Science des données II : tp3

Classification hiérarchique

Guyliann Engels & Philippe Grosjean

Université de Mons, Belgique
Laboratoire d'Écologie numérique des Milieux aquatiques

http://biodatascience-course.sciviews.org
sdd@sciviews.org

UMONS

Classification hiérarchique

En partant d'une matrice de (dis)similarité, la classification hiérarchique permet de réaliser des regroupements. Ex: considérant 6 stations parmi les 68 que comporte le jeu de données `marphy`, pouvons-nous réaliser des regroupements selon les conditions physico-chimiques de l'eau.

Dans le cadre de ce TP, les méthodes agglomératives sont employées via :

- Liens simples
- Liens complets
- Liens moyens
- Ward

UMONS

Sélection de 6 stations du jeu de données marphy.

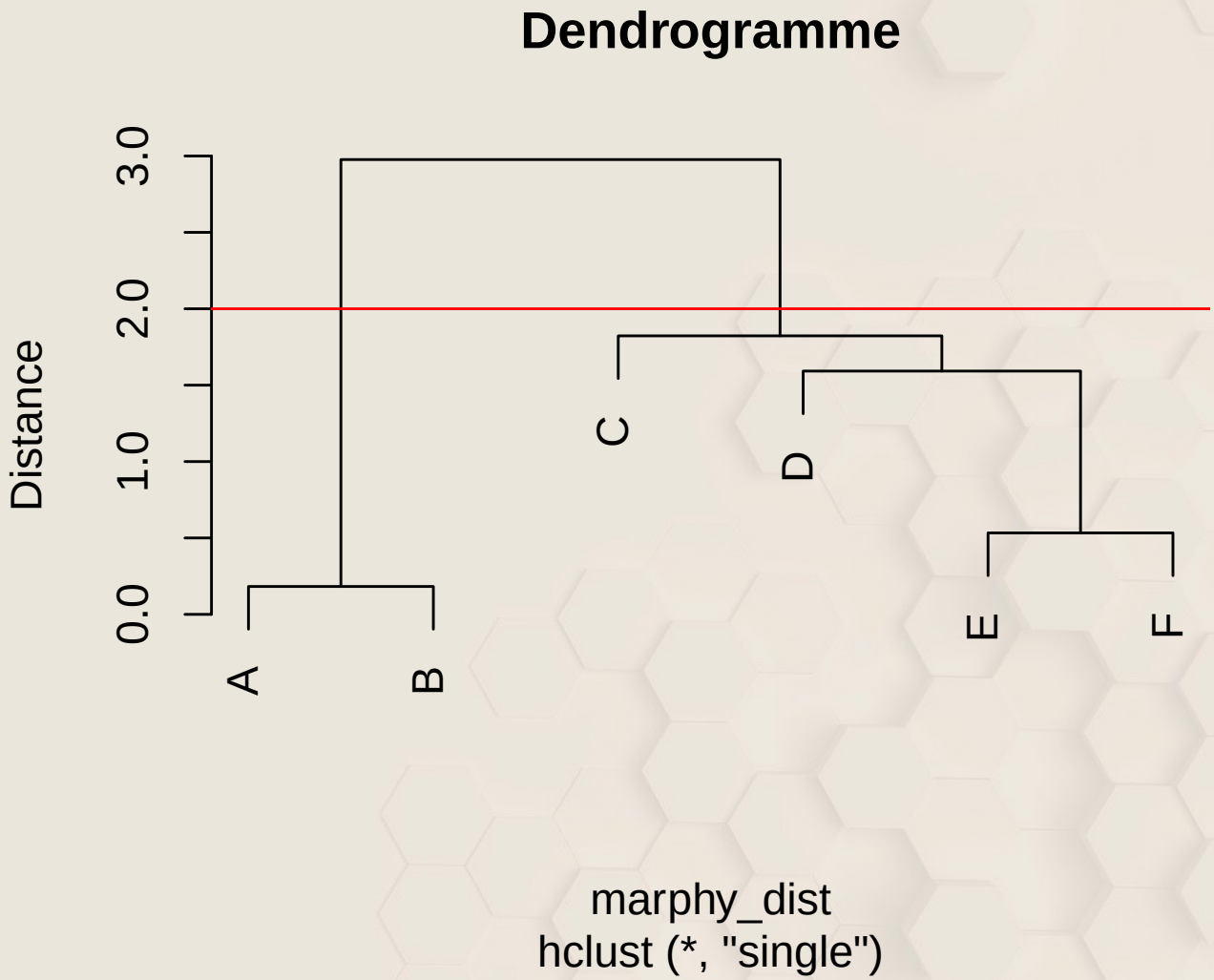
	Temperature	Salinity	Fluorescence	Density
A	13.082	38.166	0.958	28.8436
B	13.070	38.162	0.931	28.8430
C	12.868	38.283	1.552	28.9787
D	12.993	38.372	1.477	29.0218
E	13.062	38.412	0.993	29.0384
F	13.025	38.409	1.064	29.0438

