

2025

UE 901_21

PROJET TÉLÉDÉTECTION

M2 SIGMA

BERTHE Kadiatou Hawa
NIANG Aby
RODRIGUES Océane



SOMMAIRE

Introduction (p.)	1
Données (p.)	2
Description de la méthode (p.)	3
Résultats et Analyses (p.)	4
Conclusion & Discussion des résultats (p.)	5

INTRODUCTION

La télédétection est une technique essentielle pour l'observation et l'analyse des surfaces terrestres, permettant de collecter des données à distance sans contact direct avec les objets étudiés.

Elle est largement utilisée dans divers domaines, notamment la cartographie, la surveillance environnementale et la gestion des ressources naturelles. Dans ce projet, nous nous concentrerons sur l'utilisation de la télédétection pour la cartographie des essences forestières.

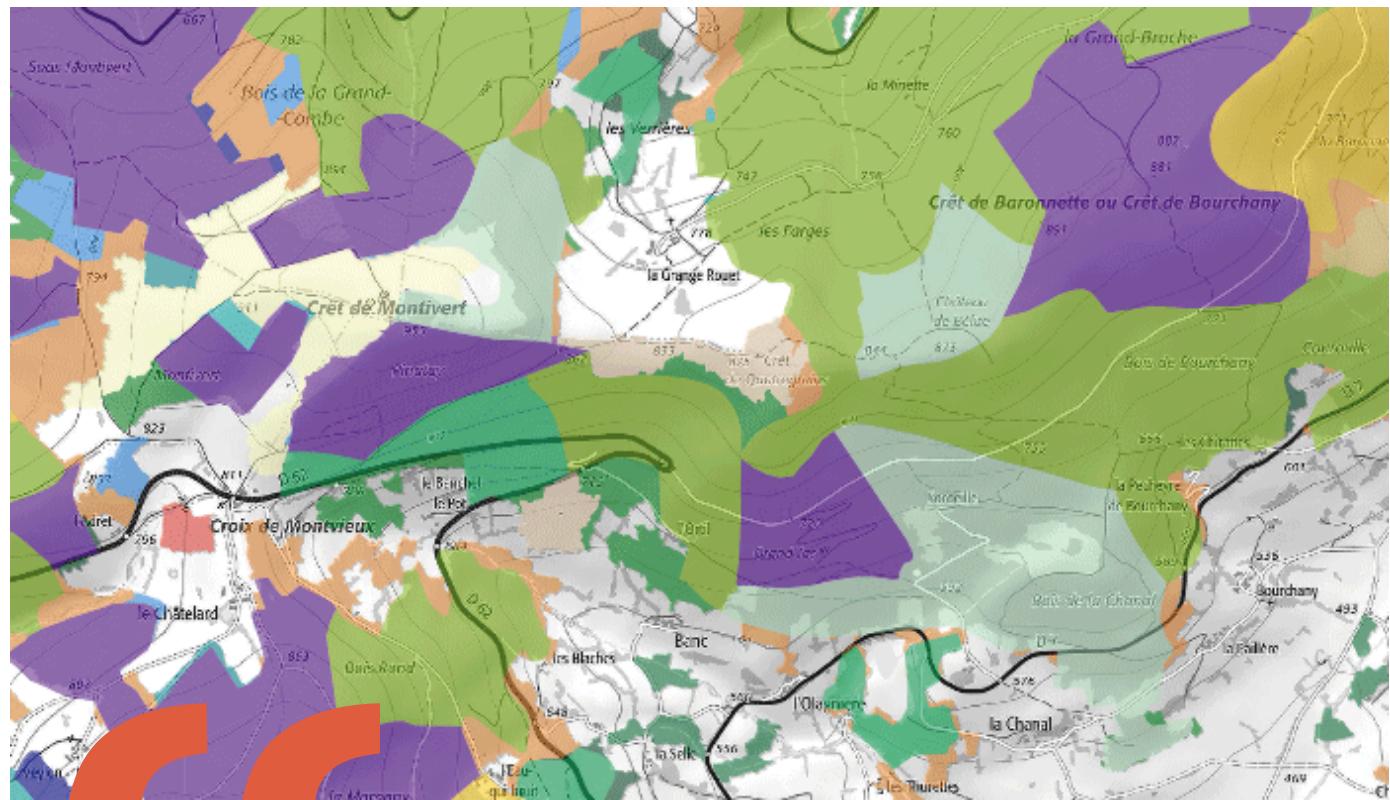


INTRODUCTION

La BD Forêt version 2.0 est une base de données de référence pour l'espace forestier et les milieux semi-naturels en France. Elle décrit les formations végétales forestières et naturelles en se basant sur la couverture du sol, la densité du couvert, la composition et l'essence dominante pour des éléments de plus de 5 000 m². Cependant, cette base de données n'est pas mise à jour régulièrement et ne permet pas de cartographier les essences forestières à une échelle intra-peuplement.

L'objectif de ce projet est d'évaluer si la BD Forêt version 2.0 peut être utilisée comme source d'échantillons de référence pour réaliser une classification supervisée de séries temporelles d'images Sentinel-2. Nous réaliserons une classification supervisée à l'échelle du pixel, puis l'information sera agrégée à l'échelle des peuplements de la BD Forêt.

Enfin, nous évaluerons la qualité des cartes produites à ces deux échelles.



BD Forêt® IGN

DONNÉES DE L'ÉTUDE



CHOIX DES DONNÉES

Les données d'entrées utilisées dans ce projet comprennent :

- Une série temporelle d'images Sentinel-2 disponibles sur la plateforme THEIA ;
- La BD Forêt® V2.0 de l'IGN (fournies pour le projet) ;
- Le masque de l'emprise de Toulouse et ses environs ;
- L'emprise de la tuile T31TCJ.

Série temporelle d'images Sentinel-2

Ces images multispectrales ont différentes résolution spatiale de 10 mètres à 20m et couvrent plusieurs bandes spectrales, notamment le bleu (Bande 2), le vert (Bande 3), le rouge (Bande 4) et le proche infrarouge (Bande 8). Nous avons travaillé avec les 10 bandes FRE (B2, B3, B4, B5, B6, B7, B8, BA, B11 et B12).

Elles ont été choisies selon les spécifications suivantes :

- Acquises entre janvier 2022 et février 2023 :
 - une date en hiver :
 - 09/02/2022 ;
 - quatre dates en automne/printemps :
 - 22/09/2022 .
 - 11/11/2022 ;
 - 26/03/2022 ;
 - 05/04/2022 ;
 - une date en été :
 - 03/08/2022 .
- Correspondent à l'emprise de Toulouse et ses environs (la tuile T31TCJ) ;
- Possède une couverture de nuages inférieure à 15% ;
- Sont en réflectance de niveau 2A.

