Institut de Science Financière et d'Assurances (ISFA)

MASTER 2 - ECONOMETRIE, STATISTIQUES : EQUADE

Projet - Apprentissage Statistique et IA

Classification orientée pixels d'images satellites en combinant le *Bootstrap Aggregating (Bagging)* et les Arbres de Décision

Etudiant :
Crésus KOUNOUDJI

Professeure :
Véronique Maume-Deschamps

Avril 2023

Table des matières

Chapitre 1

Cadre général

L'apparition de la photographie et l'essor de l'aviation puis la mise en place de satellite d'observation de la terre ont permis le développement d'ensemble de techniques utilisées pour déterminer à distance les propriétés d'objets naturels ou artificiels à partir des rayonnements qu'ils émettent ou réfléchissent, c'est la télédétection. Ce vaste champ d'étude a plusieurs applications très pratiques notamment en météorologie, en reconnaissance militaire, gestion des ressources agricoles et forestières, cartographie etc (Maya Nand Jha et al., 2019). Tout une branche de ces méthodes a pour objet l'étude de la propriété de certains milieu ou objets à partir de caractéristiques physiques mesurables notamment le rayonnement pour les identifier et donc faire de la classification. C'est donc assez logiquement que l'utilisation de méthodes statistiques et autres modèles d'apprentissages comme les arbres de décision a fait son essor notamment pour la caractérisation des cultures ou la cartographie (Gbodjo et al. (2020), Kumar et al. (2018), Zafari et al. (2017) et Gislason et al. (2005)). Dans le cadre de notre projet, nous avons comparé un modèle de forêt aléatoire à un modèle de forêt aléatoire avec bootstrap aggregating pour la classification orienté pixel d'images GeoTIFF de la zone de berlin alentour selon. Le Bootstrap Aggregating (Bagging) étant un algorithm d'apprentissage ensembliste permettant d'améliorer la stabilité et la précision des algorithmes d'apprentissage automatique en réduisant la variance et permettant d'éviter le surapprentissage. On cherche donc à déterminer si l'algorithme Bagging combiné au modèle classique apporte une amélioration significative dans l'utilisation de ces méthodes dans le domaine de la classification d'image satellites.

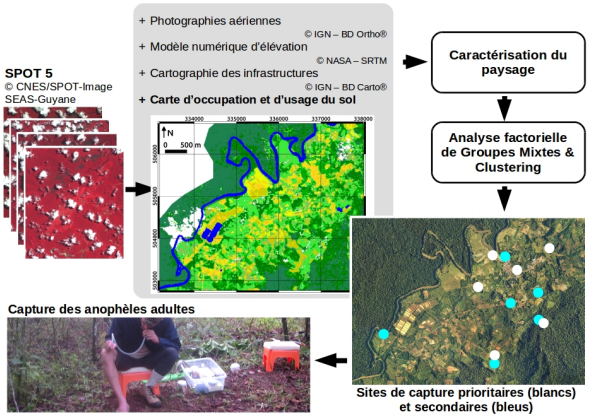

FIGURE 1.1 – Illustration d'utilisation de méthode statistique pour de la télédétection (caractérisation de sol) Herbreteau et al. (2018)

1.1 Objectifs du projet

Ce projet s'inscrit dans le cadre du cours d'Apprentissage Statistique et IA et s'inspire d'un module du séminaire Introduction to Remote Sensing destiné aux étudiants en géographie de la Humboldt-Universität zu Berlin. L'objectif générale est d'améliorer un modèle d'Arbre de décision pour classer des images satellites en le combinant avec l'agorithm Bagging. Il s'agit plus spécifiquement d'implémenter un modèle de Random Forêt sur des images satellites de la zone de Berlin et alentours pour classer les élément sur les images pixel par pixel selon s'ils sont : Prairies/herbacées(Grassland/Herbaceous) , Végétation ligneuse à feuilles larges (Broadleafed woody vegetation), Végétation ligneuse de conifères (Coniferous woody vegetation), Construit (Built-up), L'eau (Water), Terres cultivées (Cropland) à partir des bands spectraux des images.

