

# Science des données II : cours 5a



## Introduction & matrices de distances

Philippe Grosjean & Guyliann Engels

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



<http://biodatascience-course.sciviews.org>  
[sdd@sciviews.org](mailto:sdd@sciviews.org)

# Introduction

Jusqu'ici nous nous sommes concentrés sur l'étude d'un petit nombre de variables à la fois.

Mais que faire lorsque le jeu de données est volumineux et que nous ne savons pas par où commencer ?

Dans la partie **analyse** de ce cours, nous aborderons l'**analyse exploratoire des données** ou EAD. Elle permet d'**explorer** des gros jeux de données multivariés et d'y découvrir des **structures** ou des **associations** entre individus.

# Analyse exploratoire de données

L'**analyse de données** ou **multivariée** se subdivise en deux familles de techniques complémentaires :

- Les méthodes de **classification**. Ces méthodes regroupent les objets en plusieurs sous-ensembles distincts (**groupes**), chaque individu étant plus semblable aux autres au sein de son propre sous-ensemble qu'il ne l'est par rapport aux individus des autres sous-ensembles.
- Les méthodes d'**ordination** ou de **réduction de dimensionnalité**. Elles représentent les individus sur des graphes bi- ou tridimensionnels (ou dit, une représentation sur des **cartes**, par analogie à la géographie). La distance des individus entre eux est relative à leur degré de ressemblance ou de différence (on dit : **similarité** ou **dissimilarité**).

# Analyse exploratoire de données

L'**analyse de données** ou **multivariée** se subdivise en deux familles de techniques complémentaires :

- Les méthodes de **classification**. Ces méthodes regroupent les objets en plusieurs sous-ensembles distincts (**groupes**), chaque individu étant plus semblable aux autres au sein de son propre sous-ensemble qu'il ne l'est par rapport aux individus des autres sous-ensembles.
- Les méthodes d'**ordination** ou de **réduction de dimensionnalité**. Elles représentent les individus sur des graphes bi- ou tridimensionnels (ou dit, une représentation sur des **cartes**, par analogie à la géographie). La distance des individus entre eux est relative à leur degré de ressemblance ou de différence (on dit : **similarité** ou **dissimilarité**).

## Exemples de jeux de données traités

- Le dénombrement (quasi-)exhaustif des **espèces** rencontrées en différents lieux (**stations**).
- L'**abondance** (nombre d'individus) dans chaque **taxon** en différentes stations échantillonnées.
- Une étude détaillée des **conditions du milieu** (paramètres physico-chimiques, terrain, etc.) où vivent les animaux étudiés en différentes stations.
- Un tableau rassemblant la **biométrie** (taille, poids, poids de certains organes, ...) d'un nombre important d'individus d'une population donnée, mesurés afin de déterminer l'évolution de ces mesures biométriques au cours de la croissance.
- Etc.

# Matrice de distances

# Matrice de distances

- La **matrice de distances** est le point de départ (première étape explicite ou implicite) de nombreuses analyses multivariées.
- Présentation des individus ou des variables aussi bien en ligne qu'en colonne => **deux points de vue possibles**.
- Les éléments de la matrice correspondent à **toutes les paires possibles, prises deux à deux**; on obtient une **matrice carrée**.
- Exemples de matrices de distances déjà abordées :
  - Matrice de **variances/covariances** = distances euclidiennes au carré.
  - Matrice de **corrélation** = distances euclidiennes au carré sur des données standardisées.

## Indice de similarité/dissimilarité

- Un **indice de similarité (similarity index)** est un descripteur statistique (nombre unique) de la similitude de deux échantillons ou individus représentés par plusieurs variables (échantillon multivarié).
- Un indice de similarité prend une valeur **comprise entre 0 (différence totale) et 1 ou 100% (similitude totale)**.
- Un **indice de dissimilarité** est le complément d'un indice de similarité ( $dis = 1 - sim$ ); sa valeur est **comprise entre 100% (différence totale) et 0 (similitude totale)**. **Attention: dans certains cas, un indice de dissimilarité peut varier de 0 à  $+\infty$**  (lorsqu'il s'agit d'une distance, euclidienne par exemple).
- Tous les indices de similarité / dissimilarité peuvent servir à construire des **matrices de distance**.



