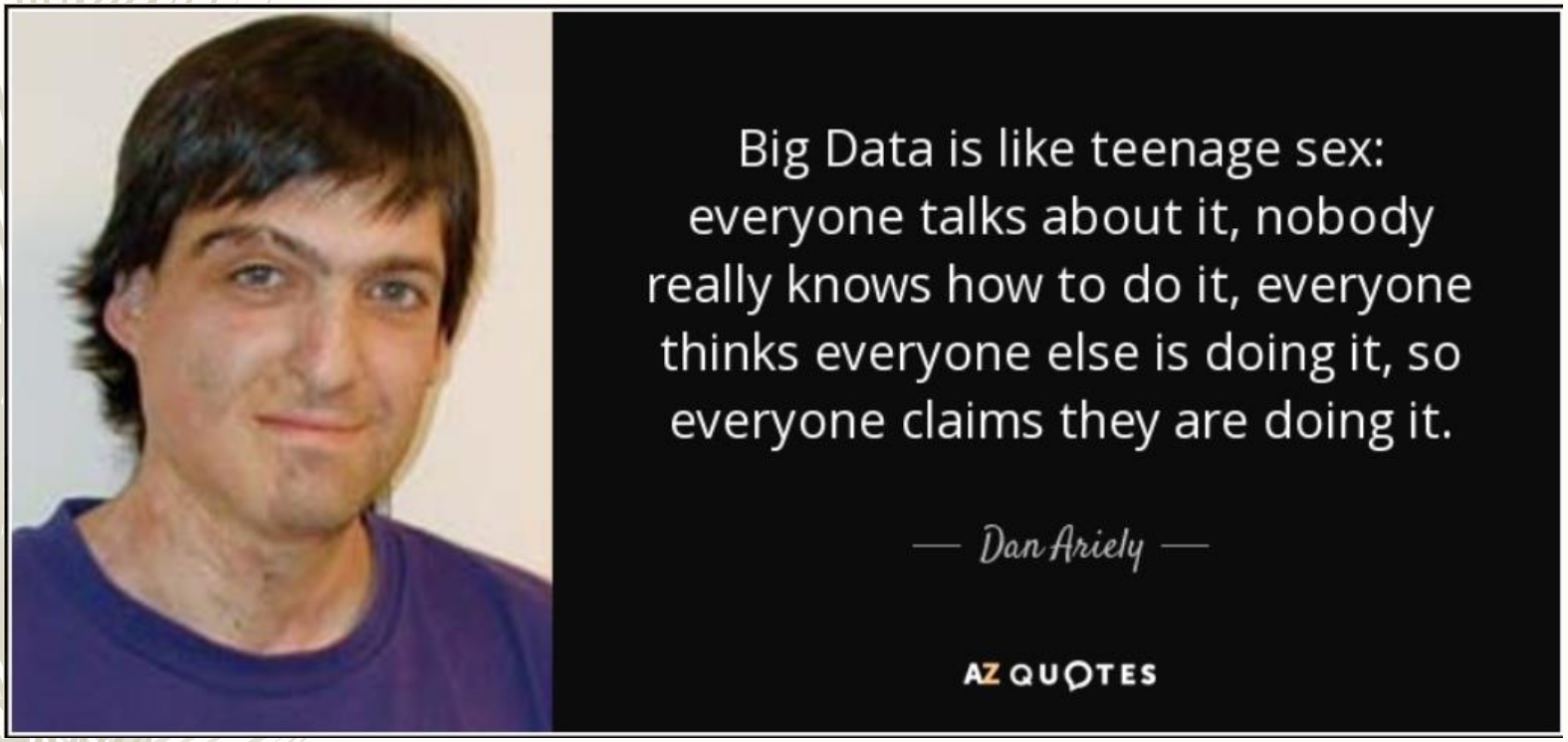
Prático

1. Duração: 1.5 a 2 horas
2. Com recurso ao computador
3. Logo com consulta
4. Logo com perguntas que implicam ligar o cérebro
5. Feito a pares (por razões logísticas e a menos que eu o consiga evitar), mas quem quiser fazer sozinho pode
6. Perguntas parecidas com o que foi sendo implementado nas TPs
7. Principal atributo a ser avaliado é a capacidade de interpretação, vs. capacidade de decorar coisas
8. Será necessário inscreverem-se no exame e selecionar o horário desejado

Teórico

1. Duração 2 a 2.5 horas
2. Precisam de uma máquina de calcular simples, não aceito uso de telemóveis!
3. Não tem consulta – tentativas de consulta não autorizadas não serão admitidas e implicam anulação do exame
4. Foca-se nos aspetos mais teóricos/conceptuais
5. No entanto, privilegia o raciocínio critico e a criatividade em detrimento da “decoração”

Ecologia Numérica - Aula Teórica 23 – 03-12-2018



https://www.azquotes.com/quote/661939

```
> sda
          1         2         3         4
2 5.010198
3 5.675741 5.938724
4 4.296223 4.906653 4.662499
5 5.926676 6.032184 6.158771 5.805220
```

Cluster Dendrogram

Cluster Dendrogram

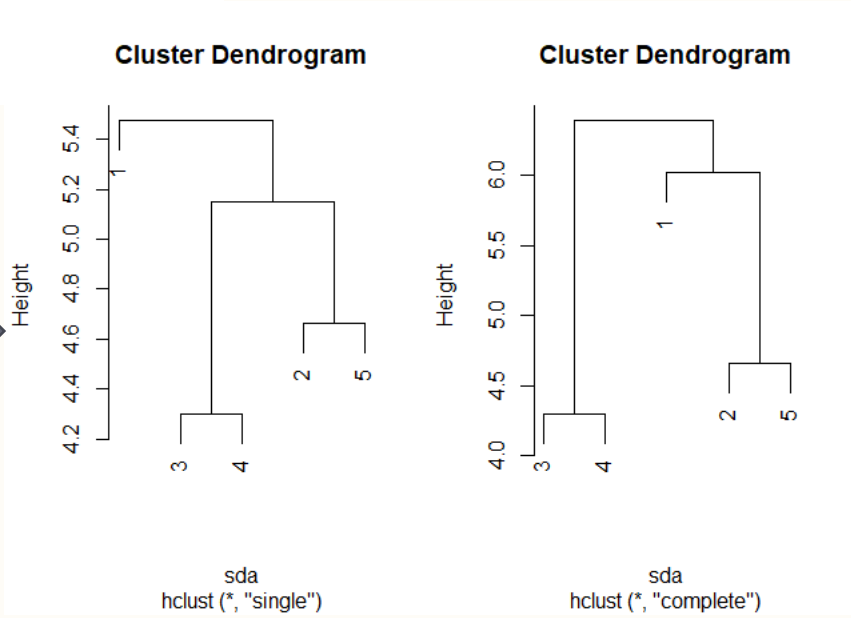
sda
hclust (*, "single")

sda
hclust (*, "complete")

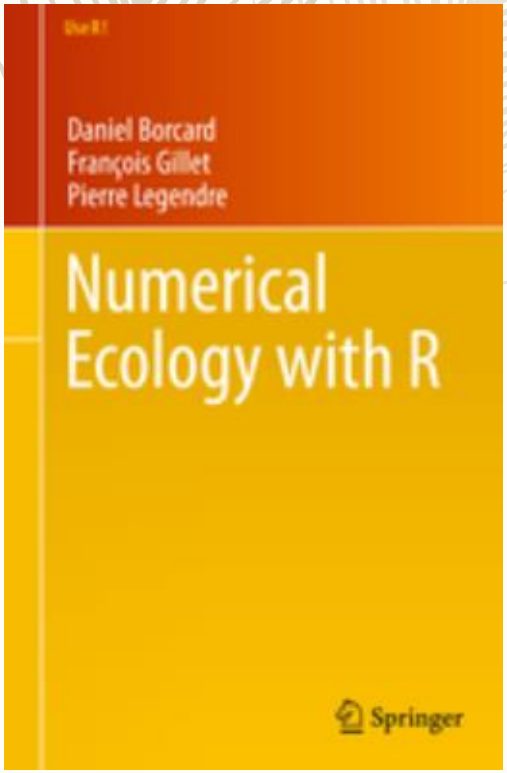
```
set.seed(2245)
abund=matrix(rpois(75,lambda=8),ncol=15,nrow=5)
#make 1 species really abundant
abund[,1]=c(1000,200,200,10,10)
#and one with the inverse pattern, but less abundant
abund[,2]=c(10,10,200,200,1000)/10
#untransformed data
da=dist(abund)
hcdaC=hclust(da,method="complete")
#transformed data
sda=dist(scale(abund))
hcsdaC=hclust(sda,method="complete")
par(mfrow=c(1,2))
plot(hcdaC)
plot(hcsdaC)
hcsdaS=hclust(sda,method="single")
par(mfrow=c(1,2))
plot(hcsdaS)
plot(hcsdaC)
```

Um “pedido” de uma árvore com estrutura de grupos distintos

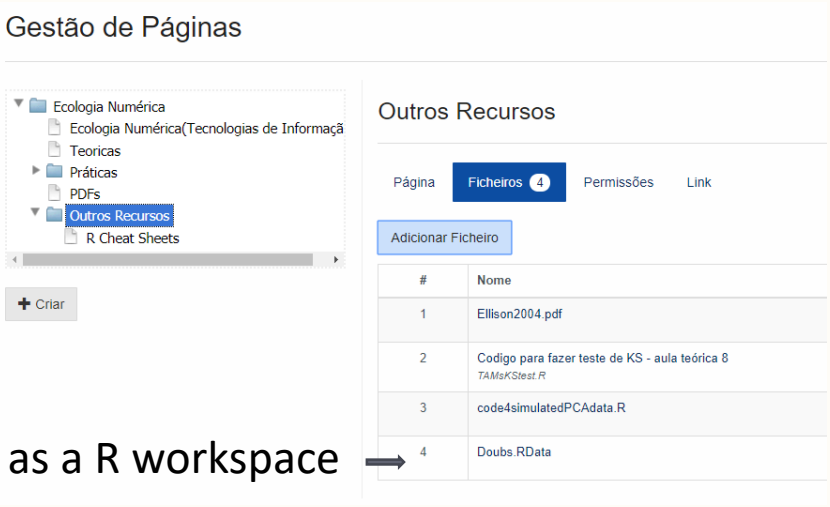
```
> round(sda,3)
      1     2     3     4
2 6.024
3 6.057 5.294
4 5.618 5.147 4.301
5 5.477 4.663 6.394 5.460
```



The Doubs data set that is used in the present book consists of three matrices containing part of the data used by Verneaux for his studies. These data have been collected at 30 sites along the Doubs River, which runs near the France–Switzerland border in the Jura Mountains. The first matrix contains coded abundances of 27 fish species, the second matrix contains 11 environmental variables related to the hydrology, geomorphology and chemistry of the river, and the third matrix contains the geographical coordinates (Cartesian, X and *Y*) of the sites. These data have already served as test cases in the development of numerical techniques (Chessel et al. 1994).



Dataset added in FENIX as a R workspace



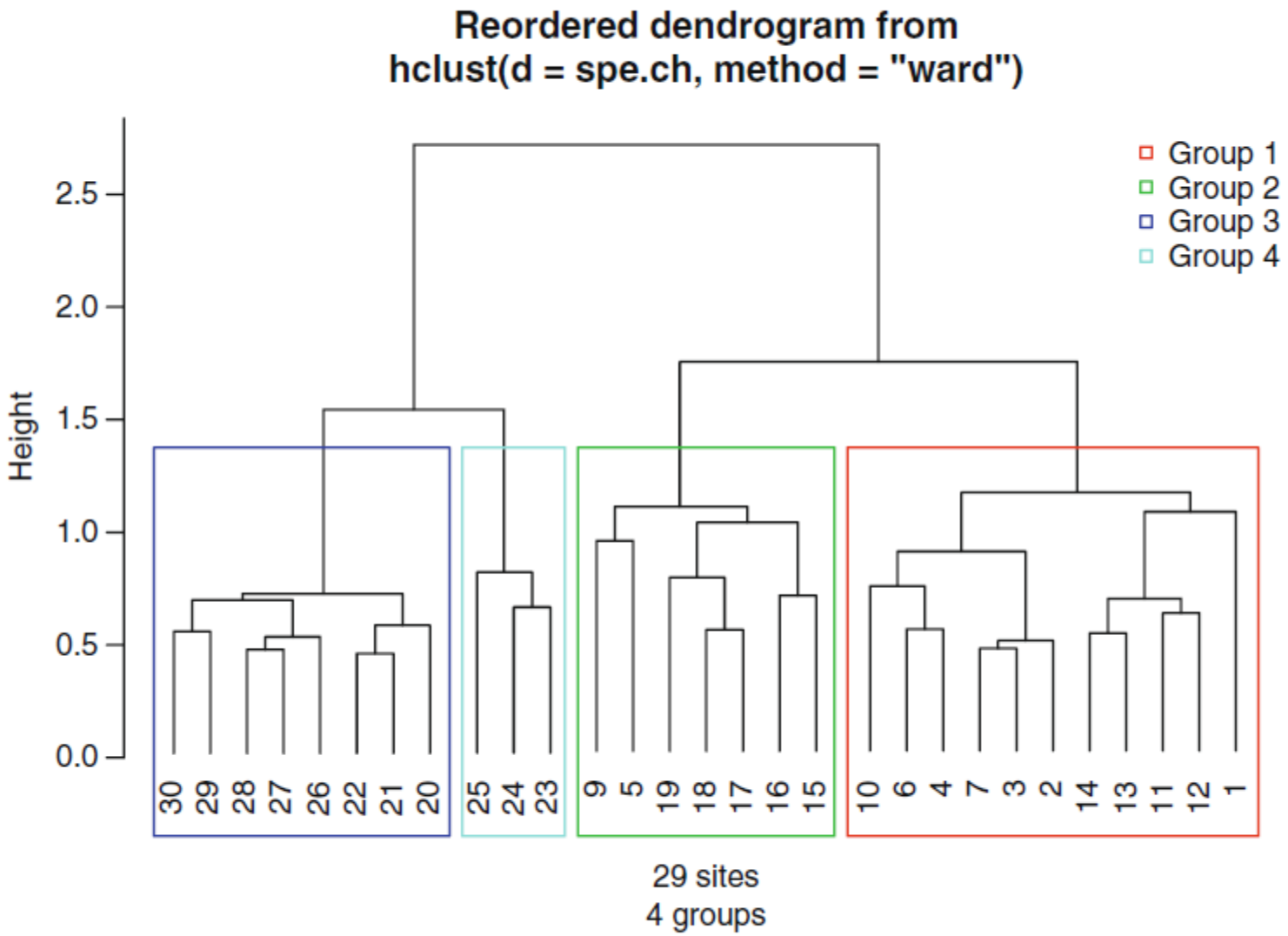


Fig. 4.11 Final dendrogram with *boxes* around the four selected groups. Species data

