Master TRIED,

Etude de cas1: Second Jalon

Analyse exploratoire sur toutes les variables

Réalisé par:

YUEHGOH Foutse

CHIBANE Lydia

Année universitaire : 2018/2019

Introduction :

Le phytoplancton (plancton végétal) regroupe les organismes photosynthétiques, microscopiques, vivant en suspension dans l'eau.

On dispose de 7 variables différentes de la base GlobColor.

La variable	Signification
SST	Température à la surface de la mer
PFT	Variétés des phytoplanctons
La chlorophyle-a	pigment présent chez tous les végétaux qui permet de capter la lumière nécessaire a la photosynthèse (Un processus très important qui permet aux végétaux de transformer le dioxyde de carbone (CO2) en matière organique).
GlobColor412	Les différentes longueurs d'ondes
GlobColor443	
GlobColor490	
GlobColor555	

L'objectif de ce deuxième jalon est l'étude des différentes données pour éventuellement mieux les comprendre, les manipuler et d'arriver à proposer une problématique à résoudre en utilisant ces résultats.

Dans le jalon précédent, on a fait l'étude exploratoire sur une seule variable en guise d'initiation pour la manipulation des données.

Dans la suite, on rapporte les différentes étapes qu'on a effectuées pour avancer dans notre étude sur la variabilité des phytoplanctons.

1. Etude des données

On dispose de sept variables différentes dont six sont quantitatives et la dernière PFT est une variable qualitative.

Pour chacune des variables on dispose de 184 fichiers représentant les données moyennes de la variable sur 8 jours.

En lisant les fichiers, on trouve des données (-999) ceci voudra dire qu'on ne dispose pas d'informations.

Pour pouvoir les manipuler, on a remplacé ces mêmes données (-999) par des « nan » (données manquantes).

Dans ce cas, plusieurs solutions sont envisageables pour palier à ce problème :

- Faire l'étude en gardant l'échantillon tel qu'il est avec les données manquantes.
- Eliminer la variable qui a beaucoup de valeurs manquantes.
- Si la variable avec beaucoup de valeurs manquantes est cruciale à l'étude, on supprime les individus ayant des données manquantes.
- Faire l'imputation qui est de remplacer ces données manquantes (nan) par des moyennes par exemple ou tout autre indicateur.

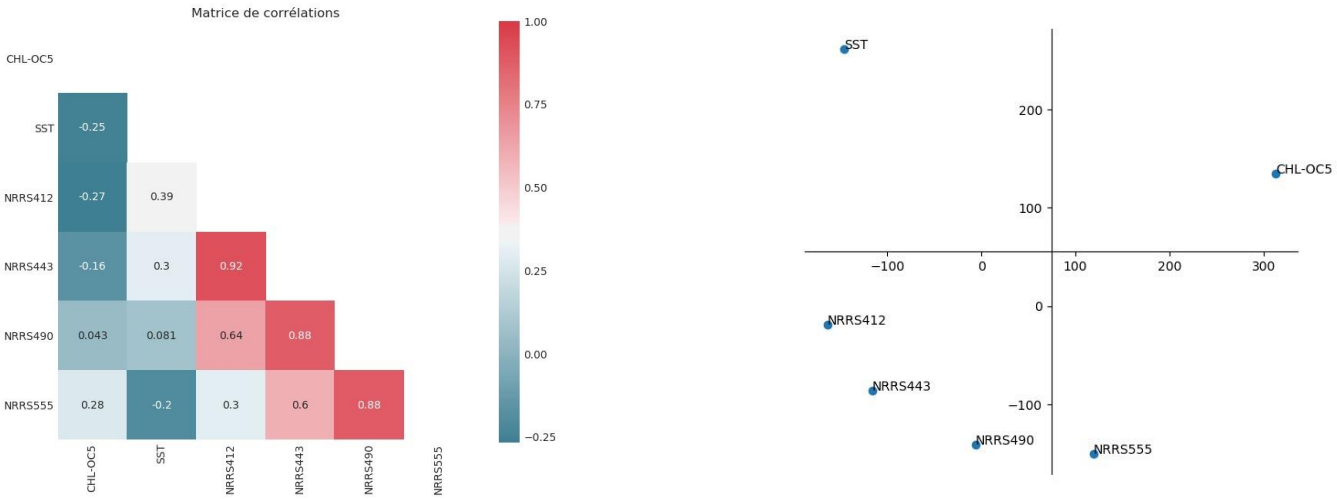
Dans notre cas, on a éliminé les individus ayant des valeurs manquantes.