Chapitre 2

Contexte Méthodologique

2.1 État de l'Art

La littérature sur les techniques et méthodes statistiques utilisées en télédétection est bien fournie. En particulier, l'utilisation d'arbres de décision a été largement étudiée. Mais dans une perspective de prospectives, il faut des résultats stable et le contrôle de surapprentissages et la précision du modèle sont impératifs. Breiman (1996) a proposé à cet effet, la méthode de Bootstrap Aggregating (Bagging) pour la prédiction à l'aide de plusieurs modèles de prédiction. Le Bagging est une technique de combinaison de modèles qui utilise un échantillonnage avec remplacement pour produire plusieurs sous-ensembles de données d'entraînement pour chaque modèle. En appliquant l'algorithme Bagging à un ensemble d'arbres de décision, on construit une forêt aléatoire avec des résultats plus stables et précis. Mise en perspectives avec d'autres méthode d'ensembles d'arbres de décision (le Boosting et la Randomisation), le Bagging apparaît efficace pour réduire la variance des prédictions (Dieterich, 2000). Ainsi les arbres de décision sont-elles très utilisées pour extraire des informations à partir des images de télédétection, tandis que que la méthode Random Forest est une méthode d'ensemble efficace pour la classification de la couverture terrestre. Pour la classification orienté apprentissage machine, on distingue principalement deux approches complémentaires celle basée sur les pixels avec un classificateur Random Forest et celle orienté objets avec une méthode de segmentation d'objet basée sur les relations hiérarchiques (Bui et Mucsi, 2020). Il apparait que les performances de modèles Random Forest comparées à d'autres méthodes de classification donnent des résultats comparables ou meilleurs (Gislason et al. 2006).

Chapitre 3

Modélisation

3.1 Random Forest : Le Bagging appliqué aux arbres de décision

Le Bagging (Bootstrap Aggregating) est une technique d'apprentissage ensembliste qui consiste à entraîner plusieurs modèles de manière indépendante sur des sous-ensembles aléatoires du jeu de données d'entraînement et à agréger leurs prédictions. L'idée est de réduire la variance des prédictions en moyennant les sorties de plusieurs modèles.

Soit $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ un jeu de données d'entraînement avec n observations et y_i la variable cible associée à l'observation x_i. On utilise le Bagging pour construire un ensemble de m modèles $h_1, \ldots, h_m$. Chaque modèle h_i est entraîné sur un sous-ensemble aléatoire de n' observations avec remise tirées uniformément dans D. Le Bagging consiste à agréger les prédictions de ces m modèles sur une nouvelle observation x par :

$$\hat{f}_{\text{Bag}}(x) = \frac{1}{m} \sum_{i=1}^{m} h_i(x) \tag{3.1}$$

où $\hat{f}_{\text{Bag}}$ est la prédiction agrégée et h_i est la prédiction du modèle i.

Le Bagging réduit la variance des prédictions en moyennant les sorties de plusieurs modèles. En effet, si les modèles sont suffisamment diversifiés, les erreurs aléatoires des différents modèles devraient s'annuler en moyenne et la prédiction agrégée devrait être plus robuste.

Pour quantifier la performance du Bagging, on peut utiliser le biais, la variance et l'erreur quadratique moyenne (MSE).