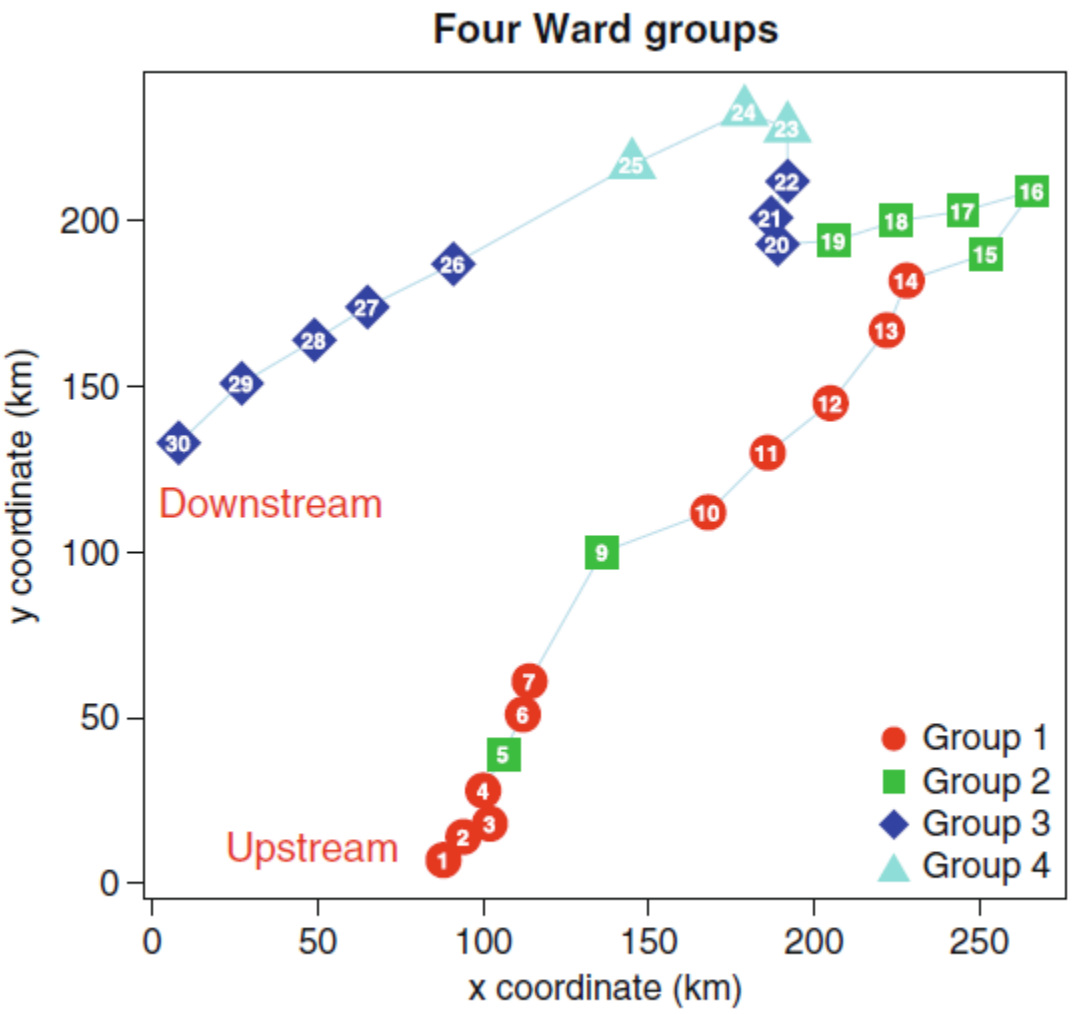
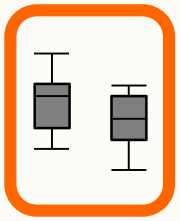


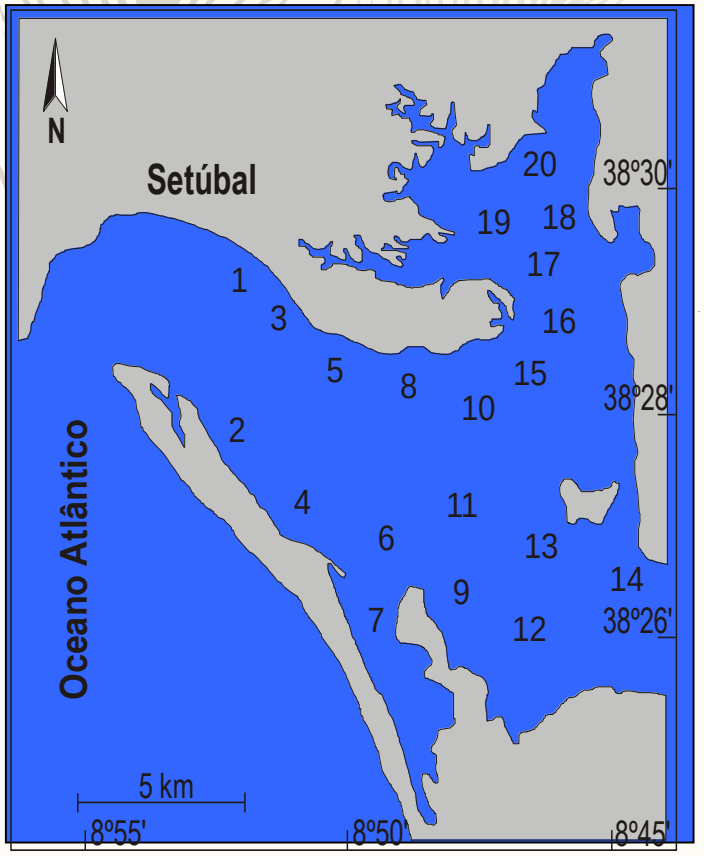
Fig. 4.12 The four Ward clusters on a map of the Doubs river



agrupamento

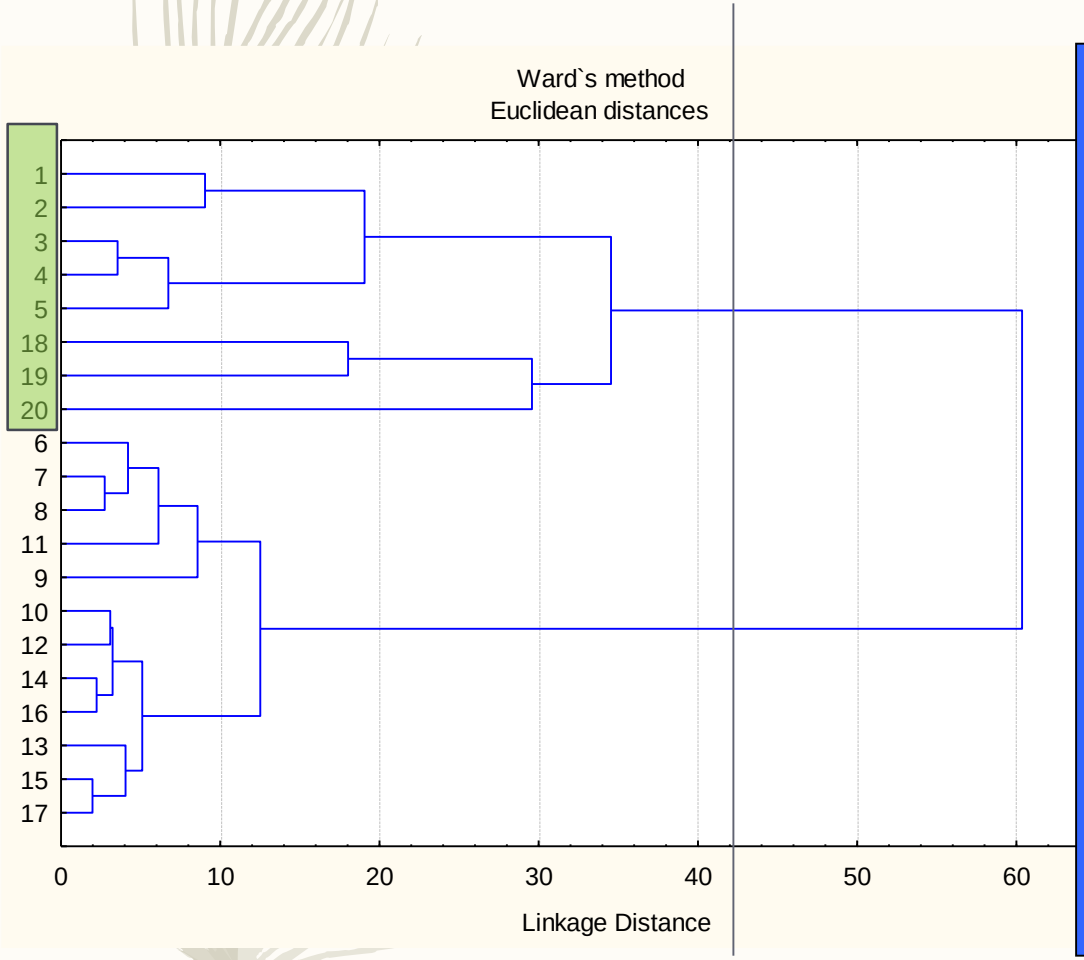
Exemplo:

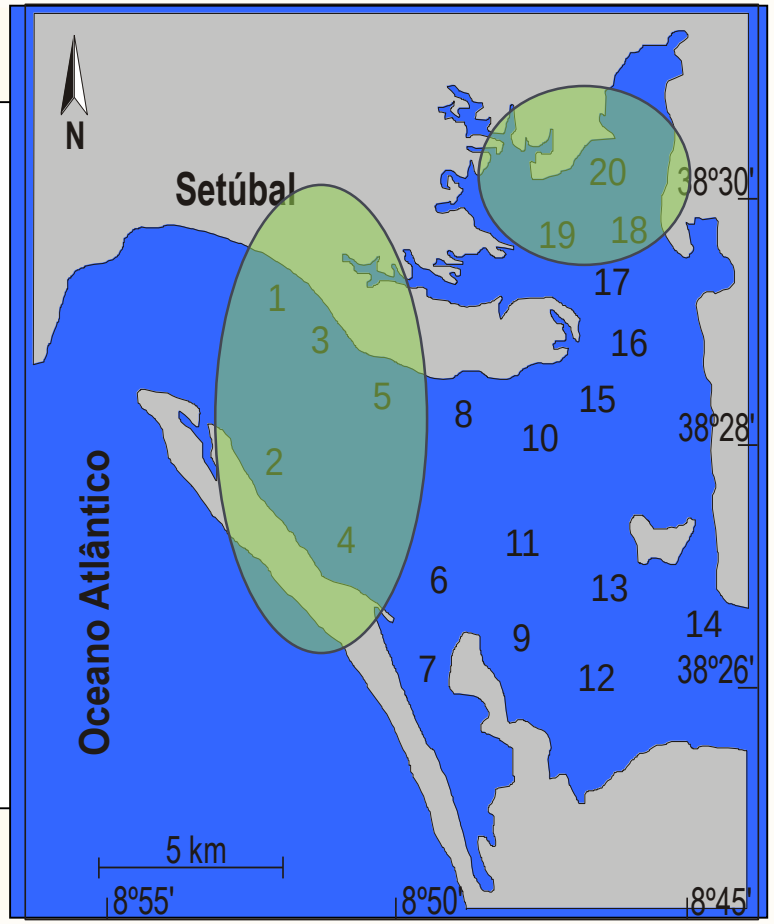
Dados de abundâncias de espécies de peixes em 20 estações de amostragem no estuário do Sado.



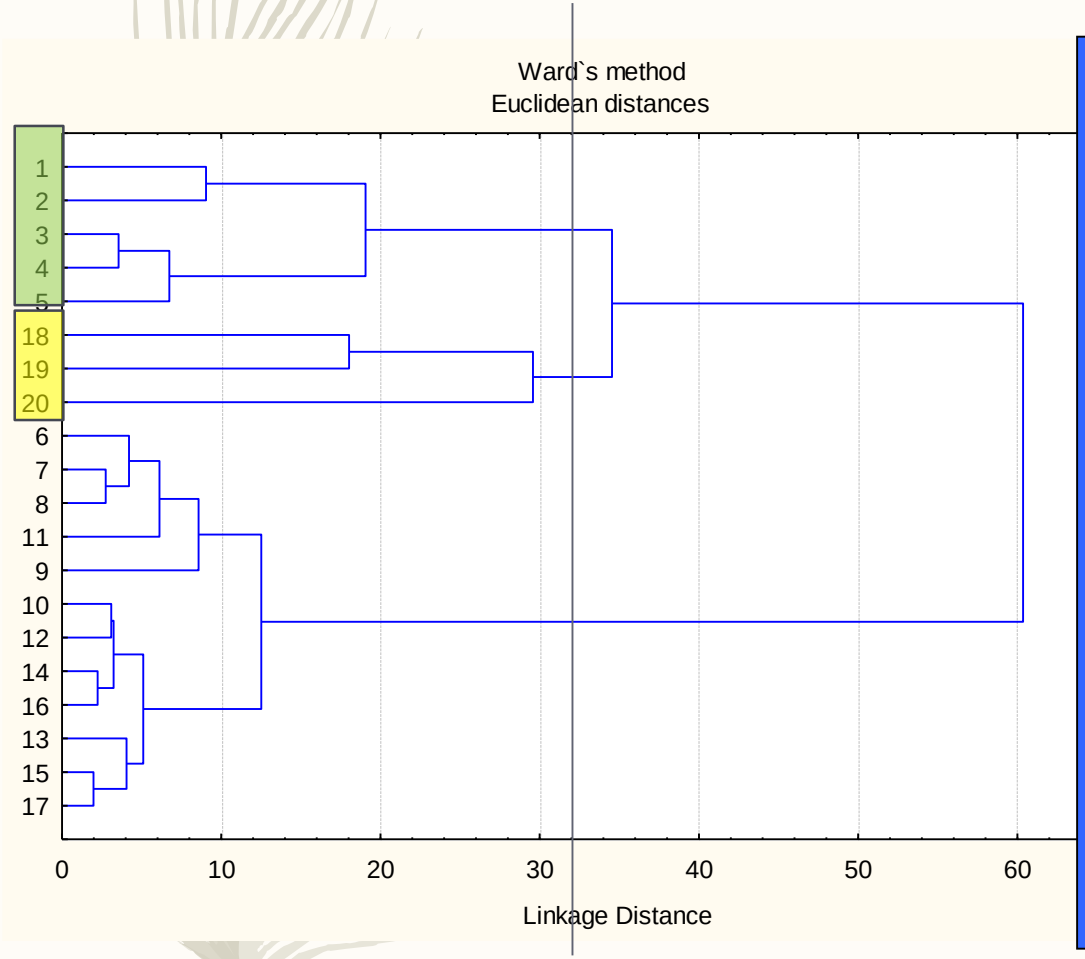
A maneira de perceber se a análise de agrupamento é sensata é ver se ela é interpretável do ponto de vista ecológico. E neste caso é, ou seja, de facto, de acordo com as abundancies de peixes (ou seja, de acordo com as comunidades de peixes) Podemos arranjar um agrupamento que muito provavelmente reflete a alteração das comunidades ao longo do gradiente de salinidade do estuário.

