ETUDE DE CAS 1 :

Classification des phytoplanctons par sélection de variables :

Par CHIBANE Lydia, YUEHGOH Foutse.

Introduction :

Le phytoplancton constitue l'ensemble des cyanobactéries et microalgues (végétaux microscopiques) présentes dans les eaux de surface et qui dérivent au gré des courants. Méconnu car invisible à l'œil nu, le phytoplancton est pourtant le poumon de notre planète. Grâce à la photosynthèse, il produit plus de la moitié de l'oxygène terrestre et consomme la moitié du dioxyde de carbone. Le phytoplancton est indispensable à la vie marine car il se trouve également à la base de la chaîne alimentaire océanique.

Il existe une grande diversité de formes et de couleurs de phytoplanctons, des milliers d'espèces ont été recensées à ce jour. Pour les distinguer, on les regroupe en plusieurs familles.

De ce fait, une classification a été établi pour avoir différents types fonctionnels des phytoplanctons (PFT). Soit les sept classes suivantes:

- Classe 1 : Haptophytes-Nanoplancton
- Classe 2 : Chlorophycées-Nanoflagellées
- Classe 3 : Cryptophycées-Nanoflagellées
- Classe 4 : Prochlorococcus
- Classe 5 : Synechococcus
- Classe 6 : Diatomées
- Classe 7 : Non identifiable

Ainsi, on dispose d'une distribution de ces différentes familles dans la mer méditerrané et ce pendant 4 années. En plus de cette distribution, on a également, pour les mêmes résolutions spatio-temporelles, la température sur la surface de la mer (SST), les réflectances marines à différentes longueurs d'ondes (Rrs) et la concentration de la Chlorophylle (Chl-OC5). Toutes ces variables sont issues du portail GlobColour et constituent notre base de données.

Dans ce qui suit, on propose de faire une classification par sélection de variables. Suite à une recherche bibliographique à ce sujet, on démarre de l'hypothèse que chaque classe de phytoplancton est efficace à des longueurs d'onde différentes, ainsi à l'issue de cette étude, on pourra réduire les variables de notre base de données initiale et de ne garder que celles qui sont pertinentes pour la détermination du type fonctionnel du phytoplancton et voir si effectivement les longueurs d'ondes suffisent pour cette classification.

Prétraitement de la base de données :

La base de données mise à notre disposition comporte les mesures satellitaires de la Chla(OC5), des 4 Rrs(λ) et de la SST sur une grille régulière avec une résolution spatiale de 4km et une résolution temporelle de 8 jours (moyenne hebdomadaire), en plus de la PFT. Toute la base couvre la mer méditerranée durant une période de 4 ans.
Après les études préliminaires unidimensionnelles, on remarque que pour chaque variable, il y'a des données manquantes dans certaines régions.
Pour pallier à ce problème, plusieurs solutions sont envisageables. Nous, on a choisi de supprimer les individus ayants des informations manquantes. Ainsi, on se retrouve avec

une base de dimension réduite certes, mais reste quand même assez riche et les informations sont exacts car c'est celles comportant les vraies valeurs provenant des enregistrements satellitaires.

Classification par sélection de variables :

Dans cette section, nous allons utiliser «les arbres de décision » ainsi que «le random forest» pour faire une classification par sélection de variables. A chaque fois qu'on déroule l'algorithme, on pourra visualiser les variables qui sont les moins influentes. Puis on fera l'apprentissage une seconde fois après élimination de cette variable pour voir si les performances sont maintenues ou pas. On étire cette étape jusqu'à ce qu'on n'obtienne que les variables les plus importantes soit celles avec lesquelles on obtient les mêmes performances que si on faisait la classification avec les six variables initiales. On pourra à l'issue de cette classification comparer dans un premier temps entre les deux algorithmes utilisés, et puis vérifier si les longueurs d'ondes seront les variables suffisantes à cette classification.

a. Choix des ensembles de test et apprentissage :

On travaille sur des séries temporelles où la composante de saisonnalité intervient.