UMONS

Matrice de distance réalisée avec la distance euclidienne :

	A	B	C	D	E	F
A	0.00	0.18	3.85	3.39	2.98	3.09
B	0.18	0.00	3.82	3.42	3.00	3.10
C	3.85	3.82	0.00	1.82	3.41	2.94
D	3.39	3.42	1.82	0.00	1.99	1.59
E	2.98	3.00	3.41	1.99	0.00	0.53
F	3.09	3.10	2.94	1.59	0.53	0.00

UMONS


Dendrogramme
Distance
0.0
1.0
2.0
3.0
A
B
C
D
E
F
marphy_dist
hclust (*, "single")

UMONS

Transect : marphy & marbio

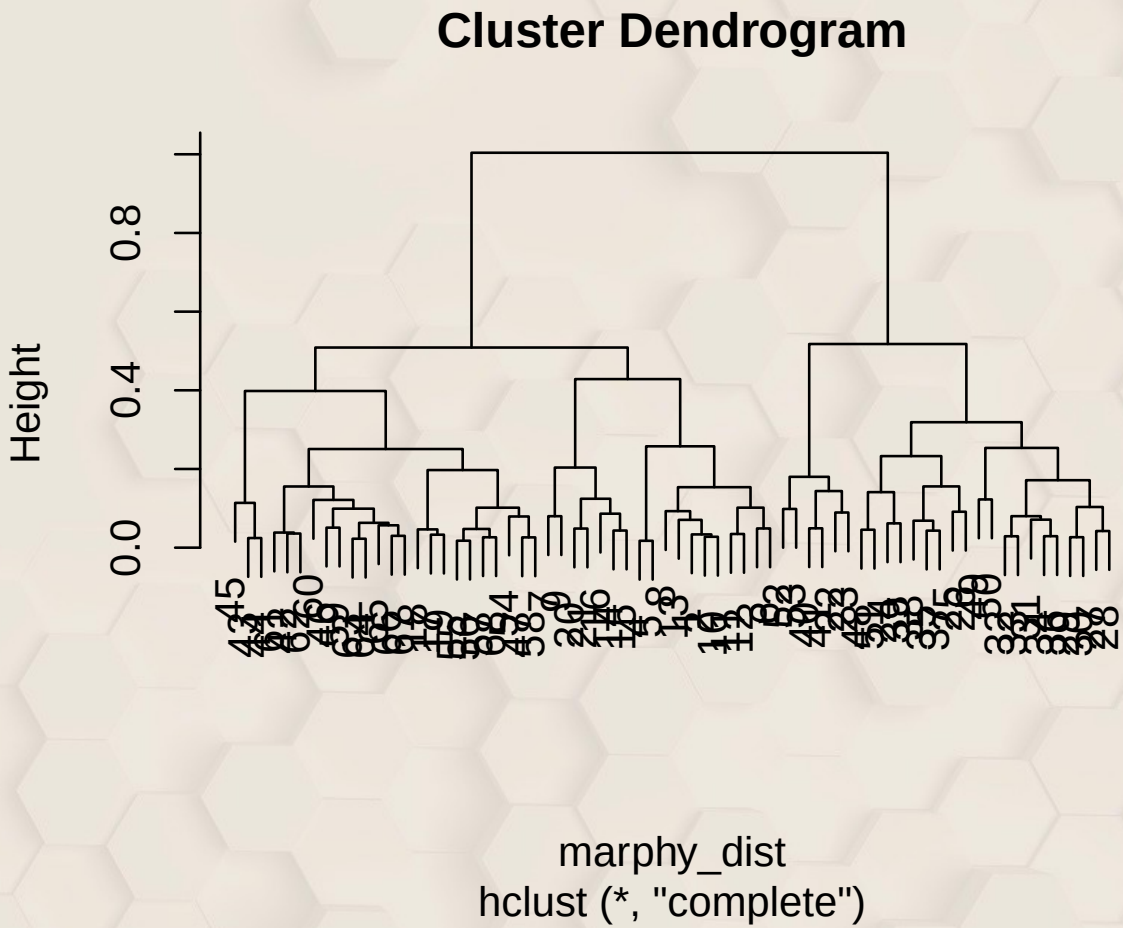
Transect entre Nice et Calvi

- Employez la fonction **`hclust()`** pour réaliser votre classification
- Employez la fonction **`plot()`** pour afficher votre classification

UMONS

Consignes

- Employez la distance euclidienne pour réaliser votre matrice.
- Employez la méthode des liens complets.