Le téléchargement des images a été une étape compliquée, principalement à cause de la taille des fichiers, qui étaient très lourds. De plus, après un certain temps, il était nécessaire de se reconnecter au site et de réappliquer les filtres pour poursuivre les téléchargements, ce qui a ralenti le processus. Une autre difficulté a été de vérifier que toutes les images contenaient bien les 10 bandes spectrales nécessaires à notre analyse. Certaines étaient incomplètes, ce qui nous a obligés à les retélécharger plusieurs fois afin d'obtenir l'ensemble des bandes requises.

ENVIRONNEMENT DE TRAVAIL

Pour ce projet, nous avons principalement travaillé sur Onyxia Jupyter, un environnement de travail en ligne qui nous a permis de coder en Python et de manipuler les données de manière collaborative.

Au début, chacun de nous a travaillé sur des parties spécifiques du projet, puis nous avons comparé nos résultats pour nous assurer que tout le monde comprenait bien chaque étape. Cela nous a permis de mieux nous approprier le projet dans son ensemble. Ensuite, nous avons fait le tri pour sélectionner les meilleurs éléments de chaque code avant de les rassembler et de les sauvegarder sur notre dépôt Git. Cette méthode de travail nous a aidés à avancer de manière efficace tout en apprenant les uns des autres.

L'ensemble des scripts de ce projet sont disponibles sur le dépôt GIT Hub de notre groupe :

https://github.com/niangaby/Projet_teledetection

Les données (images) sont quant à elles stockées dans un serveur de fichiers Onyxia : diffusion/images/

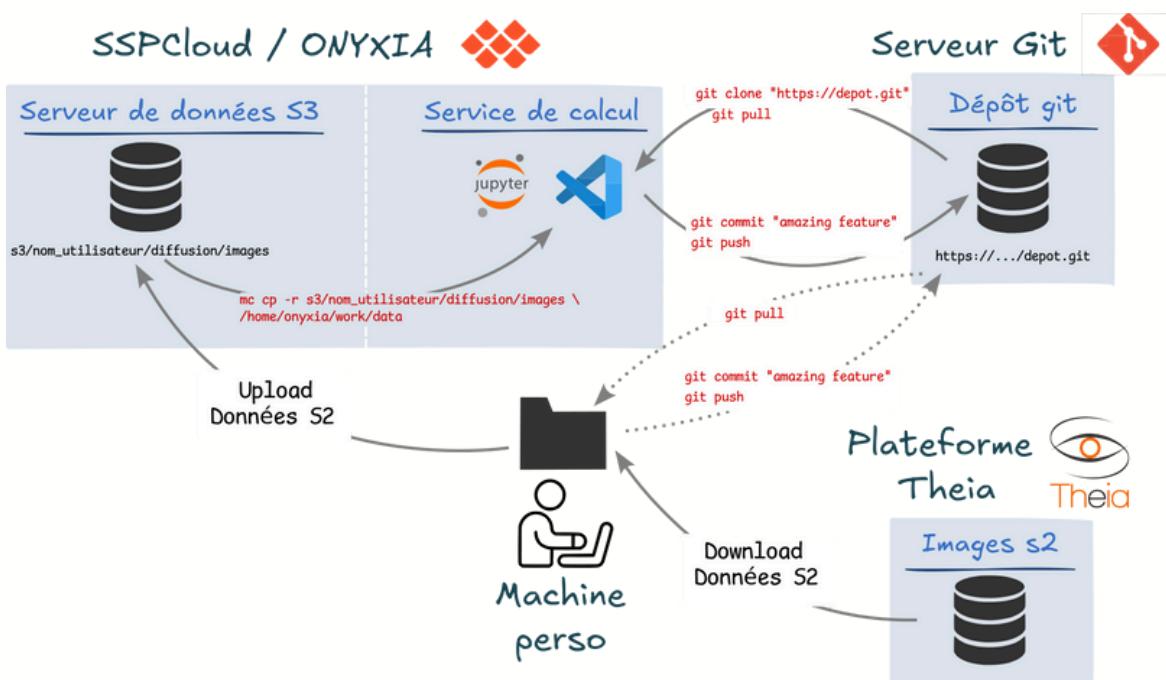


Figure n°1 : Schématisation de l'organisation de stockage et récupération de nos ressources
(sources : <https://mlang.frama.io/>)

DESCRIPTION DE LA MÉTHODE

03

MÉTHODOLOGIE DU PROJET

La méthodologie suivie dans le cadre de ce projet a été construite pour permettre de répondre à notre objectif :

“ Évaluer si la BD Forêt version 2.0 peut être utilisée comme source d'échantillons de référence pour réaliser une classification supervisée de séries temporelles d'images Sentinel-2. ”

Nous présenterons notre méthode de façon à ce que la démarche suivie soit reproductible en détaillant les traitements réalisés, les couches en entrées et les paramètres principaux choisis. Des diagrammes de flux seront également fournis pour visualiser au mieux l'enchaînement des procédés du projet.

Point clés de notre méthodologie :

- Construction d'un masque à partir de la BD Forêt® ;
- Sélection d'échantillons ;
- Analyse des échantillons ;
- Pré-traitement des images ;
- Production d'une essence forestières à l'échelle du pixel ;
- Production d'une carte à l'échelle des peuplements ;
- Analyse des résultats.

CONSTRUCTION D'UN MASQUE DE FORET

Diagramme de flux

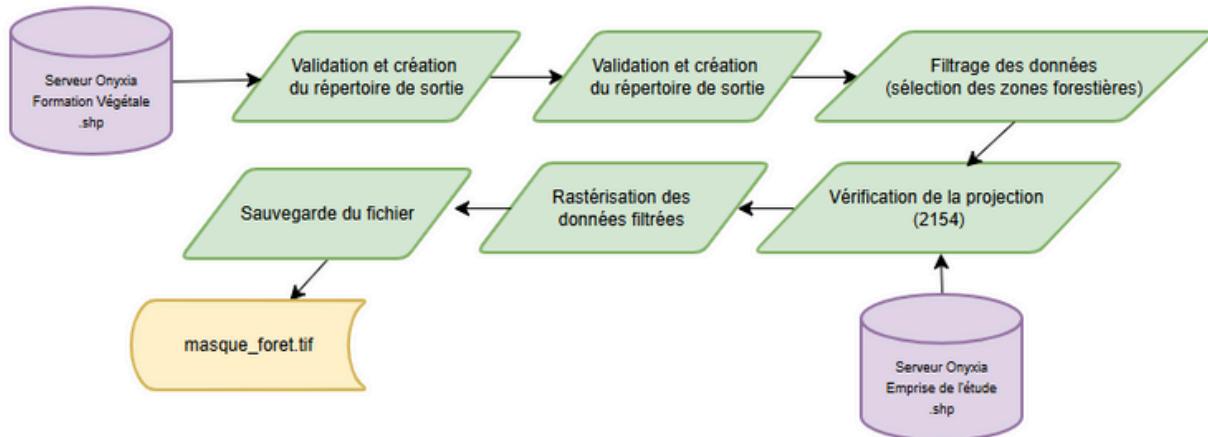


Figure n°2 : Diagramme de flux du workflow de construction d'un masque

Dans cette première étape du projet, l'objectif est de construire un masque raster permettant de délimiter les zones de forêt à classer et celles à exclure de l'analyse. Ce masque est généré à partir des données de la BD Forêt® nommée FORMATION VÉGÉTALE. Cette couche est filtrée, découpée avec l'emprise d'étude puis rastérisée, en tenant compte des critères définis dans l'énoncé.