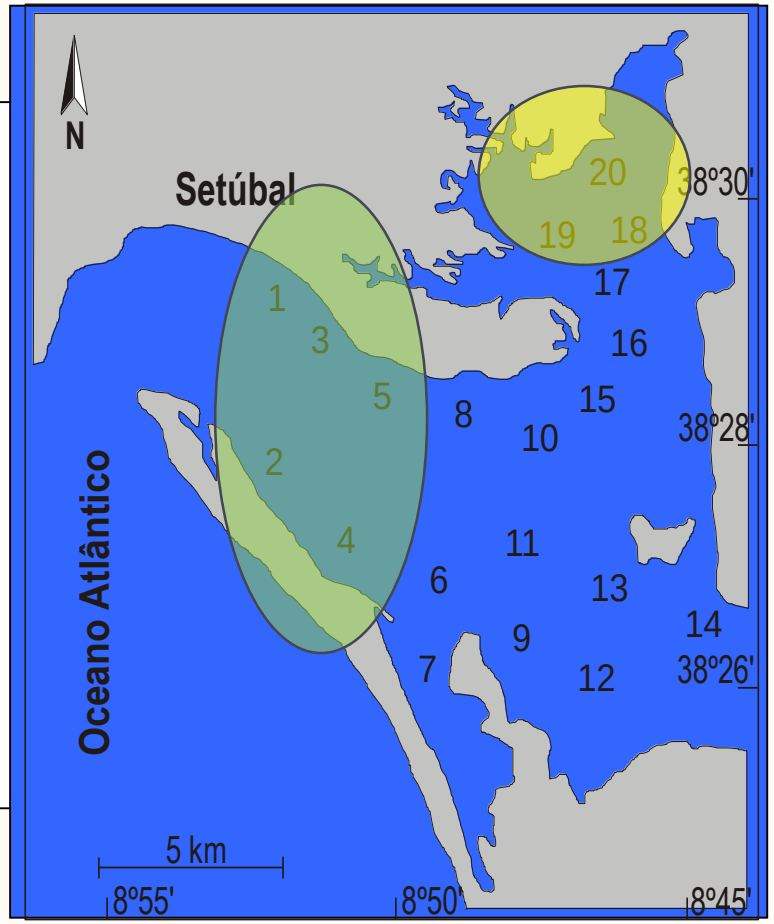
agrupamento



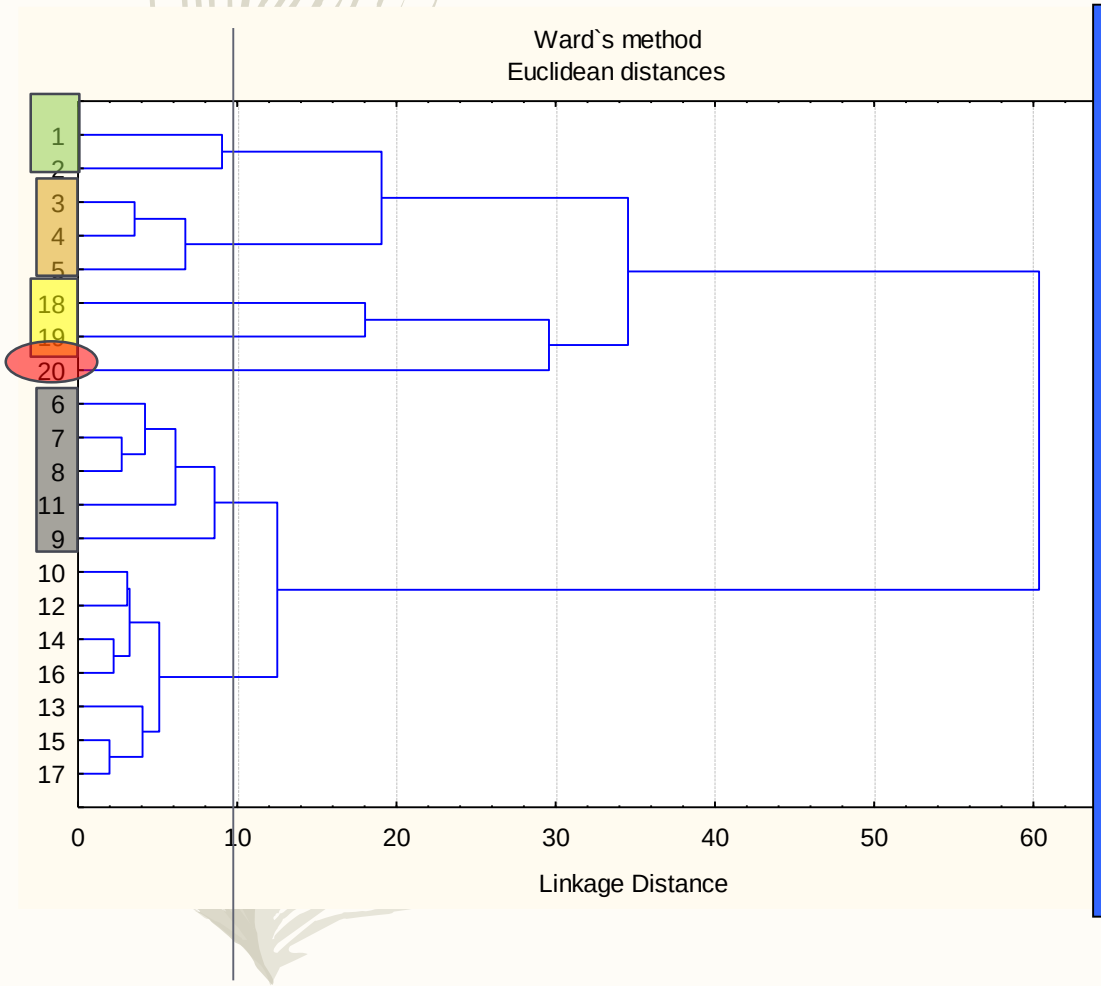


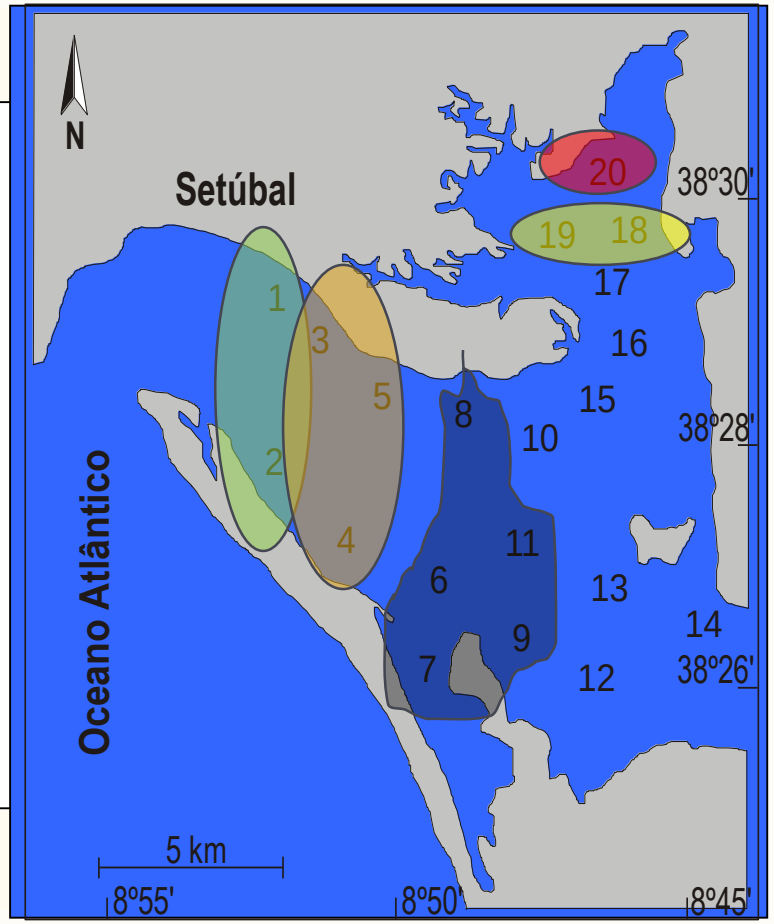
agrupamento





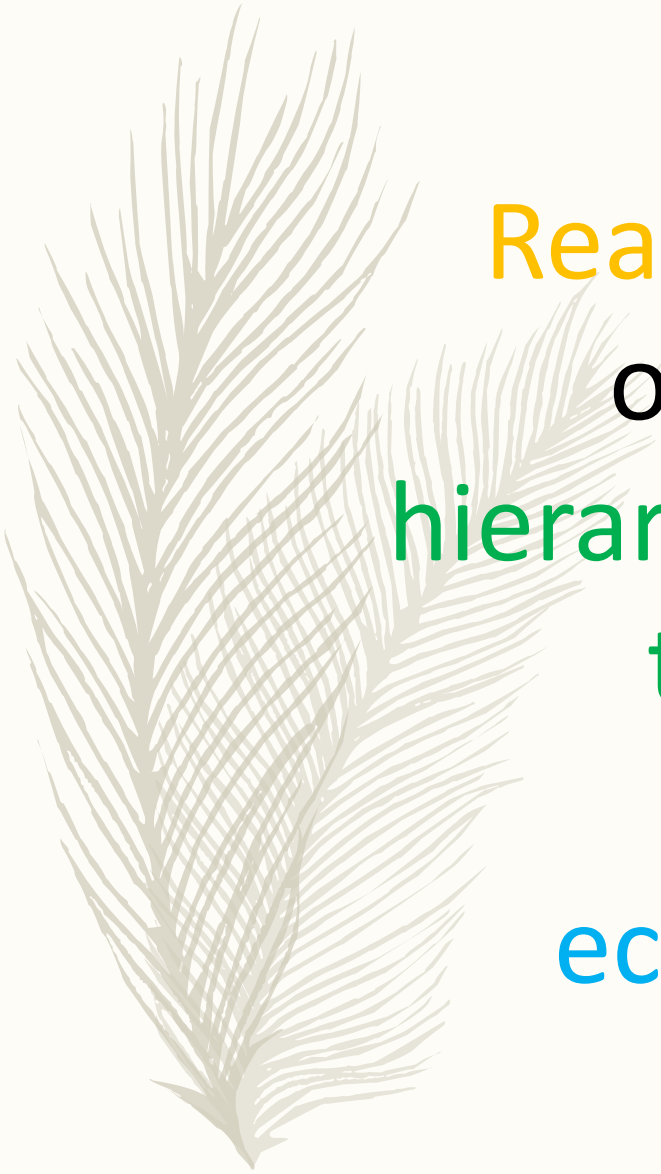
agrupamento





Remember: it is very hard to say if a given clustering procedure is any good, we can say whether one is adequate or not to represent a given data set, depending on the story we want…

1. to tell !
2. to sell !



Real life examples

of the use of

hierarchical clustering techniques

over

ecological data

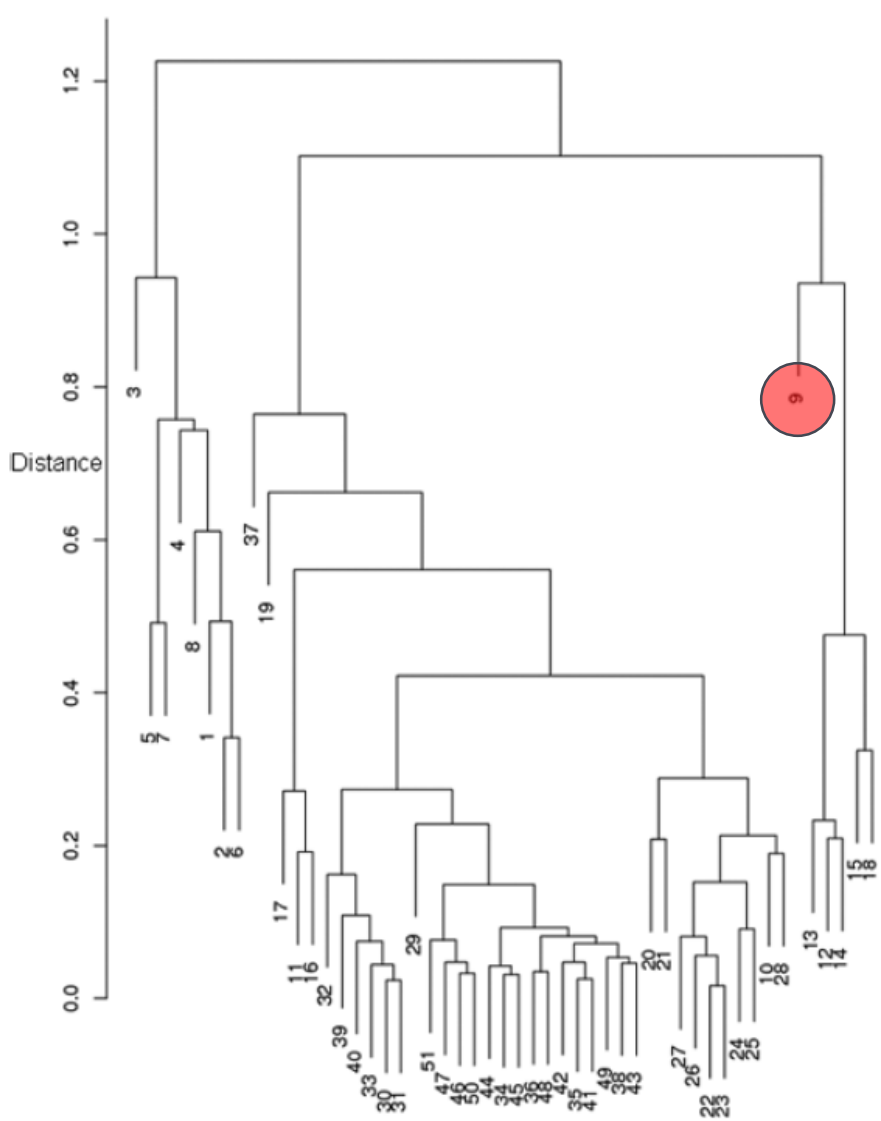


Fig. 7. Dendrogram showing hierarchical clustering of the PCA scores for the whale dives. Dive 9 occurred during ensonification.

