# Bref récapitulatif de la démarche

Abdoulaye Diabakhaté
Encadrants: Stéphane ROBIN et Mahendra MARIADASSOU

19 octobre 2018

# Plan de la présentation

# Présentation du jeu de donnée utilisé : *Genocenoses_env_parameters_all_tara_initial.xlsx*

Ce jeu de donnée que j'ai reçu de la part de Romain NARCI et sur lequel s'est fait mes études comporte initialement 644 observations sur 20 variables.

Pour toutes les études menées, on s'est d'abord focalisé sur 15 variables qui sont :
Lat,Long,T,Sal,chla,O2,NO3m,NO3,NO2,NH4,SSD,Phos,Si,depth,Fe

Les fractions de taille concernées sont au nombre de 6 :

- 3 fractions composées de 11 matrices de distance : 0-0.2, 0.22-3 et 20-180
- 3 fractions composées de 13 matrices de distance : 5-20, 180-2000 et 0.8-5

# Présentation des matrices de distance : Les 11 matrices et les 13 matrices de chaque groupe

Les 11 matrices du premier groupe de fractions sont :

1. jaccard_abundance
2. ochiai_abundance
3. sorensen_abundance
4. simka_jaccard_abundance
5. chord_hellinger_prevalence
6. jaccard_canberra_prevalence
7. kulczynski_prevalence
8. ochiai_prevalence
9. whittaker_prevalence
10. simka_jaccard_prevalence
11. sorensen_braycurtis_prevalence

Les 13 matrices du deuxième groupe de fractions sont :

1. jaccard_abundance
2. ab_jaccard_abundance
3. braycurtis_abundance
4. ab_ochiai_abundance
5. ab_sorensen_abundance
6. simka_jaccard_abundance
7. chord_prevalence
8. jaccard_prevalence
9. kulczynski_prevalence
10. ochiai_prevalence
11. whittaker_prevalence
12. simka_jaccard_prevalence
13. sorensen_braycurtis_prevalence

# Syntaxe de la méthode adonis sous R

**La syntaxe d'Adonis du package vegan de 2 façons est :**

- $adonis(matrice \sim variable1 + ... + variable15, data)$
- $adonis(matrice \sim ., data)$

On peut ajouter quelques arguments dans la fonction adonis comme :
permutations=999, qui est le par défaut, même si on ne le met pas ou method = euclidean...

# Les analyses menées

Après avoir appliqué notre méthode adonis, sur l'ensemble des 15 variables cibles, citées plus haut.

Comme la méthode adonis prenait en compte l'ordre d'inclusion des variables, et comme on ne pouvait pas tester tous les ordres d'inclusions possibles ($2^p - 1$ possibilités), cela nous a motivé a rangé les 15 variables cibles en 3 groupes comme suit :

Physique(constituait de depth et T),Géographie (constituait de Lat et Long)et Chimie(constituait de tout le reste des autres variables).

C'est à la suite de ce classement qu'on s'est retrouvé a testé 6 ordres de variables des 3 groupes : ce qui a été fait.

Par la suite on a ajouté une nouvelle variable, cette fois-ci catégorique,Génocénose et on s'intéresse à :

- Commencer par tester si l'effet génocénose est significatif une fois les autres variables déjà prise en compte.
- Et,symétriquement, voir quelles variables auraient un effet significatif une fois l'effet génocénose pris en compte.

A la suite de ces analyses on s'est intéressé également à effectuer une comparaison 2 à 2 des modalités de la variable génocénose. A cet effet, nous avons utilisé la fonction pairwise.adonis du package pairwiseAdonis.

Cependant, cette fonction n'admet que la variable facteur,Génocénose dans le modèle (en plus de la matrice de distance), et non la totalité de la matrice de design (le reste des covariables), ce qui fera l'objet d'une étude plus approfondie pour les jours qui viennent.

# La fonction bioenv (Sélection de variables) : Best Subset Of Environmental Variables With Maximum (Rank) Correlation With Comunity Dissimilarities

En effet cette fonction, qui s'opère que sur des variables quantitatives a pour syntaxe :

**La syntaxe de bioenv du package vegan est :**

- $bioenv(matrice \sim variable1 + ... + variable15, data)$
- $bioenv(matrice \sim ., data)$

Cette fonction calcule une matrice de dissimilarité de communauté en utilisant vegdist.
Ensuite, elle sélectionne tous les sous-ensembles possibles de variables environnementales, met à l'échelle les variables et calcule les distances euclidiennes pour ce sous-ensemble en utilisant dist.

Ensuite, elle trouve la corrélation entre les dissimilarités communautaires et les distances environnementales, et pour chaque taille de sous-ensembles, enregistre le meilleur résultat.
Il y'a $2^p - 1$ sous-ensembles de p-variables.

En effet, un coefficient de corrélation (typiquement le coefficient de corrélation de rang de Spearman) est calculé entre les deux matrices et le meilleur sous-ensemble de variables environnementales peut alors être identifié et soumis ensuite à un test de permutation pour déterminer la signification.

La méthode s'arrête lorsque rajouter des variables n'augmente plus la corrélation.