Objectif du masque

Le masque a pour rôle d'identifier les surfaces forestières pertinentes pour l'analyse, en excluant certains types de végétation qui ne correspondent pas à des formations forestières exploitées dans le cadre de cette étude.

Plus précisément, nous retenons uniquement les polygones forestiers de la BD Forêt, en excluant :

- Les landes et formations herbacées, qui ne sont pas considérées comme des formations forestières.
- La forêt ouverte, qui ne correspond pas aux critères d'analyse.
- La forêt fermée sans couvert arboré, qui ne présente pas un couvert forestier exploitable pour cette classification.

Caractéristiques techniques du masque

Le masque doit être conforme aux spécifications suivantes :

- Format raster GeoTiff.
- Encodage en 8 bits (valeurs binaires : forêt ou non-forêt).
- Emprise et résolution identiques aux images Sentinel-2 utilisées après le prétraitement.
- Fichier de sortie nommé masque_foret.tif et stocké dans le répertoire : results/data/img_pretraitees/
- Valeurs des pixels :
 - 1 : Zone forestière
 - 0 : Zone hors forêt

Méthodologie de construction

Pour générer ce masque, plusieurs étapes sont mises en œuvre :

1. Chargement et filtrage des données de la BD Forêt®

- Le fichier FORMATION_VEGETALE.shp est ouvert et traité pour ne conserver que les zones forestières pertinentes.
- Une vérification est effectuée pour s'assurer que les données contiennent bien des entités exploitables.

2. Définition de l'emprise d'étude

- Le fichier emprise_etude.shp est utilisé pour définir la zone d'analyse.
- Le système de projection de l'emprise est vérifié afin d'assurer la cohérence spatiale du masque généré.

3. Création du raster de sortie

- Un raster vierge est généré avec les dimensions et la résolution correspondant aux images Sentinel-2 utilisées.
- Il est initialisé pour recevoir les valeurs attribuées lors de la rastérisation.

4. Rastérisation des zones forestières

- Les polygones retenus après filtrage sont convertis en pixels du raster.
- Les valeurs sont attribuées en fonction de la nomenclature définie : forêt (1), hors forêt (0).

5. Sauvegarde et validation du masque

- Le fichier masque_foret.tif est produit et enregistré dans le répertoire cible.
- Un contrôle est réalisé pour s'assurer que la rastérisation s'est bien déroulée et que le fichier final est exploitable.

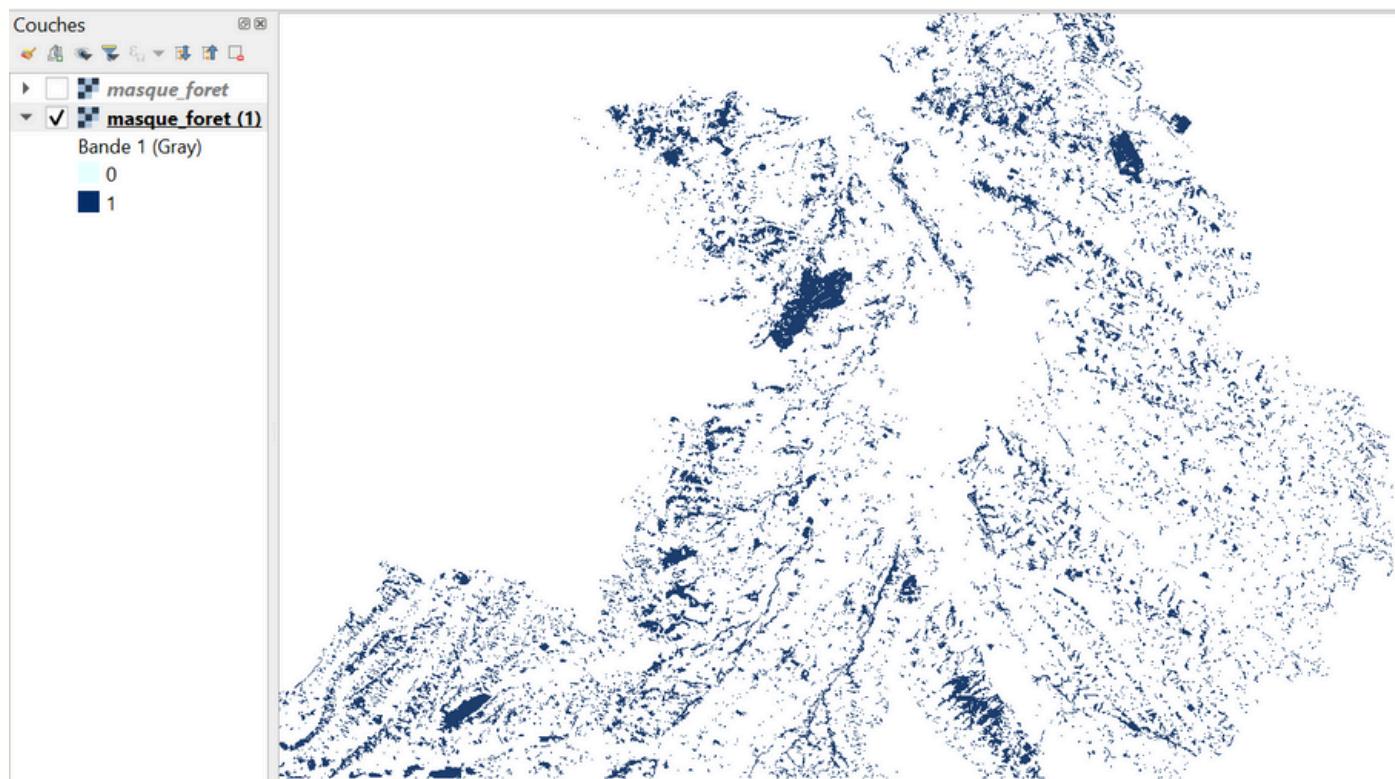


Figure n°3 : Résultat de sortie du masque

Conclusion

La construction de ce masque constitue une étape essentielle du projet, car elle conditionne la qualité des classifications futures. Un masque bien construit garantit que seules les zones d'intérêt seront analysées dans la suite du projet, réduisant ainsi les erreurs liées à des zones non pertinentes.

SÉLECTION D'ÉCHANTILLONS

Dans cette 2e phase du projet, l'objectif est de sélectionner des échantillons pertinents à partir de la BD Forêt pour entraîner et valider la classification supervisée.