Pour pallier à ce problème, on a proposé plusieurs solutions :

- Eliminer la composante de saisonnalité en utilisant le package «**saisonal_decompose** » de Python.
- Choisir sur une période de temps, 75% des données comme apprentissage et 25% pour le test et en vérifiant si les variables représentants ces deux ensembles suivent la même distribution pour s'assurer que lors de l'entrainement on prenne en compte les caractéristiques de toutes nos données.

On n'a pas pu implémenter correctement la première méthode. On a préféré travaillé sur la deuxième méthode. Le choix des données a été fait de sorte à tenir compte de toutes les variations saisonnières dans le modèle.

PS : parmi les 4 années mises à disposition, on a travaillé sur deux années dans la suite de l'étude.

b. Algorithmes utilisés pour la construction du modèle :

Les deux algorithmes utilisés pour faire la classification par sélection de variables sont :

1. Arbres de décision :

Les arbres de décisions sont des graphes non orientés acycliques et connexes. Ils emploient une représentation hiérarchique de la structure des données sous forme des séquences de décisions (tests) en vue de la prédiction d'un résultat ou d'une classe. Chaque individu, qui doit être attribué à une classe, est décrit par un ensemble de variables qui sont testées dans les nœuds de l'arbre.

2. Random Forest :

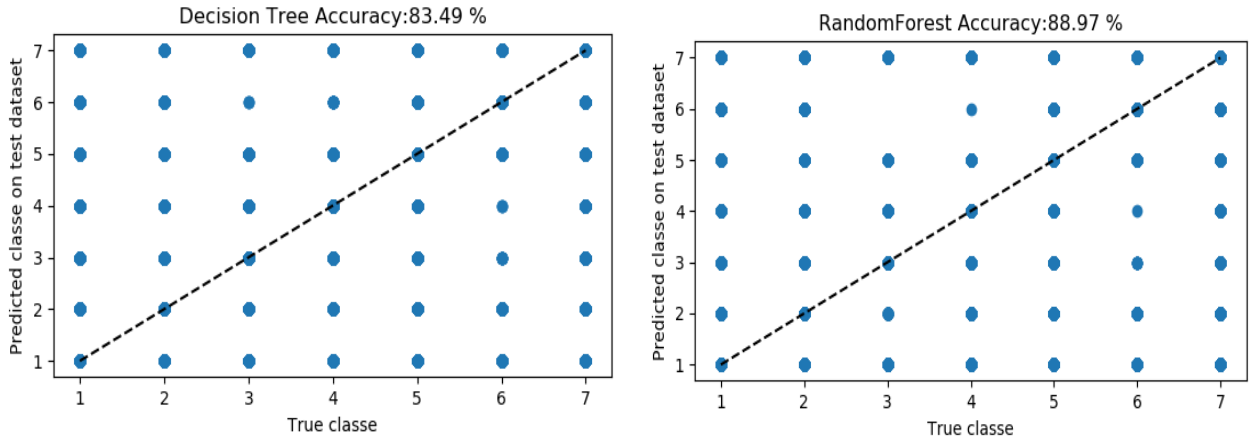
«Random Forest» est une méthode de classification où on effectue un apprentissage en parallèle sur de multiples arbres de décision construits aléatoirement et entraînés sur des sous-ensembles de données différents.

c. Méthodologie d'apprentissage :

Pour la classification par arbre de décision : On a utilisé la fonction « DecisionTreeClassifier » avec les paramètres par défaut de la fonction.