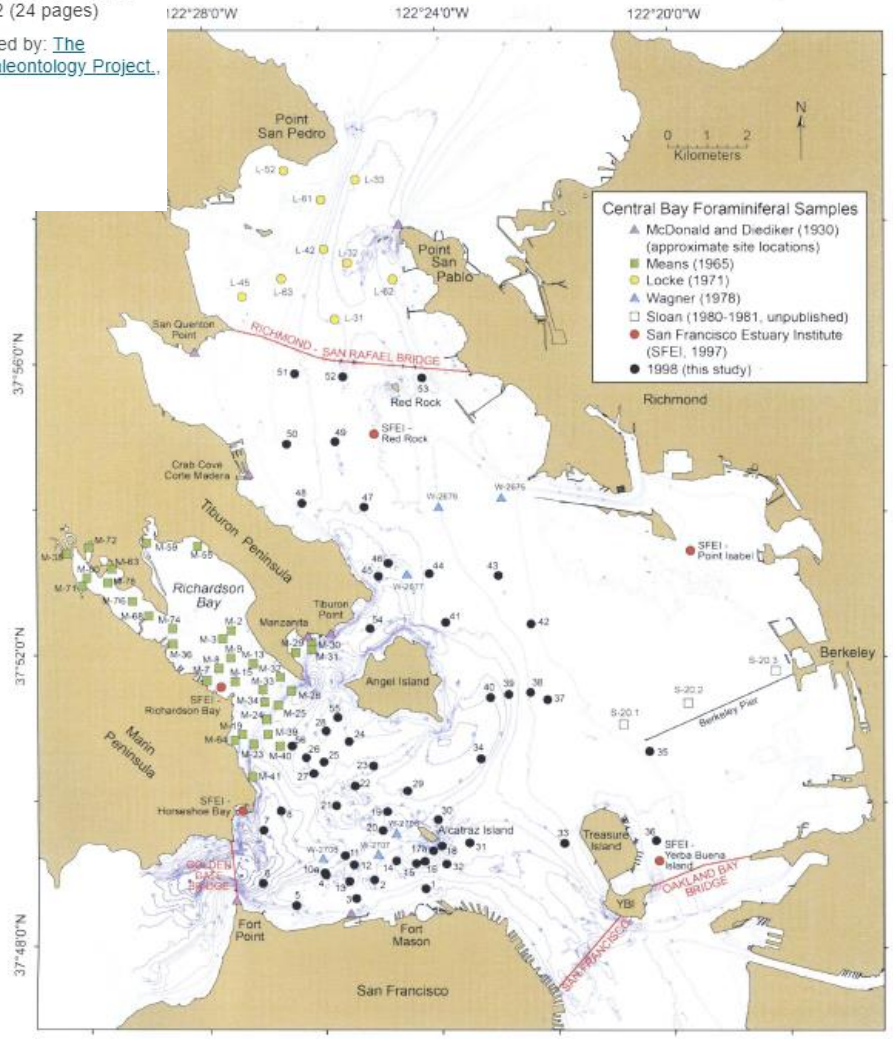
Walker, C. G.; MacKenzie, M. L.; Donovan, C. R.; Kidney, D.; Quick, N. J. & Hastie, G. D. 2011 Classification of animal dive tracks via automatic landmarking, principal components analysis and clustering *Ecosphere* **2**: art92

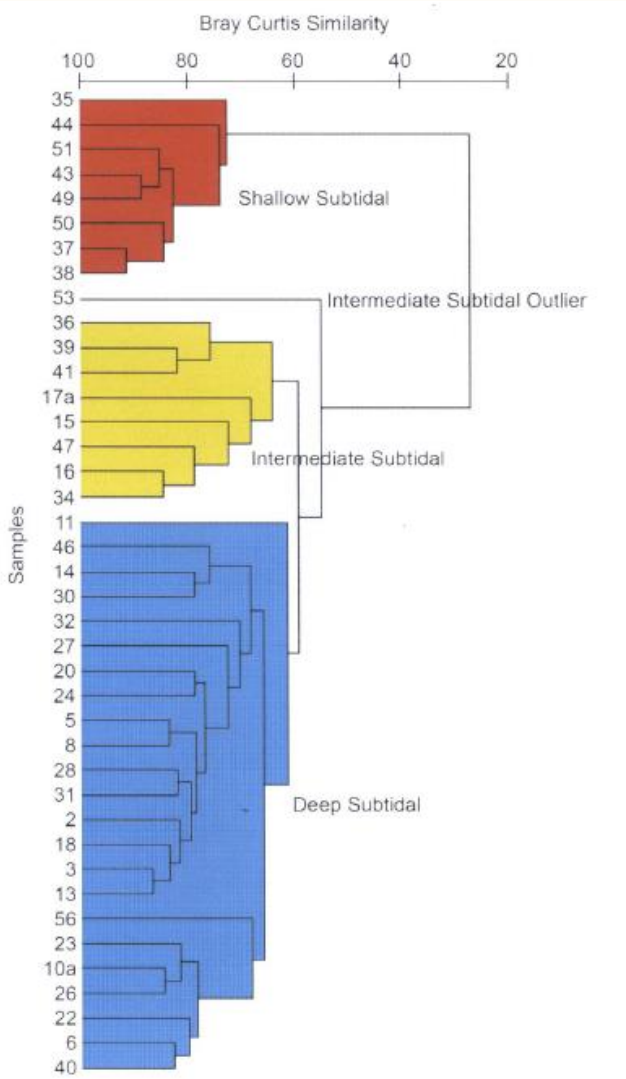
Late 20th Century benthic foraminiferal distribution in Central San Francisco Bay, California: Influence of the Trochammina hadai invasion

Mary McGann

Micropaleontology
Vol. 60, No. 6 (2014), pp. 519-542 (24 pages)

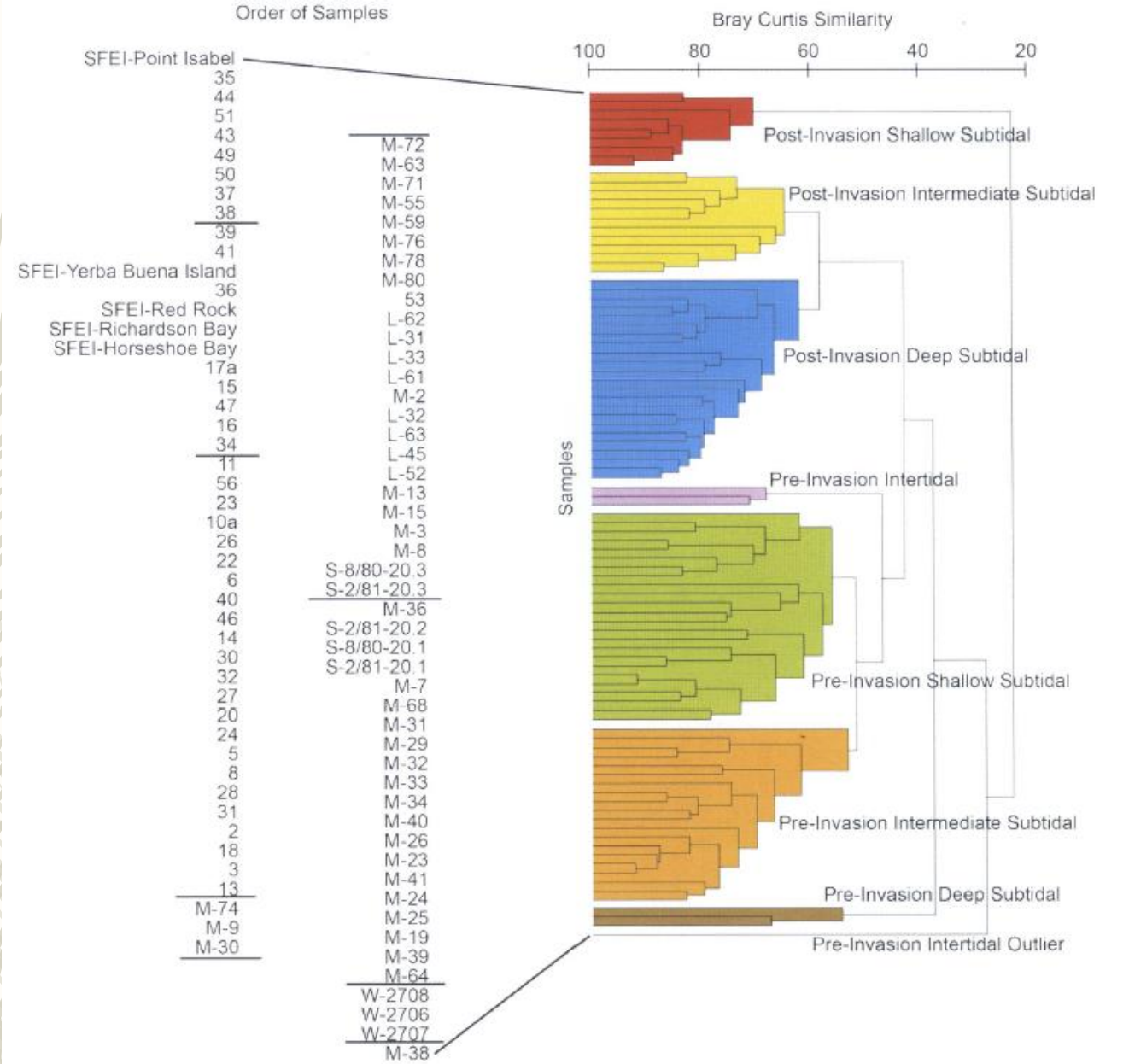
Published by: The Micropaleontology Project, Inc.





TEXT-FIGURE 8
Dendogram of the Q-mode cluster analysis of the 1998 Central Bay samples based on the quantitative foraminiferal abundances (in percent frequency). Three biofacies and one outlier are recognized.

When looking at a given time point, habitat (i.e. spatial) differences emerge

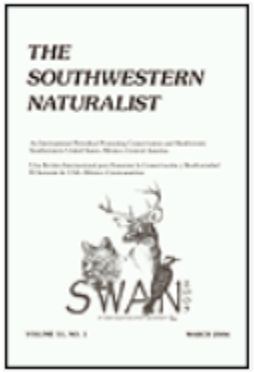


TEXT-FIGURE 10
Dendogram of the Q-mode cluster analysis of the 1965 to 1998 Central Bay samples based on the quantitative foraminiferal abundances (in percent frequency). Sites includes those collected in 1998 (this study), as well as those of Means (M-; 1965), Locke (L-; 1971), Wagner (W-; 1978), Sloan (S-; 1980-1981, unpublished data), and SFEI (SFEI-; McGann and Sloan 1999). Seven biofacies and one outlier are recognized.

When looking at data over time, invasion related patterns (i.e. temporal) differences emerge

TAXONOMIC ASSESSMENT OF THE SUBSPECIFIC STATUS OF PHRYNOSOMA ORBICULARE (SAURIA: PHRYNOSOMATIDAE) IN THE SOUTHERN PORTION OF ITS DISTRIBUTION

Ruth Moreno Barajas, Felipe Rodríguez-Romero, Alma S. Velázquez Rodríguez and Fausto R. Méndez de la Cruz



The Southwestern Naturalist Vol. 58, No. 4 (DECEMBER 2013), pp. 459-464 (6 pages)

Published by: Southwestern Association of Naturalists

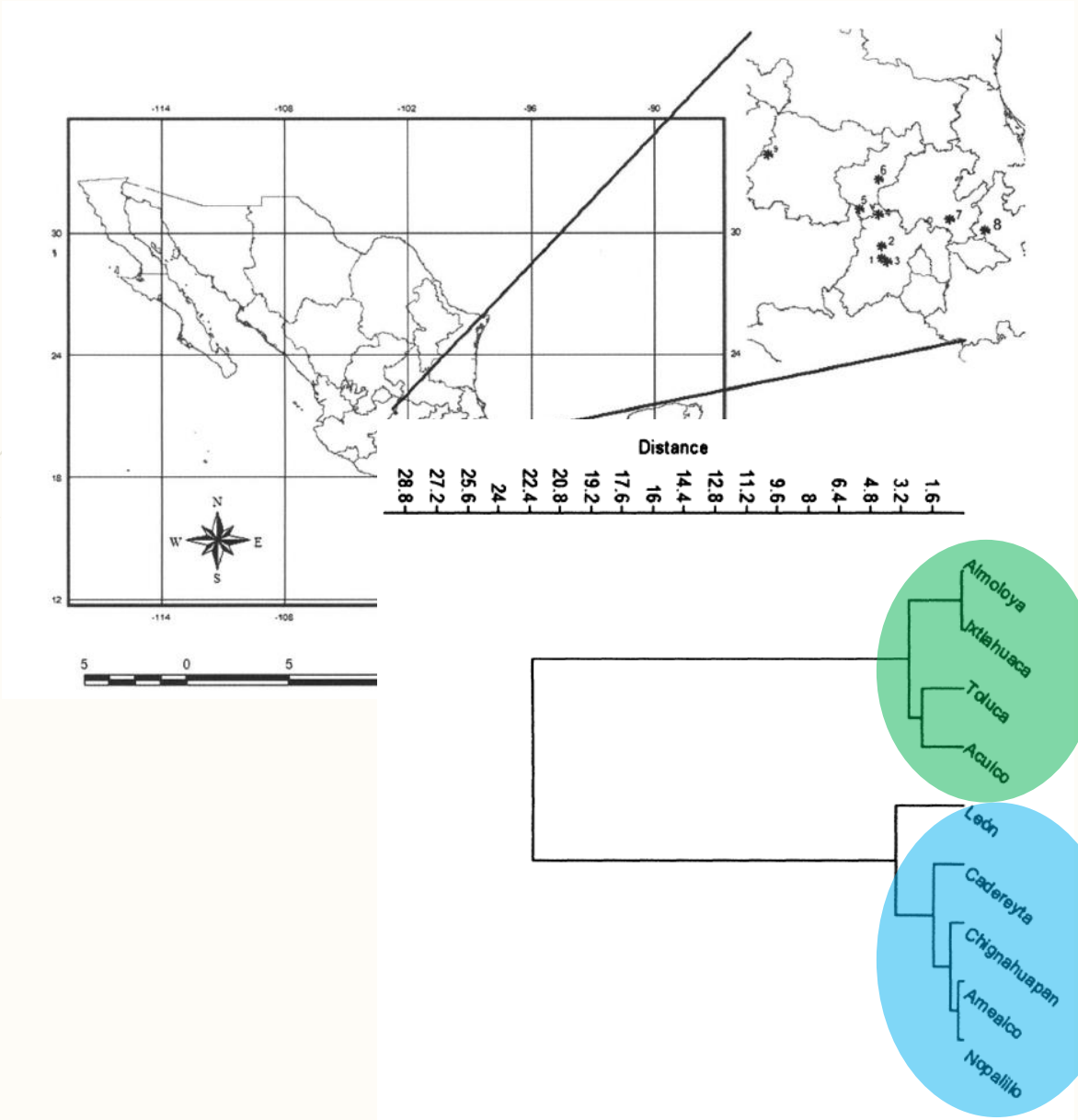


FIG. 2—Distance phenogram resulting from cluster analysis of *Phrynosoma orbiculare* from nine localities (defined in Fig. 1) in the southern portion of its distribution in Mexico. Cophenetic correlation = 0.971.