Cette sélection repose sur une correspondance entre les classes de la BD Forêt et deux niveaux de nomenclatures cartographiques :

1. Classification à l'échelle du pixel
2. Classification à l'échelle des peuplements (agrégation de la classification pixel)

Objectif de la sélection des échantillons

Les échantillons sélectionnés serviront de données d'apprentissage et de validation pour la classification des images Sentinel-2.

Tous les polygones de la BD Forêt ne sont pas utilisés, car :

- Seules les classes listées dans la nomenclature cible sont retenues.
- Certaines classes, comme "Forêt fermée de feuillus purs en îlots", sont exclues à l'échelle du pixel mais incluses à l'échelle des peuplements.

Le fichier final attendu est un vecteur contenant les polygones sélectionnés enrichi avec des attributs spécifiques.

En effet, le fichier final Sample_BD_foret_T31TCJ.shp doit contenir plusieurs attributs, dont certains proviennent directement de la BD Forêt, tandis que d'autres sont ajoutés pour structurer la classification.

Voici une description détaillée de ces attributs :

Attributs hérités de la BD Forêt®

Ces attributs proviennent directement du fichier FORMATION_VEGETALE.shp et doivent être conservés pour garantir une traçabilité des données.

ATTRIBUT	DESCRIPTION
CODE_TFV	Code de la classe forestière selon la BD Forêt
TFV	Nom de la classe forestière selon la BD Forêt
Surface	Superficie du polygone en m ²

Tableau n° 1 : Attributs hérités de la BD Forêt®

Pourquoi les conserver ?

- Ils permettent de vérifier l'origine des données et d'assurer la correspondance avec la nomenclature officielle.
- La surface est utile pour analyser la représentativité des échantillons.

Attributs ajoutés pour la classification

Ces nouveaux attributs sont créés et attribués à chaque polygone sélectionné pour structurer les deux niveaux de classification.

1. Attributs pour la Classification à l'Échelle du Pixel

Ces informations permettent d'identifier la classe à laquelle appartient chaque pixel issu de la classification supervisée.

ATTRIBUT	DESCRIPTION
Code_Pixel	Code numérique correspondant à la classe de classification du pixel
Nom_Pixel	Nom de la classe en toutes lettres

Tableau n° 2 : Attributs pour la Classification à l'Échelle du Pixel

Pourquoi ces attributs ?

- Code_Pixel : est utilisé pour entraîner et valider le modèle de classification.
- Nom_Pixel : facilite la lecture et l'interprétation des résultats.

2. Attributs pour la Classification à l'Échelle des Peuplements

Ces informations servent à agréger les résultats de la classification pixel à l'échelle des peuplements forestiers.

ATTRIBUT	DESCRIPTION
Code_Objet	Code numérique correspondant à la classification des peuplements
Nom_Objet	Nom de la classe forestière au niveau du peuplement

Tableau n° 3 : Attributs pour la Classification à l'Échelle des Peuplements

Pourquoi ces attributs ?

- Code_Objet est utilisé pour l'agrégation des pixels en unités homogènes de peuplement.
- Nom_Objet facilite l'interprétation des résultats et leur comparaison avec la BD Forêt.

Exclusion de certaines classes

Certaines classes de la BD Forêt ne sont pas prises en compte pour l'une ou l'autre des classifications :

- Par exemple, la classe Forêt fermée de feuillus purs en îlots (FF1-OO) est :
 - Exclue de la classification pixel (symbole Ø).
 - Incluse dans la classification des peuplements.

Pourquoi ce choix ?

- Il permet d'adapter la classification à l'échelle d'analyse appropriée.
- Certaines formations sont trop petites pour être identifiables à l'échelle du pixel, mais cohérentes en peuplement.

Données utilisées

Les données utilisées pour cette étape sont :

- BD Forêt (FORMATION_VEGETALE.shp) : Contient les polygones des formations forestières en France, avec des attributs détaillant leur type et leur composition.
- Emprise d'étude (emprise_etude.shp) : Délimite la zone d'intérêt, qui correspond à la couverture des images Sentinel-2 traitées dans le projet.

Bibliothèques utilisées

Les outils employés sont sélectionnés pour leur efficacité dans la manipulation de données géospatiales :

BIBLIOTHÈQUE	RÔLE
Geopandas	Gestion et manipulation des données vectorielles (shapefiles)
GDAL/OGR	Lecture/écriture avancée des fichiers SIG
Pandas	Structuration et manipulation des attributs associés aux polygones

Tableau n° 4 : Liste des bibliothèques utilisées

Caractéristiques du fichier de sortie

Le jeu d'échantillons produit doit :

- Être inclus dans l'emprise spatiale des images Sentinel-2 après découpage.
- Contenir uniquement les classes validées pour la classification.

- Conserver les attributs d'origine et ajouter :
 - Nom : Nom de la classe forestière
 - Code : Code correspondant dans la nomenclature

Format de sortie :

- Nom du fichier : Sample_BD_foret_T31TCJ.shp
- Format : Shapefile (ESRI .shp)
- Chemin de stockage : results/data/sample/

Méthodologie de sélection

Diagramme de flux

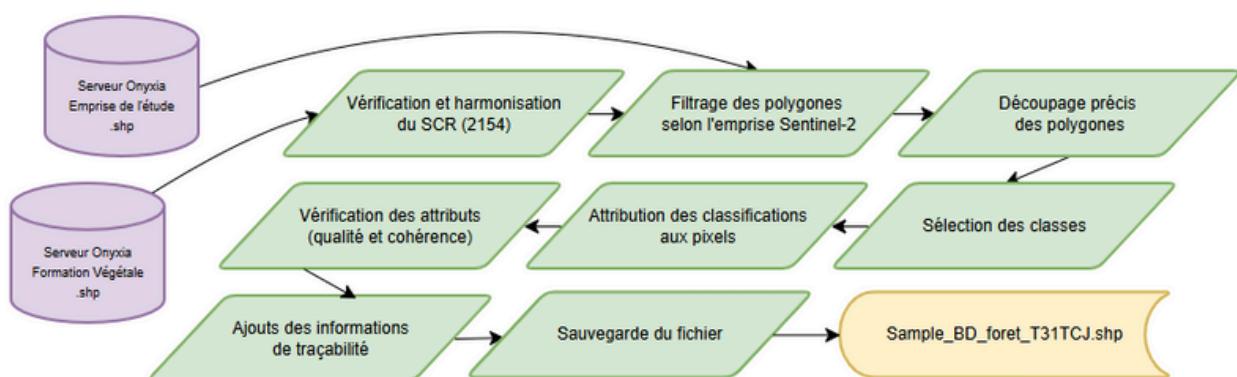


Figure n° 4 : Diagramme de flux du workflow de la sélection d'échantillons

1. Chargement et harmonisation des données

- Les fichiers BD Forêt® et emprise d'étude sont chargés en mémoire.
- Une vérification des systèmes de projection (CRS) est réalisée pour éviter toute erreur d'alignement spatial.