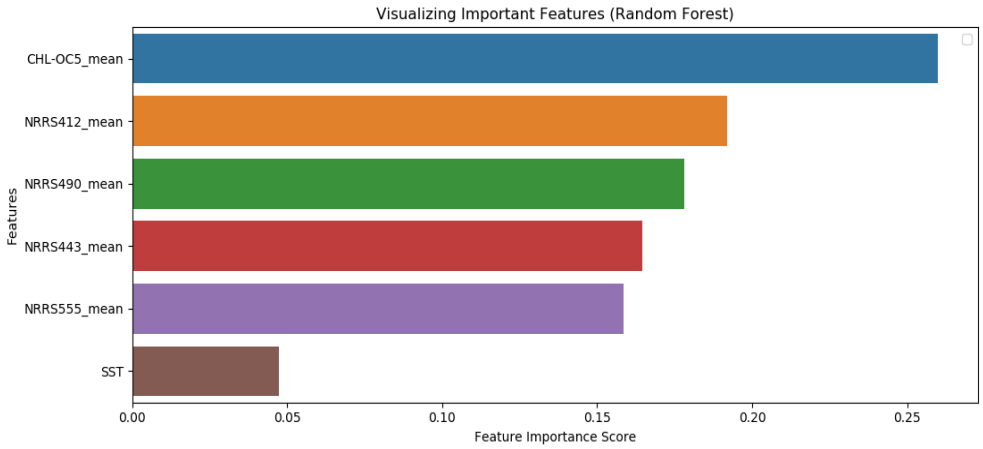
Pour la classification avec RandaomForest: On a utilisé la fonction « RandomForestClassifier » de python avec les paramètres par défaut de la fonction mais en ayant 100 arbres.

Apres entrainement des deux modèles de classification (pour les deux algorithmes choisis) sur notre ensemble d'apprentissage, on les applique sur l'ensemble test pour vérifier leurs performances. On montre les résultats de classification des données de cet ensemble dans les figures suivantes :



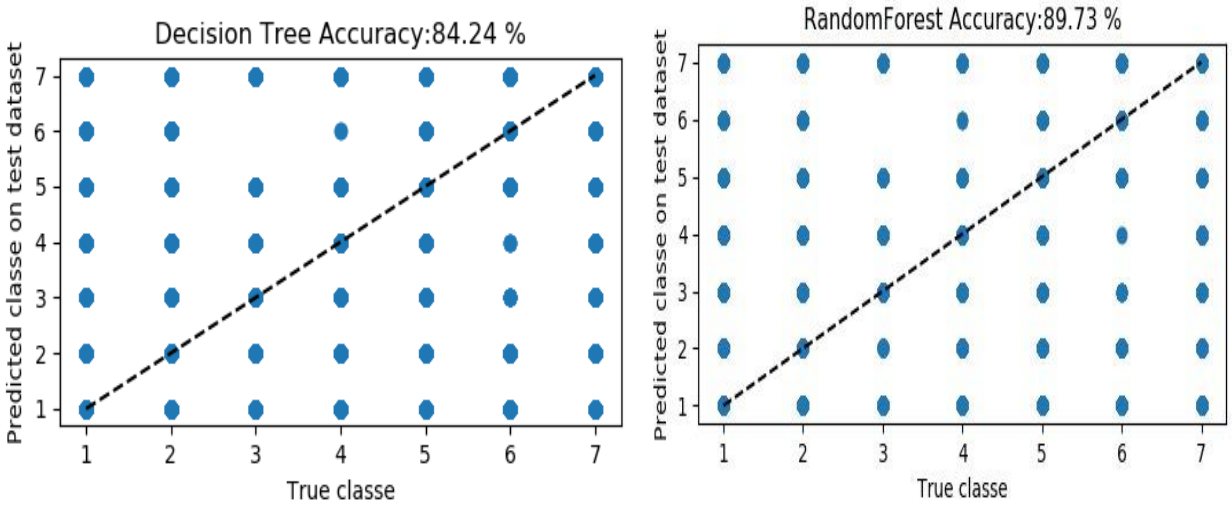
On obtient plutôt de bonnes performances surtout avec l'algorithme « RandomForest » avec un taux de 88.97% de biens classés.