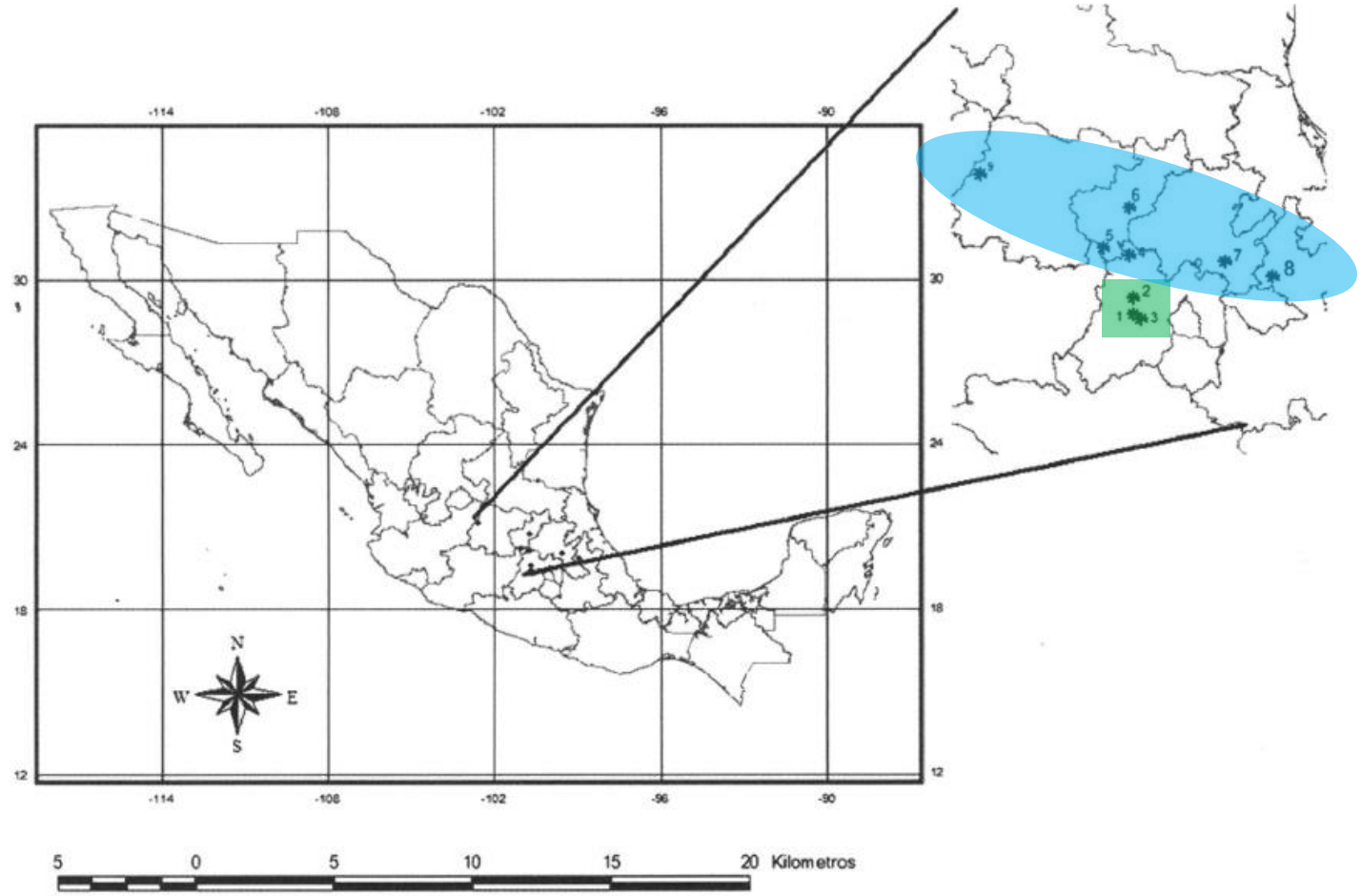


FIG. 1—Localities for *Phrynosoma orbiculare* examined in a taxonomic assessment of the subspecific status of the species in the southern portion of its distribution in Mexico: 1) Almoloya de Juárez, Mexico; 2) Ixtlahuaca de Rayón, Mexico; 3) Toluca de Lerdo, Mexico; 4) Aculco de Espinoza, Mexico; 5) Amealco, Queretaro; 6) Cadereyta, Queretaro; 7) Nopalillo, Hidalgo; 8) Chignahuapan, Puebla; 9) León, Guanajuato.

Incomplete song divergence between recently diverged taxa: syllable sharing by Orchard and Fuertes' orioles

Natasha D. G. Hagemeyer, Rachel J. Sturge, Kevin E. Omland and J. Jordan Price

Journal of Field Ornithology
Vol. 83, No. 4 (DECEMBER 2012), pp. 362-371 (10 pages)

Published by: Wiley on behalf of Association of Field Ornithologists

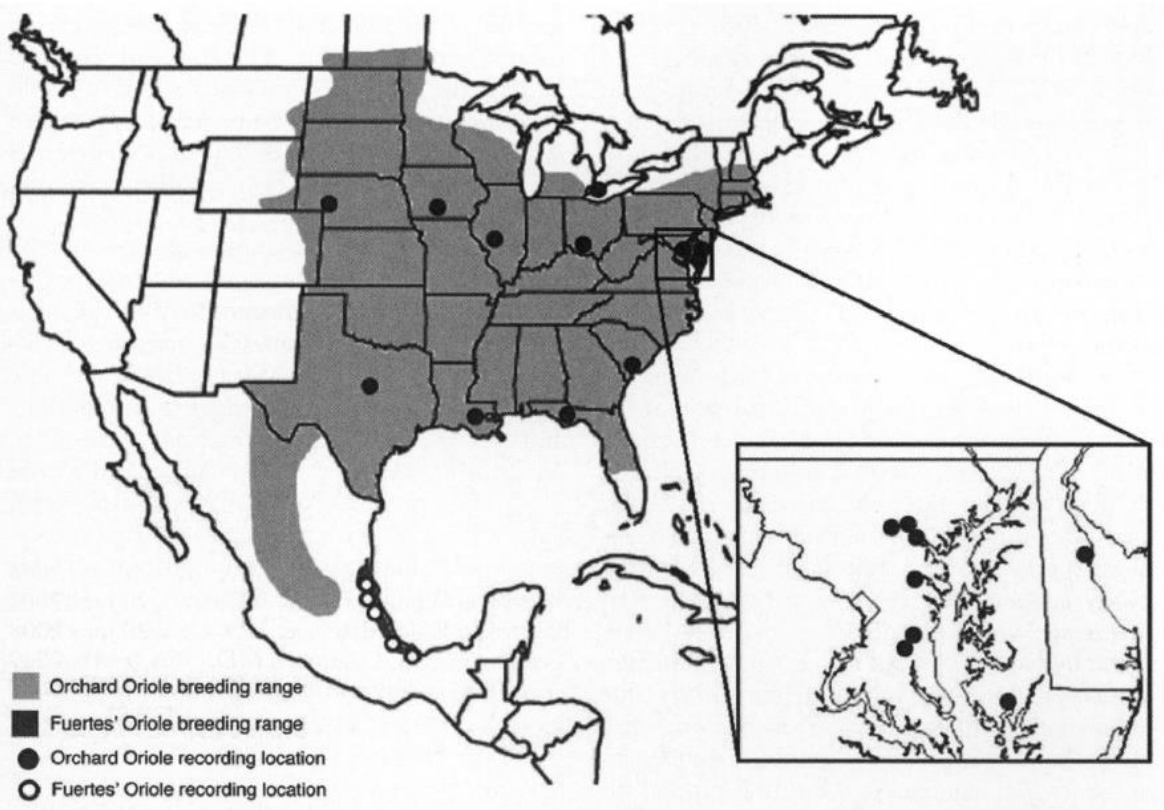


Fig. 1. Recording locations for Orchard Orioles (filled circles) and Fuertes' Orioles (open circles) across their ranges in eastern North America. The breeding range of Orchard Orioles is indicated in light gray and the breeding range of Fuertes' Orioles in dark gray (adapted from Baker et al. 2003). Inset: Locations where Orchard Orioles were recorded in Maryland and Delaware.

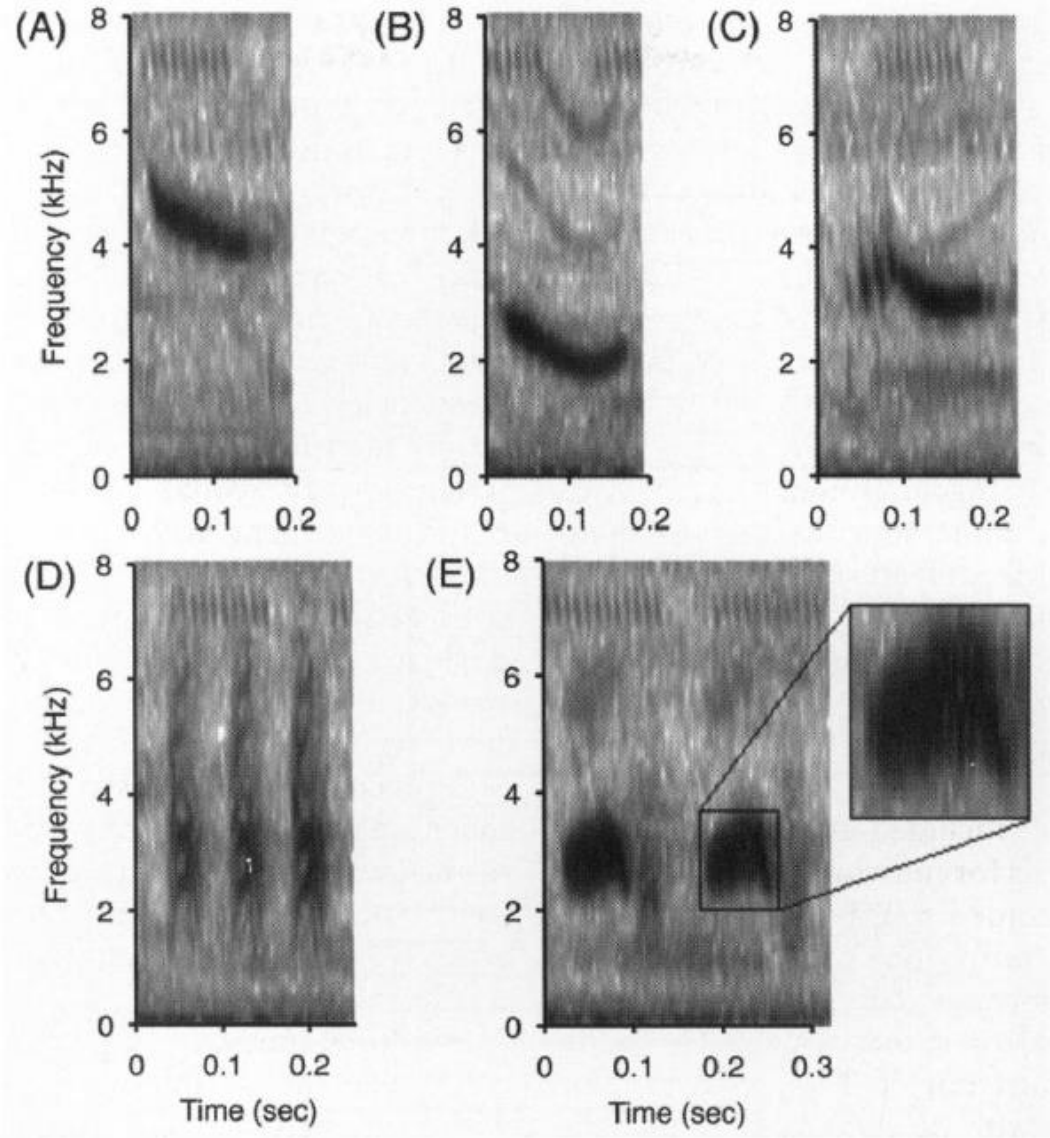


Fig. 2. Sound spectrograms showing examples of syllable types: (A and B) syllables comprising single whistled notes, (C) a syllable including multiple whistled notes, (D) syllables with visible harmonics, and (E) buzzed syllables with rapid frequency modulation (magnified in inset). Top syllables (A–C) are song-type syllables that occurred only in songs, whereas the bottom syllables (D and E) are call-type syllables that sometimes also occurred as calls. These examples are from a single Orchard Oriole song, but syllables with these characteristics occurred in both taxa.