Le biais est défini comme la différence entre la moyenne des prédictions du modèle et la valeur réelle :

$$\text{Bias}(x) = f(x) - \frac{1}{m} \sum_{i=1}^{m} h_i(x) \tag{3.2}$$

La variance mesure la variabilité des prédictions d'un modèle autour de sa moyenne :

$$\text{Var}(x) = \frac{1}{m} \sum_{i=1}^{m} (h_i(x) - \frac{1}{m} \sum_{j=1}^{m} h_j(x))^2 \tag{3.3}$$

L'erreur quadratique moyenne (MSE) mesure la distance moyenne entre les prédictions et les valeurs réelles :

$$\text{MSE}(x) = \frac{1}{m}\sum_{i=1}^{m}(h_i(x) - y)^2 \tag{3.4}$$

L'avantage du Bagging est qu'il permet de réduire la variance des prédictions sans augmenter le biais. En effet, si les modèles sont suffisamment diversifiés, le biais du modèle agrégé devrait être similaire à celui des modèles individuels. Les forêts aléatoires sont une extension du bagging qui utilise des arbres de décision comme modèles d'apprentissage.

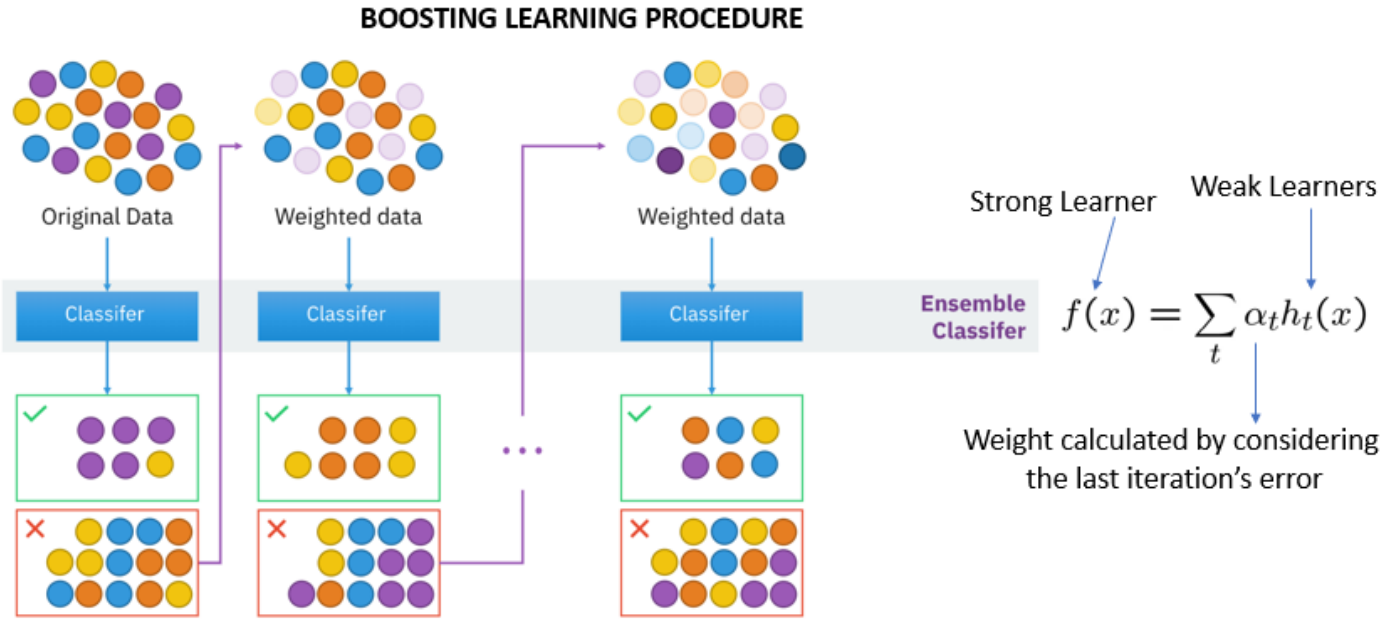

FIGURE 3.1 – Illustration de l'Algorithm Bagging

L'idée est de créer B arbres de décision en utilisant un échantillon aléatoire différent pour chaque arbre, et en choisissant une sous-régression aléatoire (random subset) des variables d'entrée à chaque nœud