

Science des données II : tp2



Git & GitHub

Indice de distance

Guyliann Engels & Philippe Grosjean

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



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

UMONS

La gestion de version est un outil indispensable dans votre recherche

La présentation se trouve dans :

- moodle -> Les nouveaux outils de la science des données -> Git et GitHub

Les mots essentiels sont :

- Git
- GitHub
- GitHub Desktop
- Commit
- Push
- Pull
- Fork
- Branch
- Merge

Analyse multivariée : matrice de distance

En partant d'un tableau de type Espèce/Station, quelles sont les stations les plus similaires ? Ce type de questions nécessite l'utilisation d'outils liés à l'analyse multivariée.

	espece_1	espece_2	espece_3	espece_4
station_1	5	0	0	2
station_2	2	2	3	0
station_3	0	0	1	10
station_4	0	3	4	3

Le point de départ de nombreuses analyses multivariées est **la matrice de distance**.

Les différents indices

Différents indices de similarité et de dissimilarité sont employés pour composer la matrice de distance.

■ Similarité

- Bray-Curtis : $S_{jk} = 1 - \sum_{i=1}^p \frac{|y_{ij} - y_{jk}|}{\sum_{i=1}^p (y_{ij} + y_{jk})}$
- Canberra : $S_{jk} = 1 - \frac{1}{NZ} \sum_{i=1}^p \frac{|y_{ij} - y_{jk}|}{(y_{ij} + y_{jk})}$

Ces deux indices sont à privilégier lors de **dénombrements d'espèces**

■ Dissimilarité

- Distance euclidienne : $D_{ij} = \sqrt{\sum_{i=1}^p (y_{ij} - y_{jk})^2}$
- Manhattan : $D_{ij} = \sum_{i=1}^p |y_{ij} - y_{jk}|$

Ces deux indices sont à privilégier lors de **mesures environnementales**

Calcul de matrices de distances

Calculez les matrices de dissimilarité entre les stations suivantes avec la distance euclidienne et l'indice de Bray-Curtis

	espece_1	espece_2	espece_3	espece_4
station_1	5	0	0	2
station_2	2	2	3	0
station_3	0	0	1	10
station_4	0	3	4	3

Quels sont les deux stations les plus proches ? Selon Bray-Curtis ? Selon la distance euclidienne ?

Transect entre Nice et Calvi

- Etude sur 68 stations
 - **Murphy** comprend les mesures de température, de salinité, de fluorescence et de densité.
 - **Marbio** comprend le dénombrement de différents groupes au sein du zooplancton.

Les données se trouvent dans le package **pastecs**
Réaliser un projet afin d'étudier ce transect



Employez la fonction **vegdist()** du package **vegan** afin de calculer vos matrices de distances sur les données proposées :

- marphy
- marbio

Pour faire appel à l'aide de la fonction, il suffit d'écrire cette dernière précédé d'un point d'interrogation (**?vegdist**) dans la console R. Employez un indice cohérent en fonction des données proposées.