2. Sélection des polygones d'intérêt

- Filtrage par emprise : Seuls les polygones intersectant la zone d'étude sont conservés.
- Découpage précis : On utilise un découpage des polygones à l'emprise afin d'assurer une cohérence géométrique.

3. Filtrage des classes de la BD Forêt®

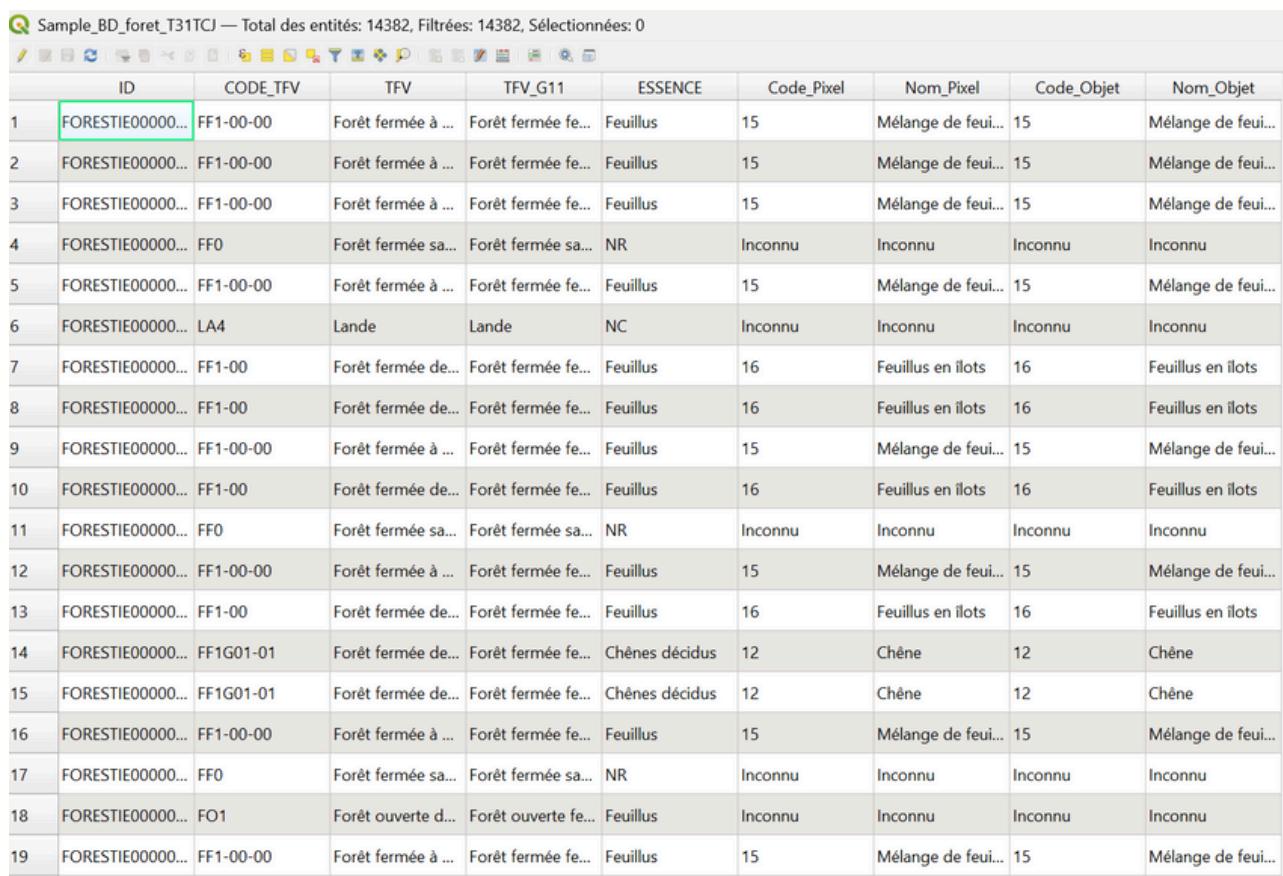
- Seules les classes nécessaires à la classification sont conservées.
- Un mapping personnalisé est appliqué pour attribuer les codes des deux niveaux de classification (pixel et peuplement).

4. Attribution des codes et noms

- Un dictionnaire de correspondance est utilisé pour associer chaque polygone à un Code_Pixel et un Code_OBJET.
- Les polygones non concernés par l'une des classifications reçoivent une valeur neutre ("Inconnu").

5. Vérification et sauvegarde

- Un contrôle est effectué sur :
 - Le nombre de polygones sélectionnés.
 - La présence correcte des attributs.
 - La cohérence des valeurs attribuées.
- Le fichier est ensuite enregistré en Shapefile (.shp) pour être exploité dans la suite du projet.



The screenshot shows a QGIS interface with a table titled "Sample_BD_foret_T31TCJ". The table has columns: ID, CODE_TFV, TFV, TFV_G11, ESSENCE, Code_Pixel, Nom_Pixel, Code_Objet, and Nom_Objet. The first column contains sample IDs starting from 1. The second column contains codes like FF1-00-00, FF1G01-01, etc. The third column contains descriptions like "Forêt fermée à ...". The fourth column contains "Forêt fermée fe...". The fifth column contains "Feuillus". The sixth column contains values like 15, 16, 12, etc. The seventh column contains "Mélange de feui...". The eighth column contains "15". The ninth column contains "Mélange de feui...". The table shows 19 rows of data.

ID	CODE_TFV	TFV	TFV_G11	ESSENCE	Code_Pixel	Nom_Pixel	Code_Objet	Nom_Objet
1	FORESTIE00000...	FF1-00-00	Forêt fermée à ...	Forêt fermée fe...	Feuillus	15	Mélange de feui...	15
2	FORESTIE00000...	FF1-00-00	Forêt fermée à ...	Forêt fermée fe...	Feuillus	15	Mélange de feui...	15
3	FORESTIE00000...	FF1-00-00	Forêt fermée à ...	Forêt fermée fe...	Feuillus	15	Mélange de feui...	15
4	FORESTIE00000...	FF0	Forêt fermée sa...	Forêt fermée sa...	NR	Inconnu	Inconnu	Inconnu
5	FORESTIE00000...	FF1-00-00	Forêt fermée à ...	Forêt fermée fe...	Feuillus	15	Mélange de feui...	15
6	FORESTIE00000...	LA4	Lande	Lande	NC	Inconnu	Inconnu	Inconnu
7	FORESTIE00000...	FF1-00	Forêt fermée de...	Forêt fermée fe...	Feuillus	16	Feuillus en îlots	16
8	FORESTIE00000...	FF1-00	Forêt fermée de...	Forêt fermée fe...	Feuillus	16	Feuillus en îlots	16
9	FORESTIE00000...	FF1-00-00	Forêt fermée à ...	Forêt fermée fe...	Feuillus	15	Mélange de feui...	15
10	FORESTIE00000...	FF1-00	Forêt fermée de...	Forêt fermée fe...	Feuillus	16	Feuillus en îlots	16
11	FORESTIE00000...	FF0	Forêt fermée sa...	Forêt fermée sa...	NR	Inconnu	Inconnu	Inconnu
12	FORESTIE00000...	FF1-00-00	Forêt fermée à ...	Forêt fermée fe...	Feuillus	15	Mélange de feui...	15
13	FORESTIE00000...	FF1-00	Forêt fermée de...	Forêt fermée fe...	Feuillus	16	Feuillus en îlots	16
14	FORESTIE00000...	FF1G01-01	Forêt fermée de...	Forêt fermée fe...	Chênes décidus	12	Chêne	12
15	FORESTIE00000...	FF1G01-01	Forêt fermée de...	Forêt fermée fe...	Chênes décidus	12	Chêne	12
16	FORESTIE00000...	FF1-00-00	Forêt fermée à ...	Forêt fermée fe...	Feuillus	15	Mélange de feui...	15
17	FORESTIE00000...	FF0	Forêt fermée sa...	Forêt fermée sa...	NR	Inconnu	Inconnu	Inconnu
18	FORESTIE00000...	FO1	Forêt ouverte d...	Forêt ouverte fe...	Feuillus	Inconnu	Inconnu	Inconnu
19	FORESTIE00000...	FF1-00-00	Forêt fermée à ...	Forêt fermée fe...	Feuillus	15	Mélange de feui...	15