de l'arbre. La combinaison des prédictions des différents arbres permet de réduire la variance de l'estimation et d'améliorer la précision de la prédiction. Formellement, soit X = x1, x2, ..., xd l'ensemble des variables d'entrée, et soit D = (x1, y1), (x2, y2), ..., (xn, yn) l'ensemble de données d'entraînement. Pour chaque arbre j = 1, 2, ..., B, un échantillon aléatoire d'observations est créé en rééchantillonnant les observations avec remplacement, de sorte que chaque échantillon contienne environ deux tiers des observations. Ensuite, pour chaque nœud de l'arbre, un sous-ensemble aléatoire S de variables d'entrée est choisi, et la meilleure division selon l'un de ces variables est choisie pour partitionner les observations en deux sous-ensembles. La procédure de division est répétée jusqu'à ce que tous les nœuds atteignent une taille minimale ou que la partition ne peut plus être améliorée. Les prédictions des différents arbres sont ensuite combinées en prenant la moyenne (pour la régression) ou le vote majoritaire (pour la classification).

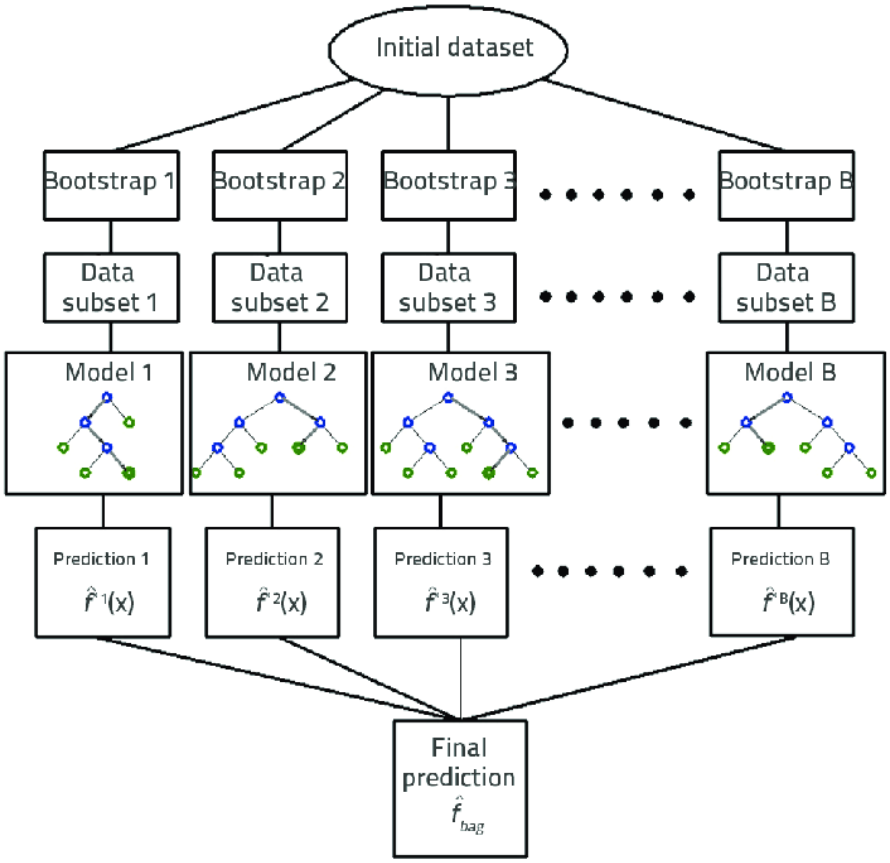

FIGURE 3.2 – Illustration Random forest : Bootstrap Aggregating (Bagging) sur plusieurs arbres de décision