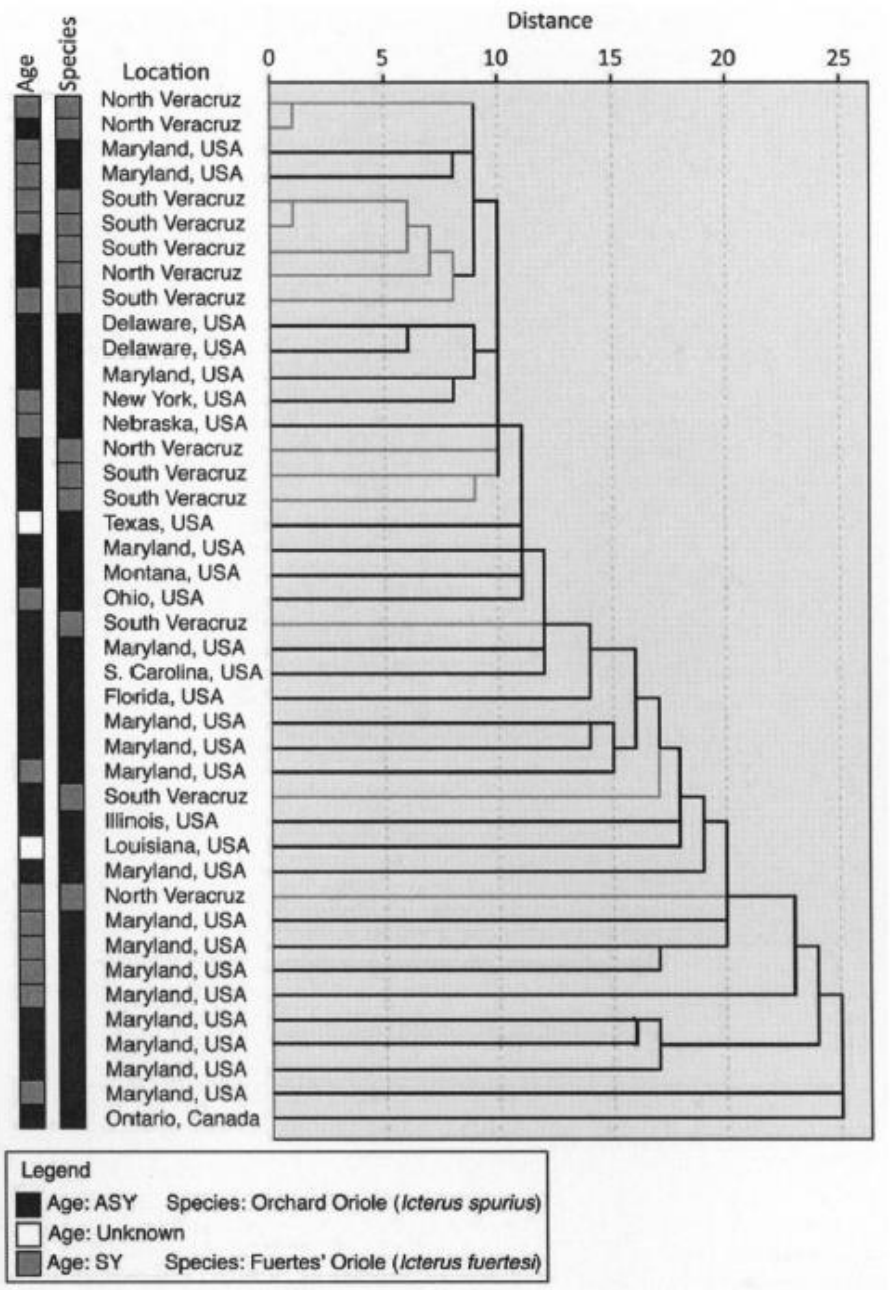


Fig. 3. Hierarchical cluster analysis dendrogram based on the presence or absence of 529 syllable types among SY (second-year) and ASY (after second year) Orchard Orioles ($N = 29$) and Fuertes' Orioles ($N = 13$). Distance between individuals on the tree reflects fusion values based on the presence or absence of syllable types in each bird's songs. Individuals from the two oriole taxa were intermixed.

No separation by either age or species !

Using hand proportions to test taxonomic boundaries within the Tupaia glis species complex (Scandentia, Tupaiidae)

Eric J. Sargis, Neal Woodman, Aspen T. Reese and Link E. Olson

Journal of Mammalogy
Vol. 94, No. 1 (February 2013), pp. 183-201 (19 pages)

Published by: American Society of Mammalogists

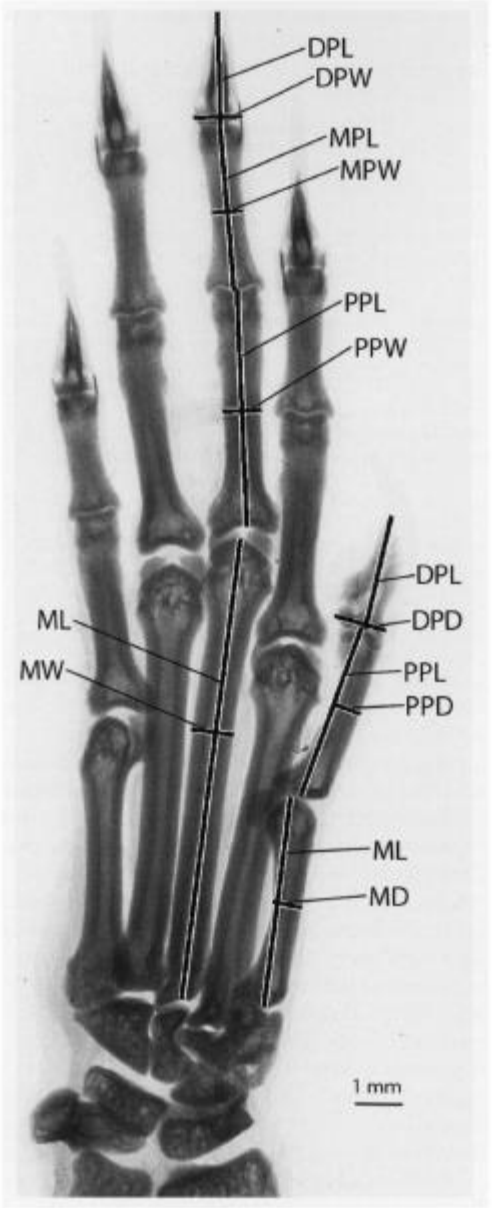


FIG. 2.—Digital X-ray of the right manus (plantar view) of *Tupaia belangeri* (USNM 201431), illustrating the measurements used in this study. DPD = distal phalanx depth; DPL = distal phalanx length; DPW = distal phalanx width; MD = metacarpal depth; ML = metacarpal length; MW = metacarpal width; MPL = middle phalanx length; MPW = middle phalanx width; PPD = proximal phalanx depth; PPL = proximal phalanx length; PPW = proximal phalanx width. Original negative was converted to a positive image.

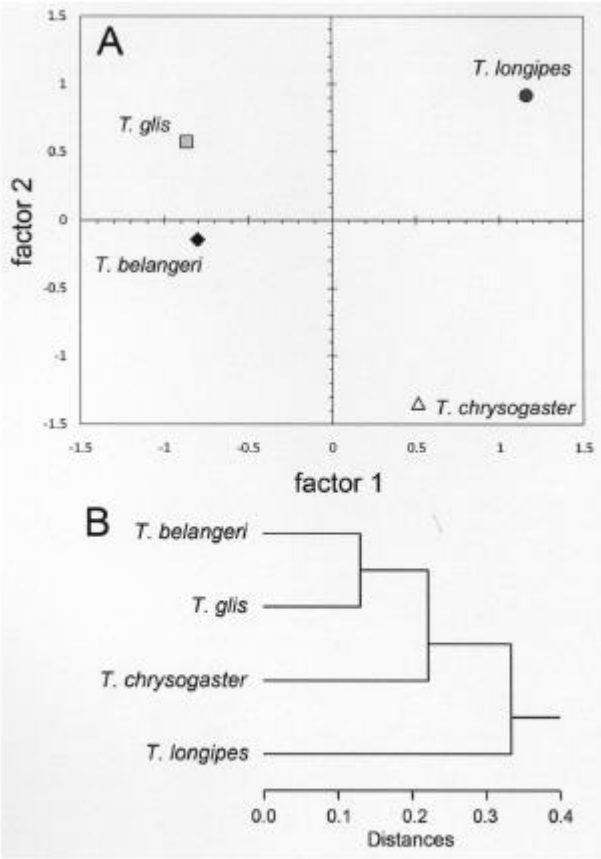


FIG. 3.—Plots illustrating the distinctiveness of 4 currently recognized species, *Tupaia belangeri*, *T. chrysogaster*, *T. glis*, and *T. longipes*. A) Plot of factor scores on first 2 axes from principal components analysis of means of 8 variables from ray IV (Table 3). All 4 taxa plot in different quadrants. B) Phenogram from cluster analysis of 38 variables from all 5 rays.

Depth distribution of epilithic cyanobacteria and pigments in a mountain lake characterized by marked water-level fluctuations

Marco Cantonati, Graziano Guella, Jiří Komárek and Daniel Spitale

Freshwater Science
Vol. 33, No. 2 (June 2014), pp. 537-547 (11 pages)

Published by: The University of Chicago Press on behalf of Society for Freshwater Science

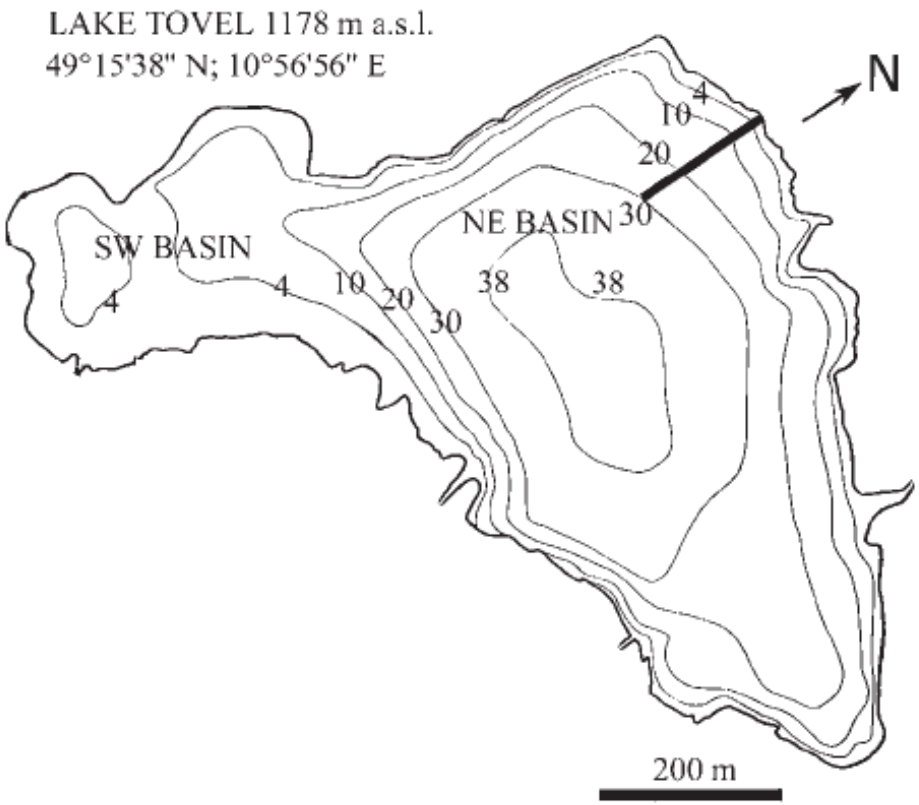


Figure 1. Bathymetric map of Lake Tovel in northern Italy (southeastern Alps). The depth transect sampled is represented by the heavy black line.

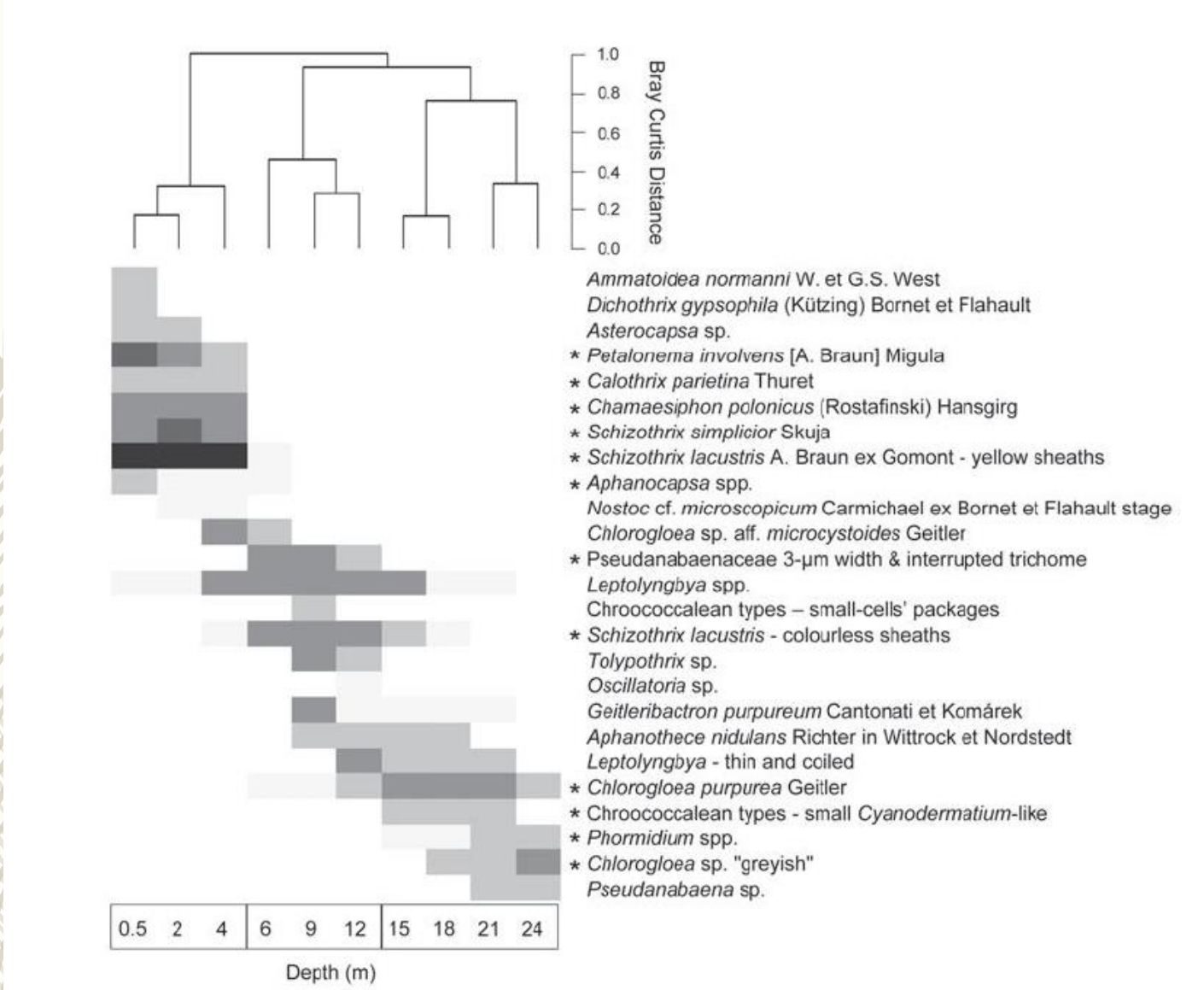


Figure 2. Depth distribution of epilithic cyanobacteria in Lake Tovel, and cluster analysis of the assemblages found at the different depths. Gray tones (from light to dark) identify the semiquantitative scale (1–5) used to estimate the abundance of the taxa.
* indicates taxa selected by the indicator value analysis (IndVal) as indicators for the 3 depth zones (delimited by vertical bars on the depth scale).