Figure n° 5 : Résultats de sortie de sélection des échantillons

ANALYSE DES ÉCHANTILLONS SÉLECTIONNÉS

Cette 3e étape du projet vise à analyser les échantillons sélectionnés dans le fichier Sample_BD_foret_T31TCJ.shp créé précédemment, en se concentrant sur la répartition des polygones et des pixels par classe.

Objectif de l'analyse

L'objectif est d'obtenir une meilleure compréhension de la distribution spatiale des échantillons avant la classification.

Notre analyse repose sur trois types de visualisations :

1. Diagramme en bâtons : Nombre de polygones par classe

- Permet d'identifier la répartition des polygones par type de couverture forestière.

2. Diagramme en bâtons : Nombre de pixels par classe

- Évalue l'importance relative de chaque classe en fonction de la surface totale des polygones.

3. Violin plot : Distribution du nombre de pixels par polygone, par classe

- Met en évidence la variabilité des tailles des polygones au sein de chaque classe.

Toutes les figures sont enregistrées dans le fichier : results/figure/

Données et Prétraitement

La donnée analysée est le fichier de sortie précédent Sample_BD_foret_T31TCJ.shp, qui contient les polygones sélectionnés pour la classification.

Seules les classes valides définies dans l'énoncé sont conservées, tandis que les polygones avec Code_Pixel = "Inconnu" sont exclus pour éviter d'introduire des biais.

Création des graphiques

Les graphiques sont générés à l'aide de Matplotlib ou Plotly (choix interactif disponible dans le script).

1. Diagramme en Bâtons : Nombre de Polygones par Classe

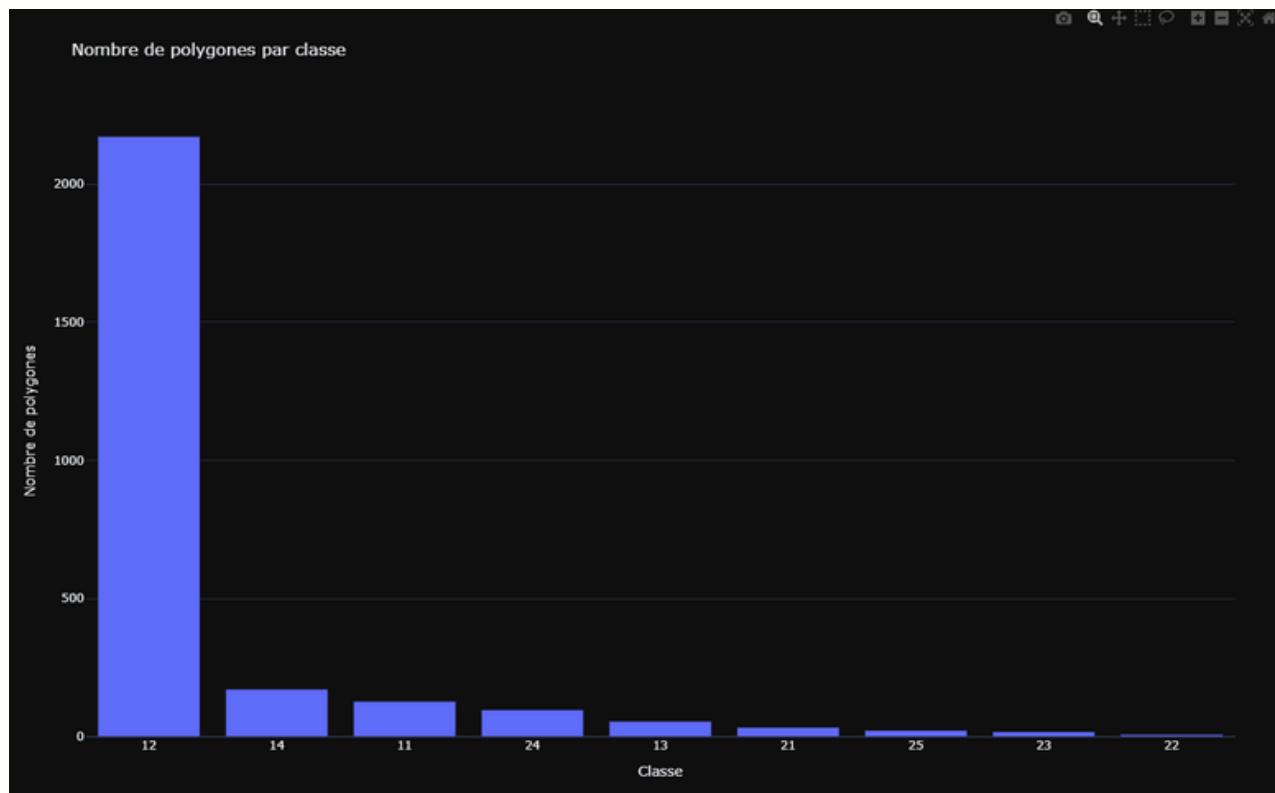


Figure n° 6 : Diagramme en Bâtons - Nombre de Polygones par Classe

- Représente le nombre de polygones par classe forestière.
 - Met en évidence les classes les plus et les moins représentées.
- Il permet de s'assurer que les classes sont suffisamment représentées pour la classification.