3.1.1 Modèle Random Forest (RF) : Bootstrap Aggregating (Bagging) sur plusieurs arbres de décision

Dans ce projet, on a défini un modèle d'arbre de décision puis un modèle de forêt aléatoire constitué de cent (100) arbre de décision avec une profondeur maximale de 20 auxquels on applique la procédure du Bagging. La façon d'entraîner le modèle ici est différente de ce qui est observé dans la littérature en cela que au lieu de choisir au hasard des pixels d'entraînement et de test dans la même image, on utilise toute la matrice de pixel (Raster) de différentes images GeoTIFF pour entraîner le modèle qui sera tester/valider sur d'autres matrices entrières d'image différentes

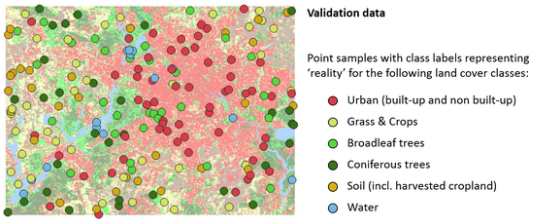

FIGURE 3.3 – Illustration d'entraînement et de validation sur les pixels d'une même image

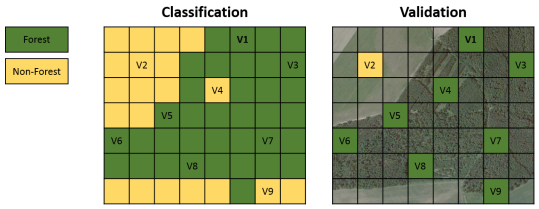

FIGURE 3.4 – Illustration de matrice de confusion sur les pixels d'une même image

3.2 Le jeu de données

Les données brutes utilisées dans ce projet sont des images satellites prises par Sentinels2, satellite du programme Copernicus de l'agence spatiale Européenne. Ces données ont été téléchargé sur la plate-forme Copernicus Open Access Hub en délimitant manuellement des zones géographiques sur Berlin et un peu au tour avec une couverture nuageuse inférieure à 10% entre Janvier 2019 et Décembre 2019. Les données ainsi acquises sont constituée de neuf (09) dossiers compressées (chacune d'environ 1 Go) contenant tous les données disponible et mesurée sur ces zones. N'étant intéressé que par les images, ces dossiers zip téléchargeable sur le Git repository que j'ai créé pour le projet ont été prétraité dans le script pour extrair les fichiers d'images JP2 selon une résolution choisi (R60m). Chaque zone pour la résolution R60m à quinze (15) fichiers JP2 qui représentes différentes bandes d'image (Rouge, Vert, Infrarouge, etc.). Pour chaque zone, les quinze (15) bandes ont été fusionnée en un seul et unique fichier GeoTIFF qui sera ensuite étiqueté pour les besoins du modèle. L'étiquetage à consisté à classer chaque pixel selon six (06) classes codées avec des chiffres : 0 - Grassland/Herbaceous , 1 - Broadleafed woody vegetation, 2 - Coniferous woody vegetation, 3- Built-up, 4- Water, 5-Cropland. Le classement est fondé sur le calcule à partir des bandes de l'image et l'application des seuils à l'indice de végétation NDVI (Normalized Difference Vegetation Index) et aux longueurs d'ondes courtes SW1 et SW2 utilisées pour détecter l'eau et les zones humides.