Phylogeny and a revised classification of the Chinese species of Nyssa (Nyssaceae) based on morphological and molecular data

Nian Wang, Richard I. Milne, Frédéric M.B. Jacques, Bao-Ling Sun, Chang-Qin Zhang and Jun-Bo Yang

Taxon
Vol. 61, No. 2 (April 2012), pp. 344-354 (11 pages)

Published by: International Association for Plant Taxonomy (IAPT)

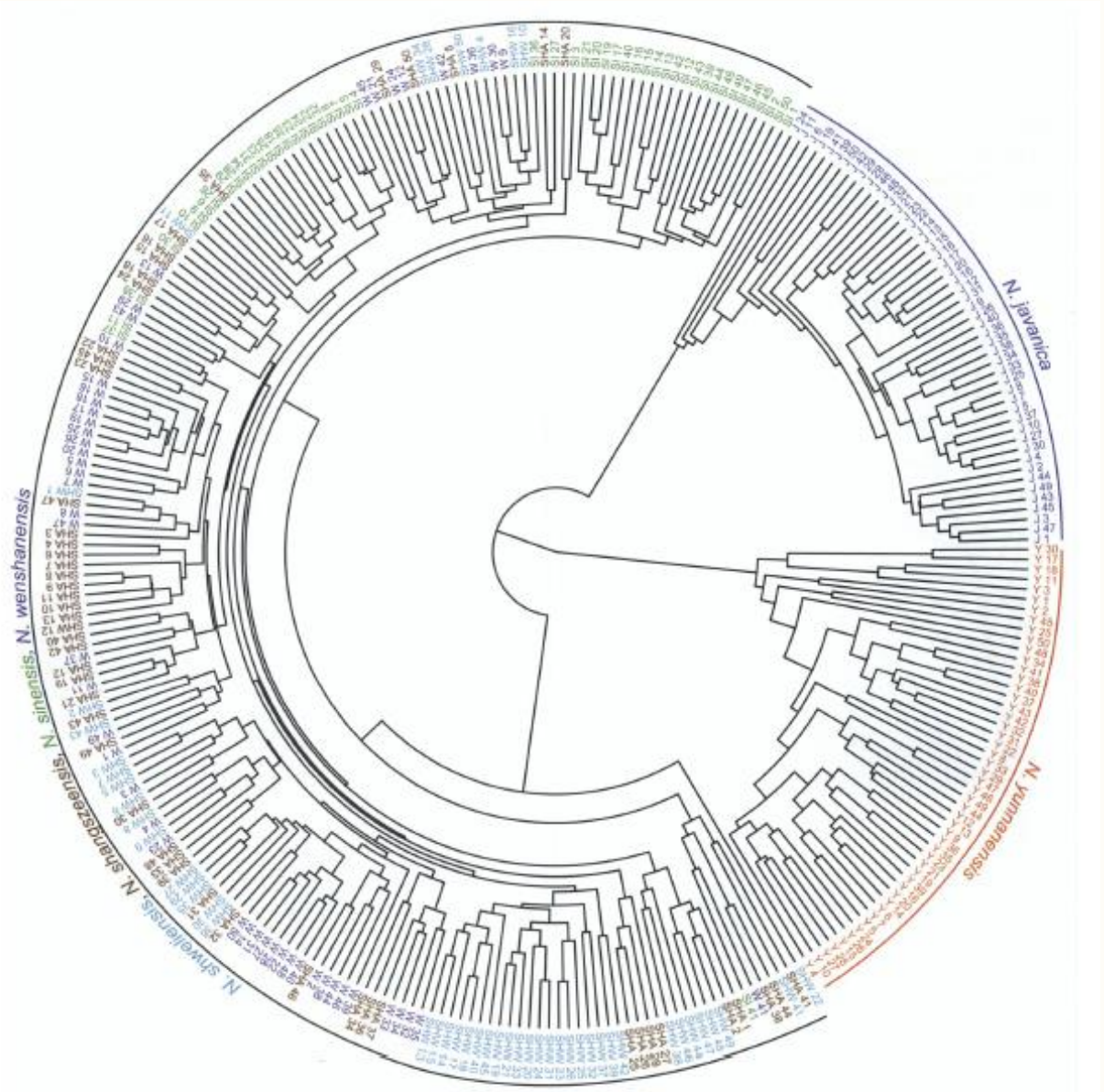


Fig. 2. Cluster analysis using UPGMA of morphological measurements of 52 characters of 50 accessions each of six *Nyssa* species. Color of accessions corresponds to color of species names.

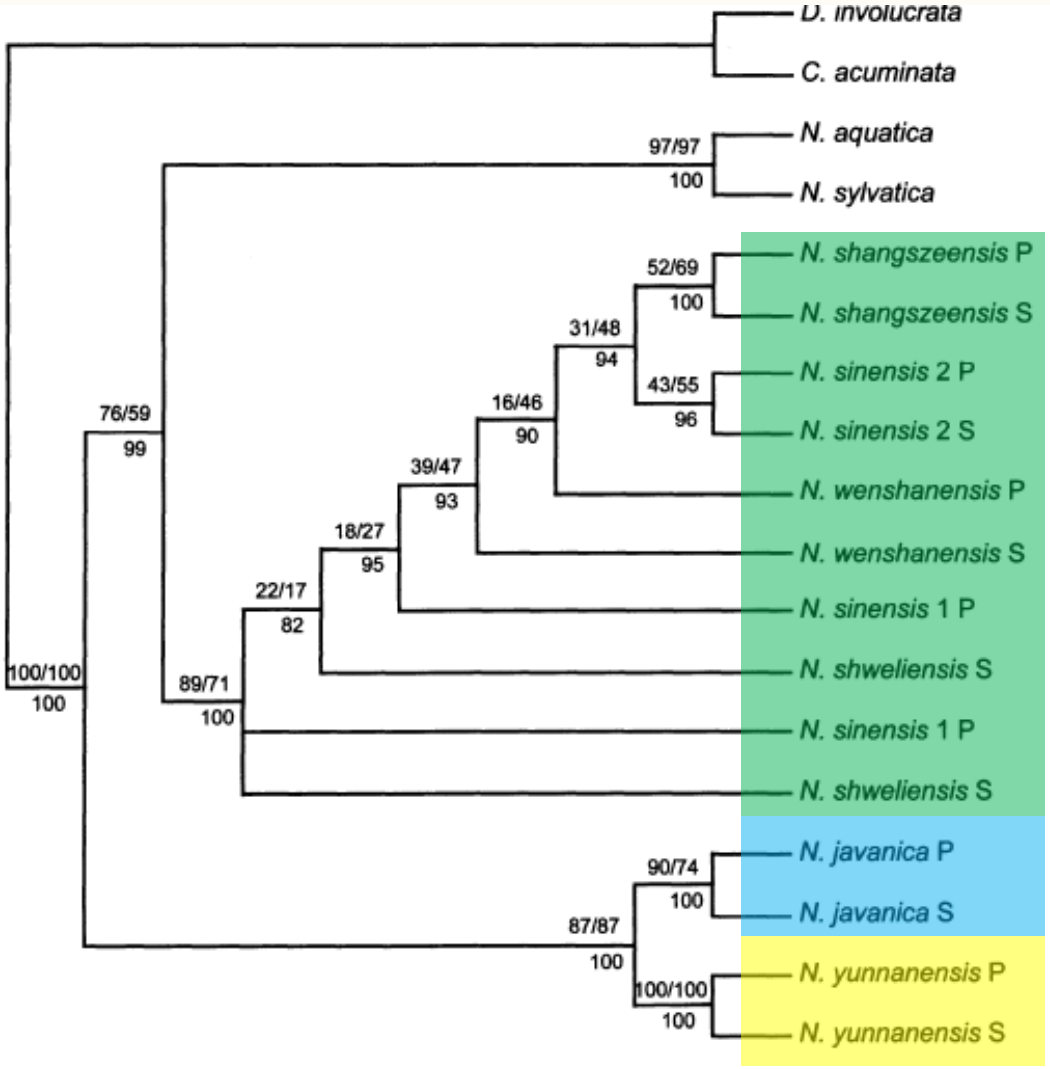
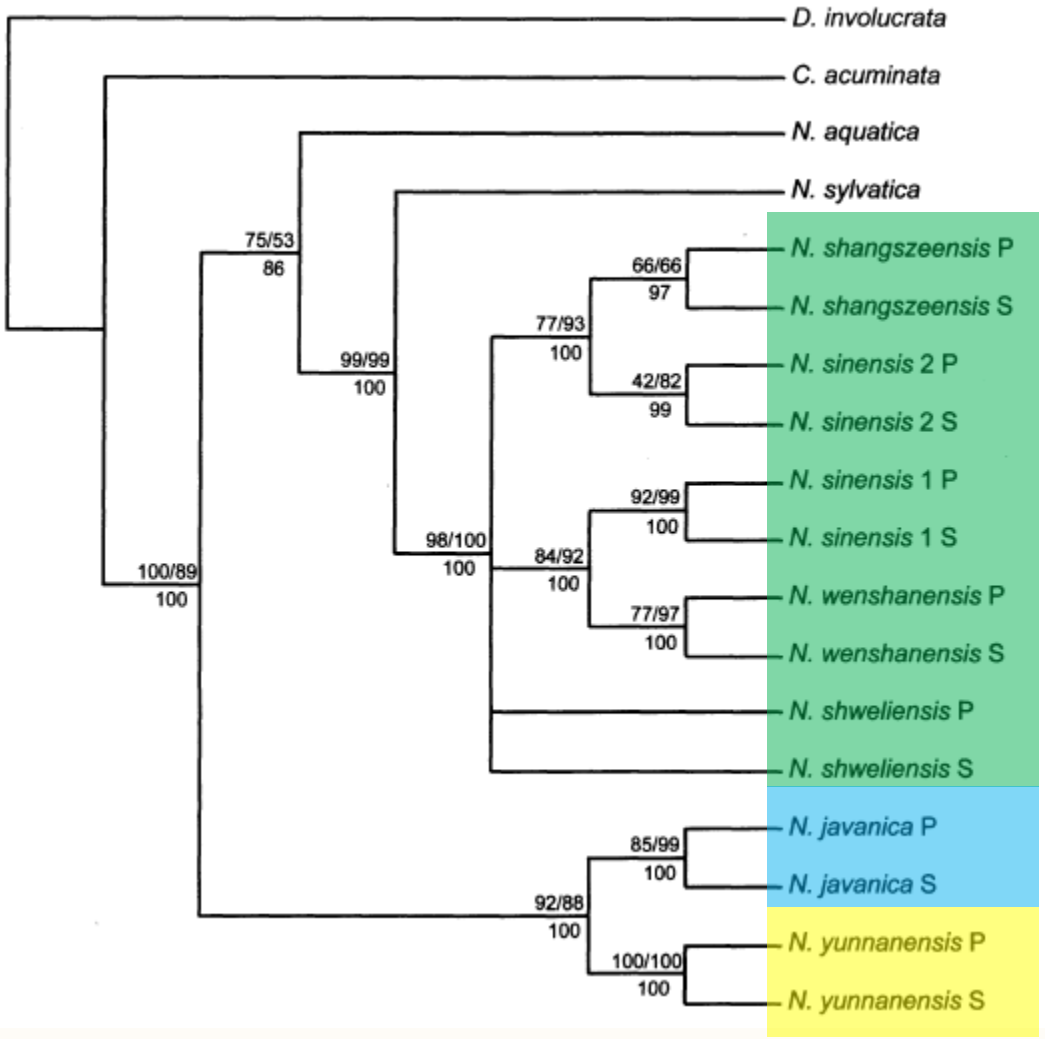


Fig. 5. Strict consensus tree based on the combined sequences of five cpDNA regions (*trnH-psbA*, *rps16r-f*, *trnL-rps32F*, *trnS-G*, *trnL-F*). Figures above branches show maximum parsimony bootstrap/ maximum likelihood bootstrap support; figures below branches show Bayesian posterior probability, all as percentages.

Fig. 4. Strict consensus tree based on ITS sequences. Figures above branches show maximum parsimony bootstrap/ maximum likelihood bootstrap support; figures below branches show Bayesian posterior probability, all as percentages.

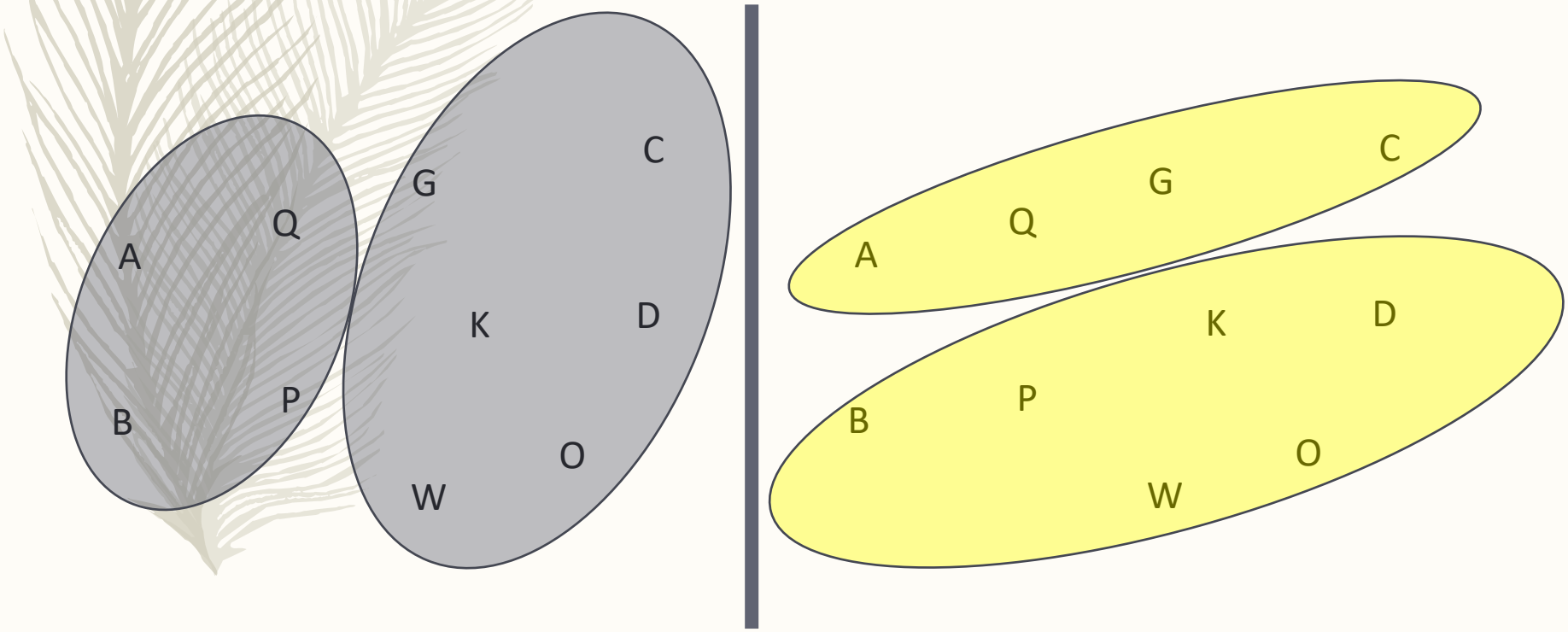


Métodos não hierárquicos

- O número de grupos a constituir é fixado a priori;
- O procedimento é sequencial e inicia-se com a definição de uma semente (valor inicial, nota que como o procedimento tem uma base aleatória, podem obter resultados diferentes se o correrem várias vezes!)
- Os objectos vão sendo agrupados através de vários métodos:
  - » Limiares sequenciais
  - » Limiares paralelos
  - » Optimização

A ideia fundamental que procuramos grupos que maximizem a variabilidade entre grupos enquanto que minimizam a variabilidade dentro dos grupos.