2. Diagramme en Bâtons : Nombre de Pixels par Classe

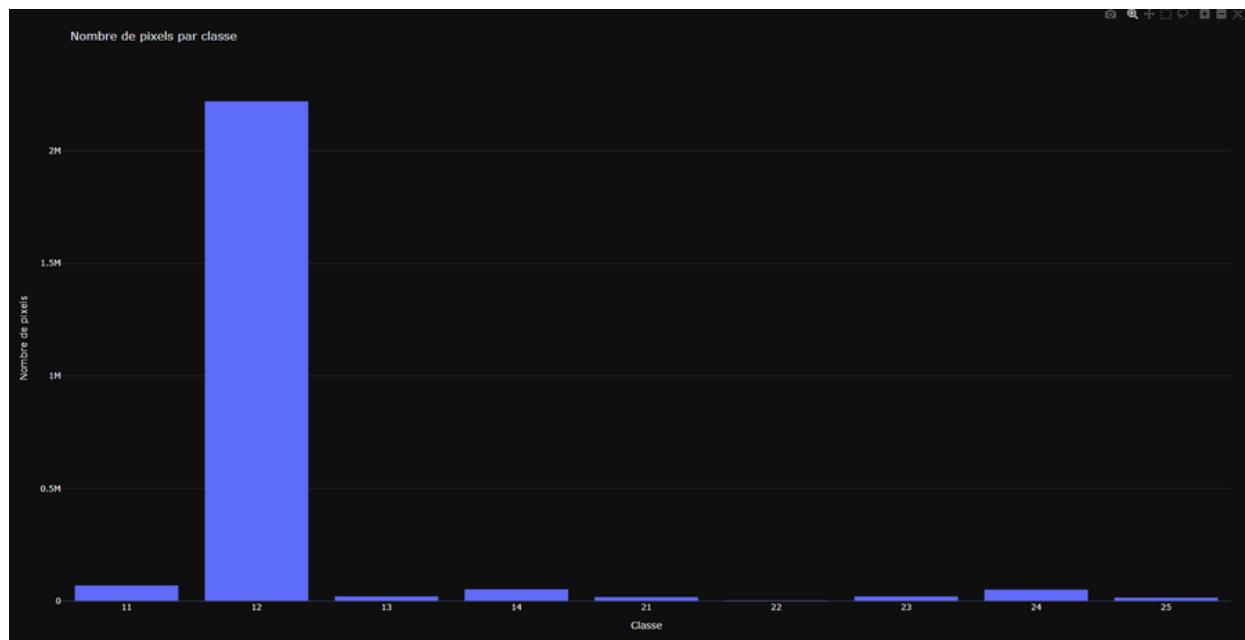


Figure n° 7 : Diagramme en Bâtons - Nombre de Pixels par Classe

- Montre la quantité totale de pixels appartenant à chaque classe.
- Corrige les biais liés aux polygones de petite taille, qui pourraient être nombreux mais représenter une faible surface.

Il est essentiel pour évaluer la poids réel de chaque classe dans la classification.

3. Violin Plot : Distribution des Pixels par Polygone, par Classe

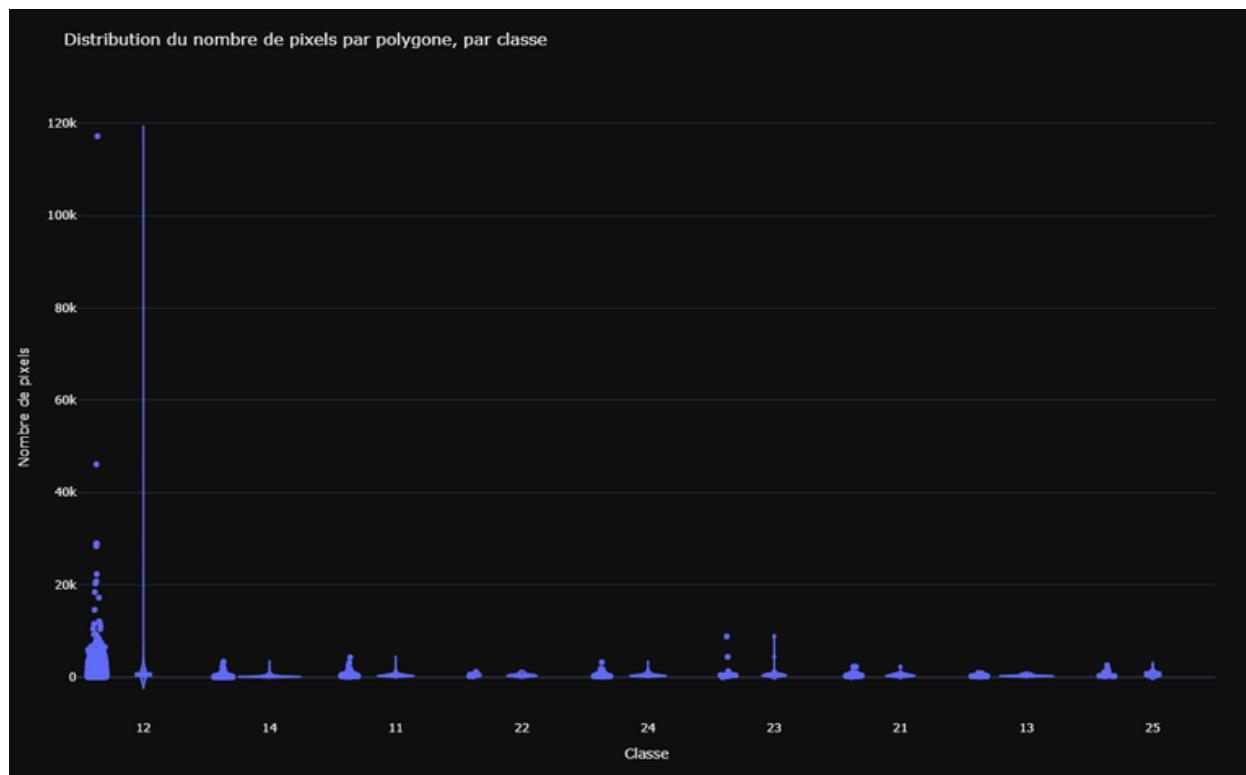


Figure n° 8 : Violin Plot - Distribution des Pixels par Polygone, par Classe

- Affiche la variabilité des tailles des polygones au sein de chaque classe.
- Permet d'identifier des classes homogènes (peu de variation) ou hétérogènes (forte variabilité).
- Une grande variabilité intra-classe peut indiquer une hétérogénéité dans la structure des peuplements, ce qui peut compliquer la classification.

Explication des choix techniques réalisés

Bibliothèques utilisées

BIBLIOTHÈQUE	RÔLE
Geopandas	Gestion et manipulation des données vectorielles (shapefiles)
GDAL/OGR	Lecture/écriture avancée des fichiers SIG
Pandas	Structuration et manipulation des attributs associés aux polygones
matplotlib	bibliothèque de visualisation en Python
os	Permet d'interagir avec le système d'exploitation
pyplot	permettant de créer des visualisations interactives facilement.
numpy	permet le traitement de données numériques et scientifiques en Python

Tableau n° 5 : Liste des bibliothèques utilisées

Paramètres Clés

- Filtrage des classes : Seules les classes spécifiées sont retenues.
- Taille de pixel fixe (10m x 10m) : Assure une estimation standardisée du nombre de pixels.
- Format des graphiques : .html avec interactivité activée.