Maintenant qu'on a le modèle, on voudrait savoir les variables les plus pertinentes pour cette classification, l'algorithme « Random Forest » est délivré avec une fonctionnalité intégrée qui permet d'afficher l'importance des variables utilisées, comme le montre la figure suivante :



Voyant cet histogramme, on remarque que la variable de température est la moins importante dans notre étude, donc on essaye de l'éliminer pour faire la classification avec seulement les cinq variables restantes et voir si ceci améliorerait les performances de la classification.

Les résultats de cette classification donnent les performances suivantes :



On remarque qu'en éliminant la variable SST, on a de meilleures performances pour les deux algorithmes.

A l'issue de cette deuxième classification, on trouve que les longueurs d'ondes ont pratiquement toutes la même importance et que la chlorophylle reste la variable la plus pertinente pour la classification.

Notre premier objectif de l'étude était de voir si avec seulement les longueurs d'ondes on pourra obtenir une différenciation entre les classes de phytoplanctons. En d'autres termes, si en faisant sélection de variables, on pourra éliminer la SST et la CHl-OC5 pour n'en garder les longueurs d'ondes. Vu que la chlorophylle est la plus pertinente, donc on arrête notre processus d'élimination de variables à ce niveau.

d. Discussion des résultats :

A l'issue des résultats de classification présentés ci-dessus, on peut en conclure sur certains points :

- ✓ **On a de meilleurs résultats en utilisant « Random Forest » :**

L'algorithme des « forêts aléatoires » est un algorithme de classification qui réduit la variance des prévisions d'un arbre de décision seul, améliorant ainsi leurs performances. Pour cela, il combine de nombreux arbres de décisions dans une approche de type bagging. Elles donnent de bons résultats surtout en grande dimension.

- ✓ **La sélection de variable faite :**

En éliminant dans une première étape la variable « SST », on a eu de meilleurs résultats de classification. Ceci dit, la qualité de la classification ne dépend pas du nombre d'informations à disposition mais de la pertinence de ces informations. En d'autres termes, ces variables pertinentes (CHLa, les 4 longueurs d'ondes) suffisent à distinguer les différences de caractéristiques entre les observations et à les regrouper en classes.

- ✓ **Pertinence des variables représentants les longueurs d'ondes :**

Les longueurs d'ondes sont des variables importantes pour la classification mais, suffisent pas pour avoir les différents types fonctionnels des phytoplanctons. La chlorophylle est une variable importante pour cette classification.

Conclusion :

Notre premier objectif pour cette étude, était l'exploitation de données mises à notre disposition, leur compréhension, proposition d'une problématique à résoudre à la fin.

Une étude préliminaire nous a permis de comprendre ce dont on dispose et éventuellement de poser une problématique à l'issue de cette étude, qui était de voir si les seules variables pertinentes pour la classification étaient les longueurs d'ondes. On a travaillé avec deux algorithmes qui sont arbres de décision et random forest et on a trouvé que le second est meilleur pour faire la classification. Suite à cette étude, on a trouvé que seule la température à la surface de la mer pourrait être éliminée pour avoir de bonnes performances en classification. Ceci dit, cette étude a été menée pour un jeu de données où on a éliminé tous les individus ayant des valeurs manquantes. Donc cette hypothèse émise au sujet de la SST reste valable pour ce cas de figure. Cette perte d'information résultante du prétraitement effectué pourrait avoir influencé cette classification.

Bibliographie :

[1] Mory OUATTARA : thèse de doctorat : Développement et mise en place d'une méthode de classification multi-bloc Application aux données de l'OQAI, CNAM 2014.

[2] Roy El Hourany, Marie Abboud-Abi Saab, Ghaleb Faour, Julien Brajard, Michel Crépon, Sylvie Thiria : Etude de la variabilité spatio-temporelle du phytoplancton à partir des pigments secondaires estimés par les observations satellitaires.

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

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

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

[6] https://www.analyticsvidhya.com/blog/2018/08/dimensionality-reduction-techniques-

[7] https://towardsdatascience.com/decision-trees-and-random-forests-for-classification-and-regression-pt-1-dbb65a458df