2. Etude bidimensionnelles :

On a organisé nos données dans un tableau à 3 dimension car pour 7 variables on a 184 fichiers et chacun contenant plusieurs variables mais on prend seulement la variable qui nous intéresse et qui sera présenté dans des matrices 433*769 et cette dernière on l'a réordonné à une dimension avec 332977 valeurs.

Donc 332977 données pour chaque variable et pendant une seule semaine.

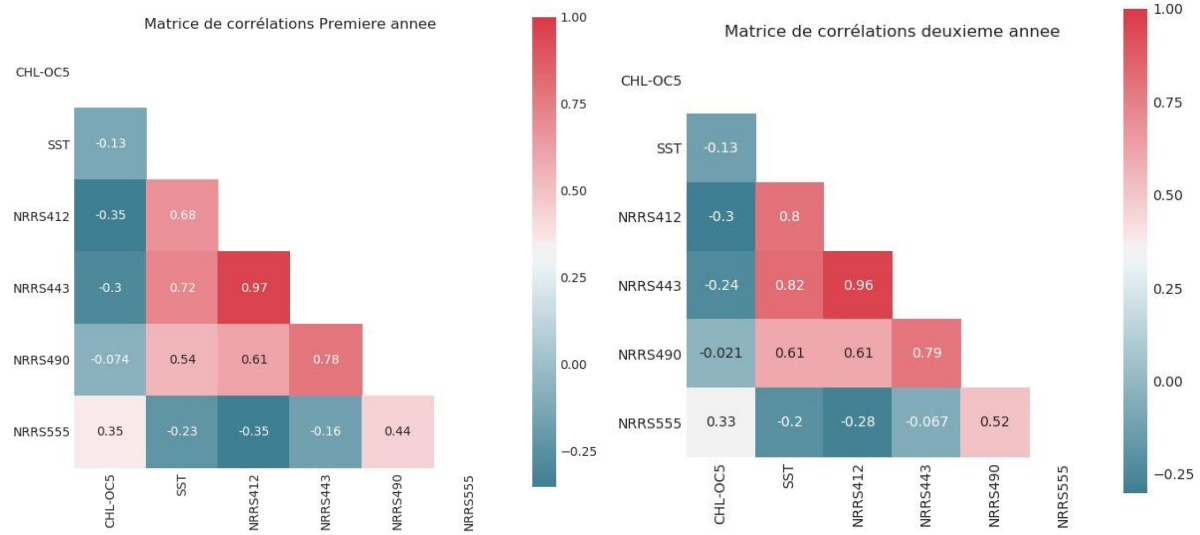
Dans un premier temps pour une seule semaine, on obtient les corrélations suivantes :