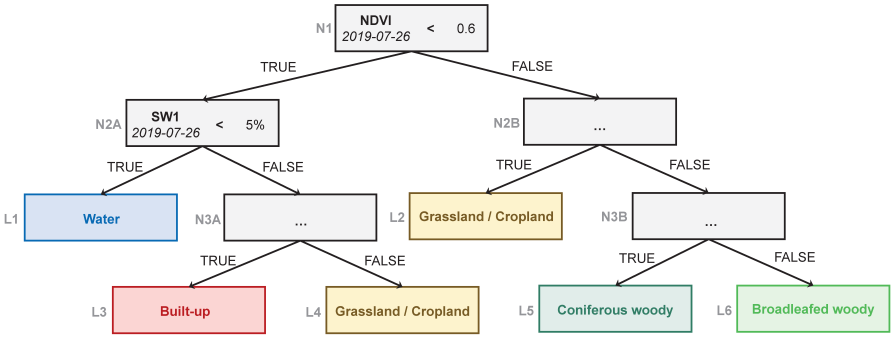

FIGURE 3.5 – Illustration des règles de décision de l'étiquetage des pixels

La nouvelle image GoeTIFF ainsi construite sera chargée en séparant les quinze(15) bandes d'origine qui représentent les *Features* du modèle et la seizième bande le *Target* à prédire. Notons que sur les neufs (09) images GoTIFF construite, une paire de trois images sont des captures d'une même exacte zone géographique mais à des période différente pour étudier la différence saisonnière au niveau spectral (mesurée par les bandes de couleurs) et servir d'une par pour l'entraînement, d'autre par pour le test et les trois dernières images sans paires pour validation.

3.3 Résultats

3.3.1 Analyse descriptive

La figure 3.6 ci-dessous donne un aperçu en fausses couleurs, de deux des cartes de GeoTIFF créées en fusionnant les différentes bande pour chaque zone. D'un autre côté, la figure 3.7 montre deux carte d'une même zone capturée en été 2019 et en hiver 2019 ainsi que la différence de profile spectrale entre les deux saison. Il est assez évident que entre les deux saisons, les mesures spectraux ne sont pas du tout les mêmes encore plus entre deux saisons aussi différente en

température, couverture et autres. Il serait donc intéressant de pourvoir tenir compte de la temporalité dans les mesures pour des modèle plus avancés.

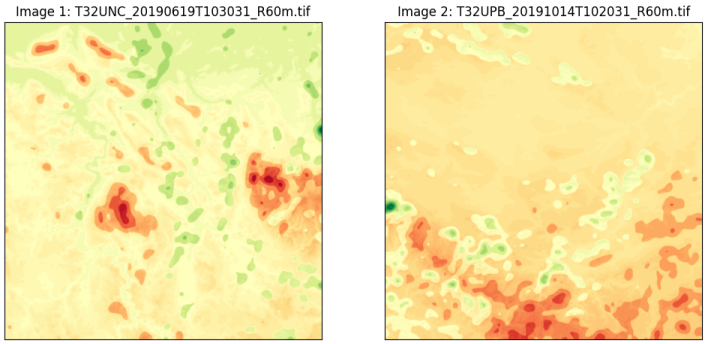

FIGURE 3.6 – Aperçu de deux différentes carte GeoTIFF créées à partire des quinzes (15) fichier JP2 correspondant

3.3.2 Modèles de prédiction

La cross-validation a permis de construire les matrices de confusion **??** ci-dessous et de calculer les performance des modèles. On obtient 99.67% en moyenne de précision pour les deux modèles. Il la classification par arbre de décision est déjà exceptionnellement précise. Certaines valeurs ont des précision nulle mais cela est dû à la prépondérance d'un nombre limité de classe par image (rarement tous en même temps et souvent plus de forêt et prairie ou construction de d'eau par exemple.

