

Science des données II : cours 5b



Classification hiérarchique

Philippe Grosjean & Guyliann Engels

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



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

Classification hiérarchique

- En anglais : **hierarchical clustering**
- Permet de faire des **groupes** à partir d'une matrice de (dis)similarité.
- En taxonomie : **espèces ou sous-espèces** et **arbres phylogénétiques**.
- En écologie, on a plus souvent des continuums, des gradients, mais cette méthode peut servir à détecter des **changements plus brutaux dans les écosystèmes**.
- Analyses souvent **complémentaires** aux méthodes d'ordination (présentation des individus sur des « cartes » que nous verrons plus loin)

Classification hiérarchique - Principe

- Il existe des méthodes divisives, et des méthodes **agglomératives**. Nous traiterons de ces dernières.
- Il existe plusieurs moyens d'agglomérer les données :
 - **Liens simples** (single linkages) : plus proches voisins,
 - **Liens complets** (complete linkages) : plus lointains voisins,
 - **Liens moyens** (group-average) : liens entre moyennes,
 - (liens **médians**, **centroïdes**, ..., mais attention aux *inversions*)
 - **Ward** : ANOVA à chaque niveau => minimise la variance intergroupes. Calcul complexe (non vu au cours). Tend à faire des groupes plus homogènes en nombres d'individus.
- Résultats **très différents** selon le choix de la méthode de liaison (simple => chaînages, complets => petits groupes bien séparés).

Le dendrogramme

- Graphique associé à la classification hiérarchique. En anglais : **dendrogram**.
[Illustration sur un exemple simple : marphy pour les stations 1, 20, 35, 50, 65].
- **Arbre** représentant les rassemblements successifs des groupes à des niveaux déterminés.
- Le dendrogramme est comme un **mobile** : la position des individus par rapport aux autres sur l'axe horizontal n'est pas figée.
- On peut décider de **couper à un certain niveau** (ligne horizontale) => détermine des *groupes distincts*.
- On peut aussi **combinaison classification hiérarchique et ordination** en présentant différentes couleurs pour les groupes dans l'ordination.

Classification hiérarchique - Exemple pratique

- Classification des **iris**. Retrouve-t-on les différentes espèces ?
- Matrice de **distances euclidienne** sur données standardisées.
- Classification hiérarchique à **liens complets**.
- **Comparaison** des groupes obtenus avec les espèces.