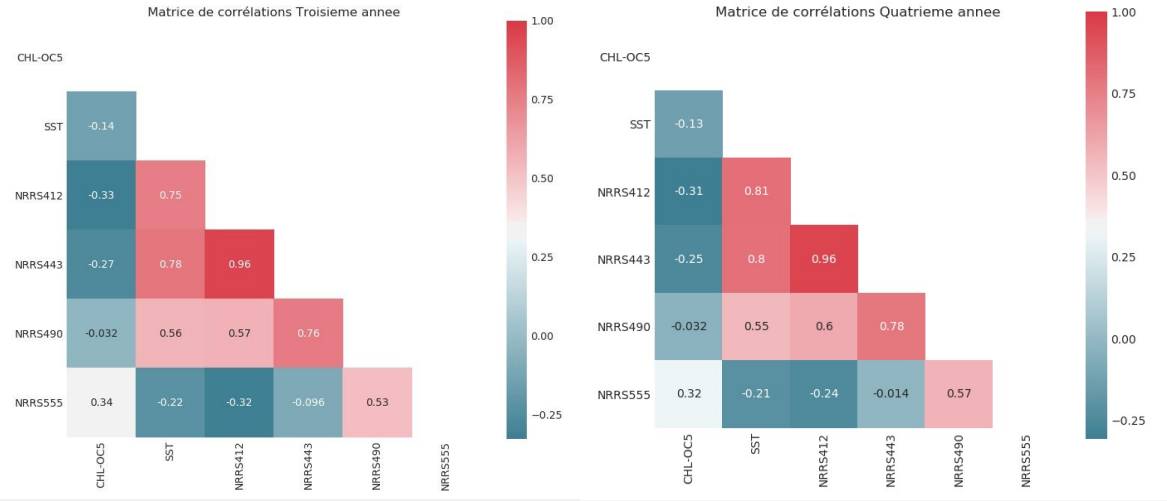


D'après la matrice de corrélation, on a les variables « longueurs d'onde » qui sont corrélées entre elles deux à deux soit entre « NRRS490 &NRRS555 » et « NRRS443 & NRRS490 » et « NRRS412 & NRRS443 ».

Et c'est la même information qu'on obtient de l'ACP en traçant le nuage des variables.

On fait les moyennes annuelles puis on retire les individus où on a des données manquantes et on calcule les corrélations entre variables deux à deux pour en tirer les éventuelles relations entre les différentes variables.





Pour les différentes années, on remarque que les longueurs d'ondes son corrélées entre elles deux à deux.

Les relations linéaires sont entre « NRRS412 & NRRS443 », « NRRS490 & NRRS443 ». On remarque aussi une corrélation entre « NRRS490 & NRRS555 » et entre « NRRS412 & NRRS490 », soit une relation moins forte que la précédente mais qui n'est quand même pas négligeable.

Il y'a également une relation linéaire entre la température à la surface de la mer et les 3 longueurs d'onde « NRRS412 », « NRRS443 » et « NRRS490 ».

Tandis que la variable CHL-OC5 n'est pas fortement liée aux autres.

En se basant sur des recherches antérieures, on peut extraire certaines relations entre toutes ces variables à étudier et les phytoplanctons :

- La croissance du phytoplancton dépend de la lumière du soleil, de la température et des niveaux de nutriments disponibles. Parce que les eaux froides ont tendance à avoir plus de nutriments que les eaux chaudes, le phytoplancton a tendance à être plus abondant là où les eaux sont froides. Ceci dit, il y'a une relation entre la quantité de phytoplancton présente dans la mer et la température à la surface de la mer.
- Des corrélations inverses ont été trouvées entre la biomasse, la productivité et la composition du phytoplancton dans les masses d'eau avec une température de surface de la mer inférieure à 19 ° C.
- La concentration en chlorophylle-a détecte la présence de phytoplancton
- chaque classe de phytoplancton est efficace à des longueurs d'onde différentes.

3. Proposition d'une problématique :

Sur ce, en se basant sur les informations qu'on a eu, on se propose de faire de la classification de phytoplancton en utilisant les longueurs d'ondes.

On va utiliser les données mises à notre disposition pour faire la classification, comparer avec la classification déjà faite pour éventuellement discuter l'information disant que chaque classe est efficace a des longueurs d'ondes différentes.

Référence bibliographiques :

https://www.sciencedirect.com/science/article/pii/S0304420315300323?via%3Dihub

https://earthobservatory.nasa.gov/global-maps/MYD28M/MY1DMM_CHLORA

http://somlit.epoc.u-bordeaux1.fr/fr/IMG/pdf/Chlorophylle_2014.pdf