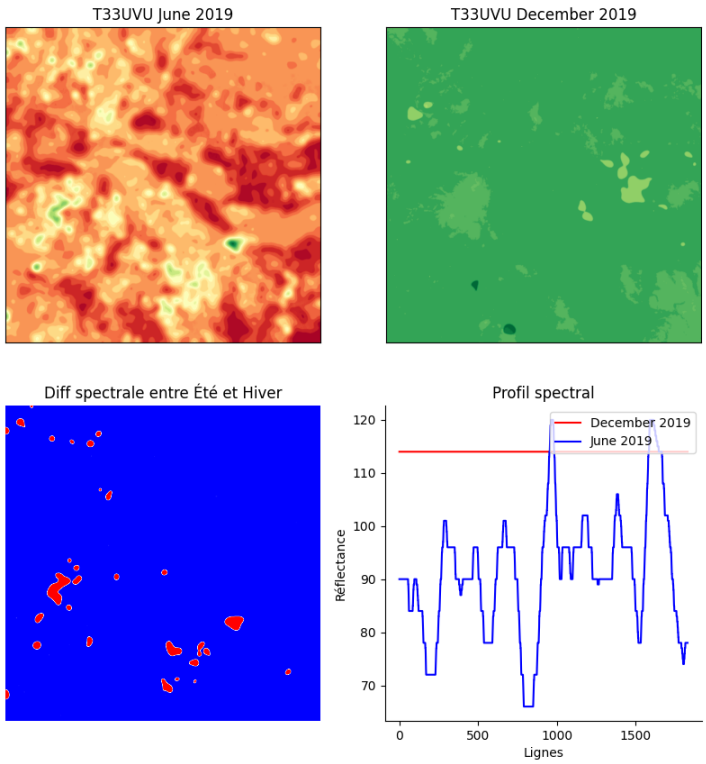

FIGURE 3.7 – Comparaison de profil spectral bi-saisonnière (Eté vs Hiver) d'une même zone géographique

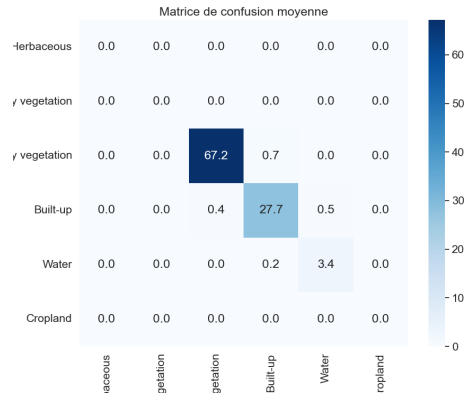

FIGURE 3.8 – Matrice de confusion du modèle arbre de décision

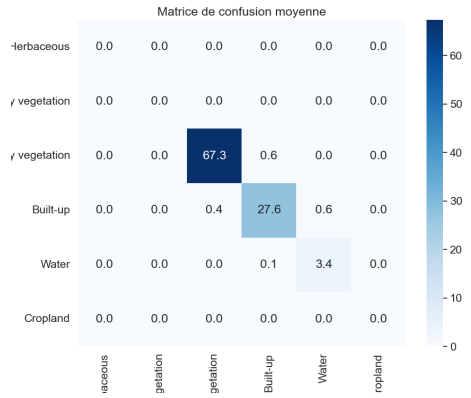

FIGURE 3.9 – Matrice de confusion du modèle random forest

Chapitre 4

Conclusion et perspectives

Le Bagging est un algorithm proposer par Breiman (1996) pour améliorer les modèles d'apprentissage et en contrôler les surapprentissage. Dans ce projet et avec des données GeoTIFF issue du satellite sentinels 2, nous avons cherché à évaluer les performances d'un modèle d'arbre de décision pour la classification d'image satellite en identifiant les éléments (eau, forêt etc.) qui ont préalablement été étiqueté à partir de leurs profile spectrale, pour ensuite améliorer ce modèle avec l'Algorithm Bagging. Il apparait le modèle d'arbre de décision fourni déjà des performances exceptionnelle (plus 99% de précision) pour une classification orientée pixel et donc l'utilisation du Bagging n'apporte rien de significatif.

Par ailleurs, le résultat final voulu qui consistait à fournir une carte qui délimite par coloration des pixels les différentes zone d'une carte en fonction des classe prédîtes n'a malheureusement pas pu être atteint dans le temps à disposition.

Une piste serait de finaliser le script sur ce point pour visualiser et comparer les performances des modèles. On pourrait aussi comparer ces modèles à d'autres méthodes statistiques pour de la télédétection.