Dada uma medida de variação entre e dentro dos grupos, podemos implementar um método de força bruta.



Do exercicio 3, FT9

bentcnh2=kmeans(dados,2)

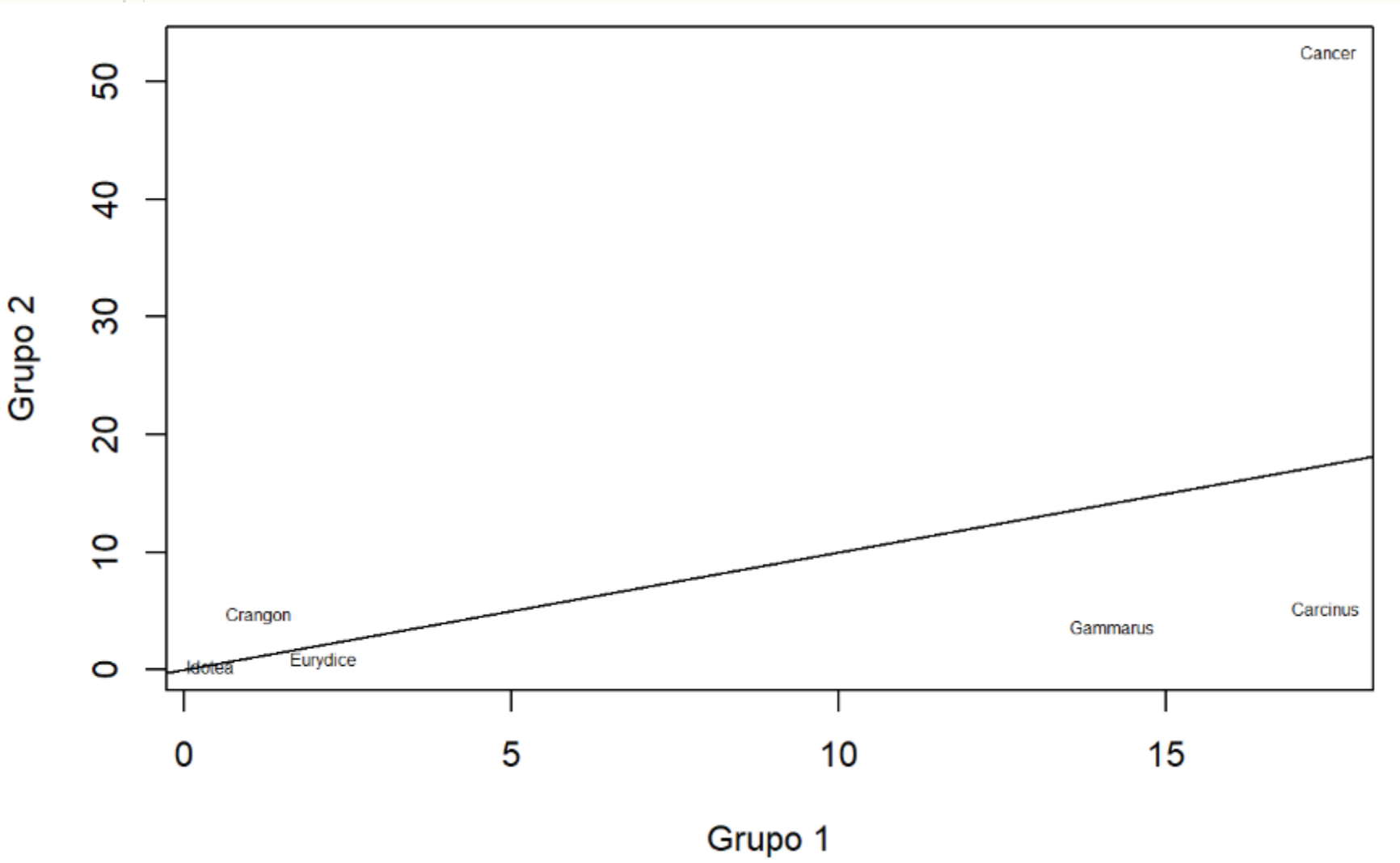
Mean of the variables per group
(i.e. which variables distinguish the groups)

```
bentcnh2
```

```
## K-means clustering with 2 clusters of sizes 9, 6
##
## Cluster means:
##   Eurydice    Idotea Gammarus  Crangon   Cancer Carcinus
## 1 2.133333 0.4111111 14.17778 1.133333 17.47778 17.43333
## 2 0.850000 0.3833333  3.70000 4.633333 52.55000  5.30000
##
## Clustering vector:
##  [1] 1 1 1 2 1 1 1 1 1 2 2 2 2 2 1
##
## Within cluster sum of squares by cluster:
## [1] 4591.3000  833.9717
##  (between_SS / total_SS =  49.9 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

Group

Between (group) variance must be large and within (group) variance therefore small)


Cancer
Crangon
Gammarus
Carcinus
Idotea
Eurydice
Grupo 2
0
10
20
30
40
50
0
5
10
15
Grupo 1

bentcnh3=kmeans(dados,3)

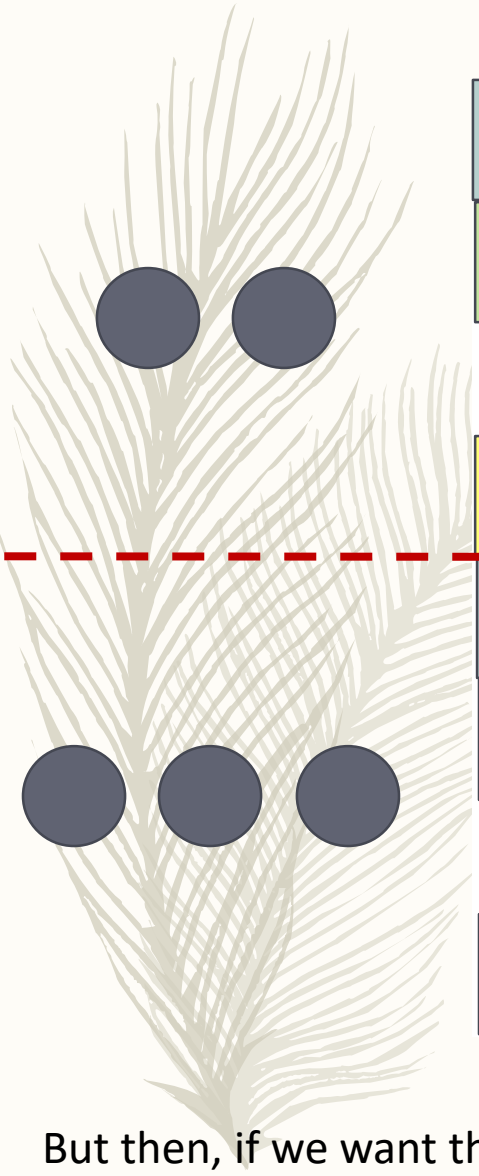
```
bentcnh3
```

```
## K-means clustering with 3 clusters of sizes 3, 6, 6
##
## Cluster means:
##   Eurydice    Idotea  Gammarus  Crangon   Cancer Carcinus
## 1 2.366667 0.2000000 37.666667 1.300000 12.00000     8.60
## 2 0.850000 0.3833333  3.700000 4.633333 52.55000     5.30
## 3 2.016667 0.5166667  2.433333 1.050000 20.21667    21.85
##
## Clustering vector:
##  [1] 1 1 1 2 3 3 3 3 3 2 2 2 2 2 3
##
## Within cluster sum of squares by cluster:
## [1]  229.8933  833.9717 1391.9083
##  (between_SS / total_SS =  77.3 %)
##
## Available components:
##
## [1] "cluster"      "centers"      "totss"        "withinss"
## [5] "tot.withinss" "betweenss"    "size"         "iter"
## [9] "ifault"
```

```
## K-means clustering with 2 clusters of sizes 9, 6
##
## Within cluster sum of squares by cluster:
## [1] 4591.3000  833.9717
##  (between_SS / total_SS =  49.9 %)
```

https://stats.stackexchange.com/questions/230989/clustering-k-means-alternatives-when-its-assumptions-do-not-hold

https://stats.stackexchange.com/questions/133656/how-to-understand-the-drawbacks-of-k-means

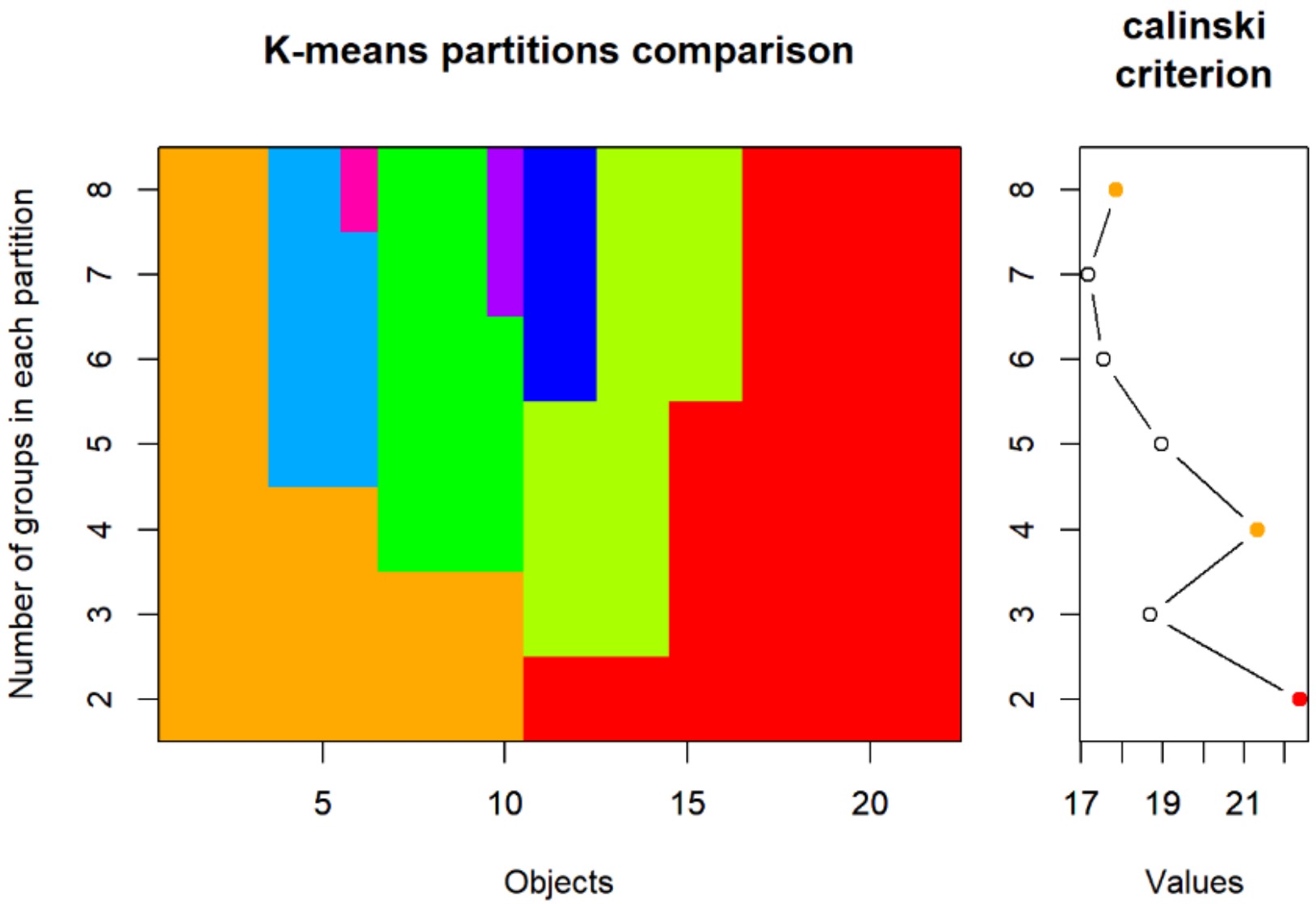


```
> cnh2$totss
[1] 1.722877
> cnh2$withinss
[1] 0.4011667 0.4121300
> cnh2$tot.withinss
[1] 0.8132967
> cnh2$betweenss
[1] 0.9095806
```

```
> cnh3$totss
[1] 1.722877
> cnh3$withinss
[1] 0.0587250 0.1096125 0.4121300
> cnh3$tot.withinss
[1] 0.5804675
> cnh3$betweenss
[1] 1.14241
```

But then, if we want the largest variance across groups and the smallest within groups, what is the optimal?

```
km2to8=cascadeKM(habs[,-1],inf.gr = 2,sup.gr = 8,criterion = "calinski")
plot(km2to8,sortg=TRUE)
```



```
km2to8=cascadeKM(habs[,-1],inf.gr = 2,sup.gr = 8,criterion = "ssi")
plot(km2to8,sortg=TRUE)
```