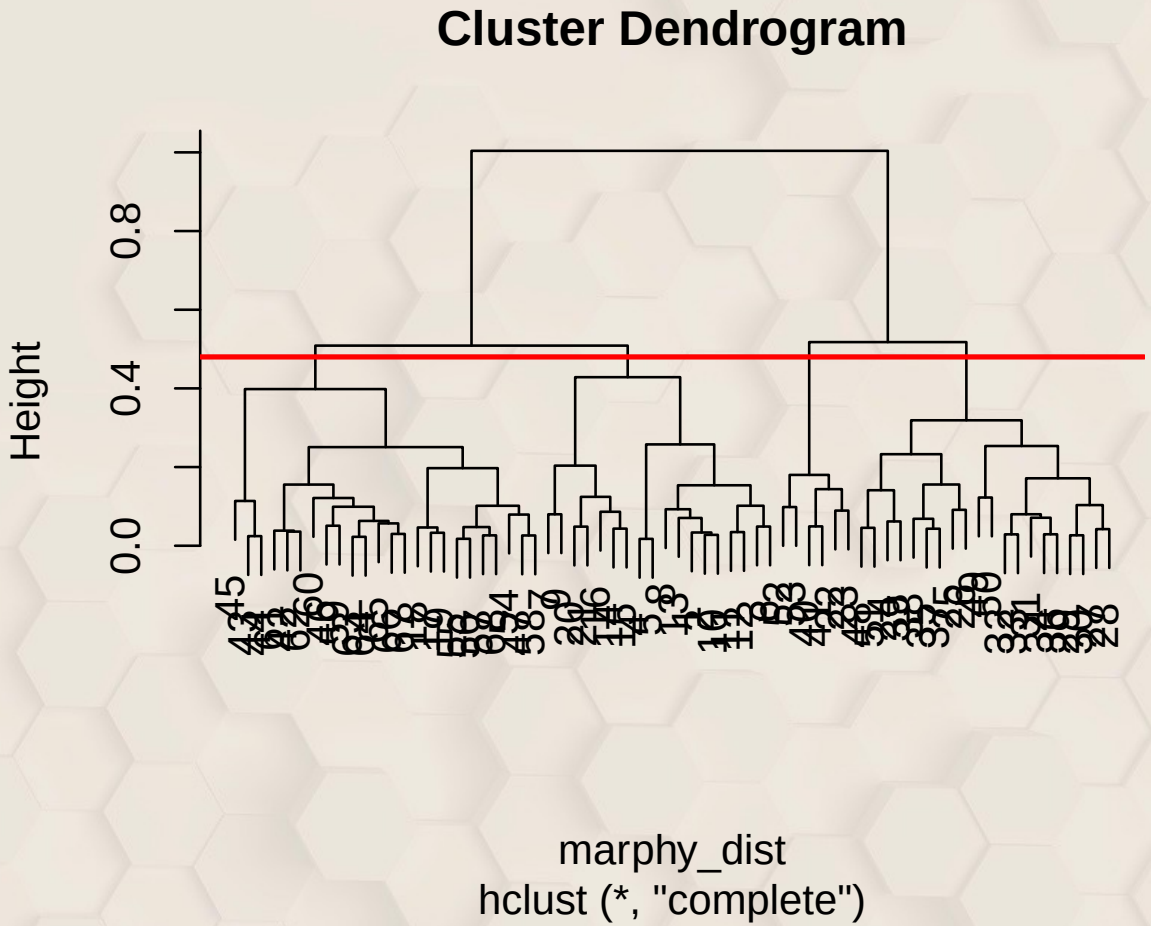


UMONS

```
plot(marphy_complete)
abline(h = 0.48, col = ”red”, lty = ”solid”, lwd = 2)
```

Consignes

- Déterminez vos groupes et indiquez le niveau de coupure en rouge sur votre graphique avec la fonction `abline()`.
- Employez différents indices et différents types de liens pour obtenir la meilleure classification selon vous.



UMONS

Classification hiérarchique : procédure

Si nous devions résumer la procédure de traitements des données, les étapes sont les suivantes :

- Transformation des données si nécéssaire
- Choix de l'indice pour la matrice de distance
- Choix de la méthode de regroupements pour le dendrogramme
- Choix du nombre de classe ou du niveau de coupure dans le dendrogramme

UMONS

Marbio

Appliquez la procédure précédente sur le jeu de données `marbio` qui provient du package `pastecs`.

UMONS