# Indice de dissimilarité : Bray-Curtis

- = coefficient de Czecanowski

$$D_{\text{Bray-Curtis}_{j,k}} = \frac{\sum_{i=1}^n |y_{ij} - y_{ik}|}{\sum_{i=1}^n (y_{ij} + y_{ik})}$$

- S'utilise pour mesurer la similitude entre échantillon sur base du **dénombrement d'espèces**. Si le **nombre d'individus est très variable** (espèces dominantes *versus* espèces rares), penser à transformer (ex:  $\log(x+1)$ , double racine carrée, ...)

## Indice de dissimilarité : Canberra

- Distance similaire à Bray-Curtis mais pondérant les espèces en fonction du nombre d'occurrences.

$$D_{\text{Canberra}_{j,k}} = \frac{1}{nz} \sum_{i'=1}^{nz} \frac{|y_{i'j} - y_{i'k}|}{|y_{i'j}| + |y_{i'k}|}$$

où  $nz$  est le nombre de valeurs non nulles.

- Toutes les espèces contribuent de manière égale  $\Rightarrow$  possibilité de surimportance d'une espèce mesurée une seule fois !
- Toute double absence n'est pas prise en compte  $\Rightarrow$  se comporte bien face aux tableaux comportant beaucoup de zéros (idem Bray-Curtis).

## Utilisation d'indices Bray-Curtis et Canberra

Seuls les indices ne dépendant pas des doubles zéros sont utilisables pour des dénombrements d'espèces ou des présence-absence.

- **Bray-Curtis** => résultat dominé par les espèces les plus abondantes.
- **Canberra** => risque de domination des espèces rares.
- Bray-Curtis sur données transformées ( $\log(x+1)$  ou double racine carrée) = **souvent bon compromis**.
- Si les volumes échantillonnés entre stations ne sont pas comparables, **il faut standardiser**.

# Indice de dissimilarité : distance Euclidienne

- Distance **géométrique** entre les points:

$$D_{\text{Euclidean}}_{j,k} = \sqrt{\sum_{i=1}^n (y_{ij} - y_{ik})^2}$$

# Indice de dissimilarité : Manhattan

- ou encore **city-block distance**.

$$D_{\text{Manhattan}_{j,k}} = \sum_{i=1}^n |y_{ij} - y_{ik}|$$

- Les indices de **dissimilarité** de Bray-Curtis et Camberra sont complémentaires aux indices de similarité correspondants (**dis** = **1 - sim**).

# Utilisation de ces indices

- Les distances euclidienne ou de Manhattan sont à préférer pour les **mesures environnementales**.
- Les distances de Bray-Curtis ou Canberra sont meilleure pour les **dénombrements d'espèces** (nombreux double zéro) !

## Indices de (dis)similarité - propriétés

- Les indices varient en 0 et 1 (0 et 100%), mais les distances sont utilisées aussi comme indices de dissimilarité et varient entre 0 et  $\infty$ .
- Un indice est dit **métrique** si :
  - **Minimum 0** :  $I_{a,b} = 0$  si  $a = b$
  - **Positif** :  $I_{a,b} > 0$  si  $a \neq b$
  - **Symétrique** :  $I_{a,b} = I_{b,a}$
  - **Inégalité triangulaire** :  $I_{a,b} + I_{b,c} \geq I_{a,c}$
- Un indice est **semimétrique** s'il répond à toutes les conditions sauf la quatrième.
- Un indice est dit **non métrique** dans les autres cas.

# Indice de (dis)similarité - métrique

Distance	Type
Bray-Curtis	semimétrique
Canberra	métrique
Euclidienne	métrique
Manhattan	métrique
Chi carré	métrique
(correlation)	(non métrique)
(variance/covariance)	(non métrique)