PRÉ-TRAITEMENT DES IMAGES

L'objectif de cette étape est de préparer les images Sentinel-2 pour une analyse cohérente et homogène. Le prétraitement comprend plusieurs opérations : sélection des bandes, découpage, reprojection, rééchantillonnage et masquage des zones non forestières.

Ces transformations garantissent que les images sont prêtes pour l'analyse suivante.

Objectifs et Résultats Attendus

Le prétraitement se décompose en deux objectifs principaux :

1. Création d'une image multi-bandes :

- Contient les 10 bandes spectrales (bandes 2, 3, 4, 5, 6, 7, 8, 9, 11, 12) pour 6 dates → soit 60 bandes au total.
- Découpée selon l'emprise du fichier emprise_etude.shp.
- Rééchantillonnée à 10m de résolution.
- Projettée en Lambert 93 (EPSG:2154).
- Masquage des zones non forestières avec masque_foret.tif.
- Encodée en uint16 (ou uint8 si nécessaire).
- Valeur NoData = 0.

Fichier de sortie :

results/data/img_pretaitees/Serie_temp_S2_allbands.tif

Création d'une série temporelle de NDVI :

- Même traitement que l'image multi-bandes.
- Conversion des images en indices NDVI pour chaque date.
- Encodage en float32, avec NoData = -9999.

Fichier de sortie :

results/data/img_pretaitees/Serie_temp_S2_ndvi.tif

Données et Paramètres Utilisés

Données en entrée :

- Images Sentinel-2 (.tif) : Ensemble d'images multi-temporelles.
- Emprise (emprise_etude.shp) : Délimite la zone d'intérêt.
- Masque de forêt (masque_foret.tif) : Exclut les zones hors-forêt.

Paramètres essentiels :

PARAMÈTRE	VALEUR
Projection	EPSG:2154 (Lambert 93)
Résolution spatiale	10m
Format d'encodage	uint16 ou uint8 (optionnel)
Valeur NoData	0 pour l'image multi-bandes, -9999 pour NDVI

Tableau n° 6 : Liste des paramètres essentiels renseignés

Méthodologie de Prétraitement

Le script pre_traitement.py suit un workflow organisé en étapes distinctes, garantissant la reproductibilité.

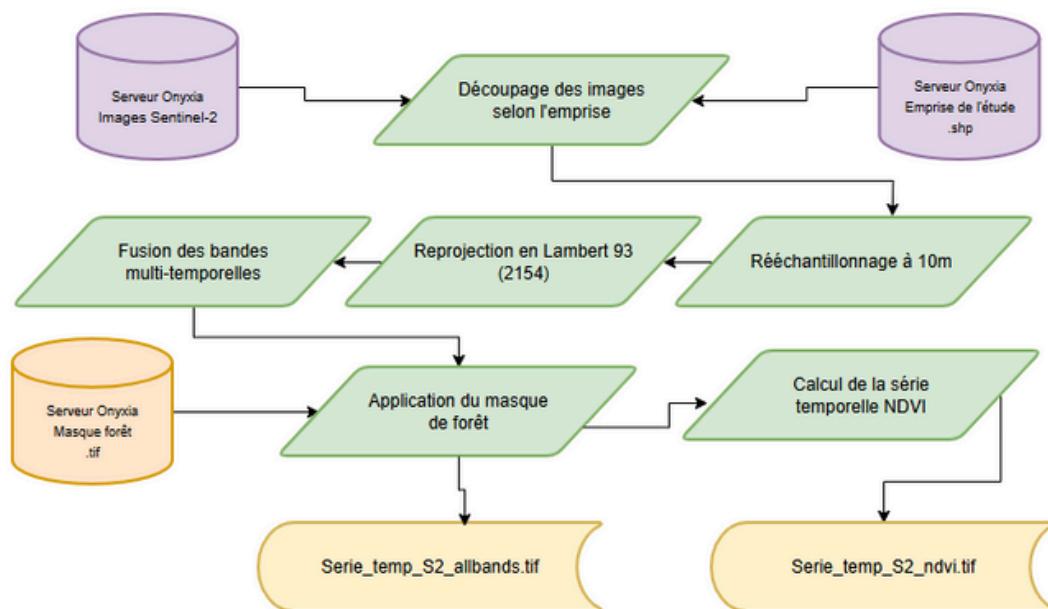


Figure n° 9 : Diagramme de flux du workflow des pré-traitements

1. Sélection et Chargement des Données

Les fichiers d'images sont identifiés et chargés depuis le dossier source

2. Découpage selon l'Emprise

L'image est découpée pour correspondre strictement à l'emprise emprise_etude.shp. Cela évite d'inclure des pixels inutiles.

3. Rééchantillonnage à 10m

Les images Sentinel-2 comportent des bandes à différentes résolutions (10m, 20m, 60m).

Un rééchantillonnage est appliqué pour homogénéiser la résolution à 10m.

4. Reprojection en Lambert 93

- Certaines images peuvent être en WGS84 (EPSG:4326) ou un autre CRS.
- Une reprojection en EPSG:2154 est effectuée pour assurer l'uniformité géographique.

5. Fusion des Bandes Multi-temporelles

- Une pile d'images multi-bandes est créée, combinant les 10 bandes sélectionnées pour chaque date.
- Le fichier final contient 60 bandes au total.

6. Application du Masque de Forêt

Le masque de forêt (masque_foret.tif) est appliqué :

- Les pixels hors-forêt sont mis à 0 (ou -9999 pour NDVI).
- Cela garantit que seuls les pixels forestiers sont pris en compte dans les analyses ultérieures.

7. Conversion des Images en NDVI

Le NDVI est calculé pour chaque date :

$$NDVI = \frac{NIR - RED}{NIR + RED}$$

- Bandes utilisées :
 - Rouge (RED) = Bande 4
 - Proche Infrarouge (NIR) = Bande 8
- Le fichier final contient une série temporelle d'indices NDVI pour chaque date.

8. Sauvegarde des Résultats

- Serie_temp_S2_allbands.tif (multi-bandes Sentinel-2)
- Serie_temp_S2_ndvi.tif (série NDVI)
- Les fichiers sont stockés dans results/data/img_pretaitees/

Explication des Choix Techniques

Utilisation de Rasterio & Geopandas

- Rasterio permet la lecture, transformation et reprojection des images raster.
- Geopandas est utilisé pour gérer l'emprise et aligner les projections.

Encodage en uint16 ou uint8

- uint16 est plus précis mais génère des fichiers volumineux (~10 Go).
- Option de conversion en uint8 pour alléger le stockage (~5 Go).

Masquage des zones non forestières

- L'application du masque de forêt permet de réduire la complexité de l'analyse en excluant les zones non pertinentes.

Rééchantillonnage homogène à 10m

- Assure une cohérence spatiale entre toutes les bandes.

Projection en Lambert 93

- Norme française assurant la compatibilité avec d'autres données SIG.

Conclusion

Ce prétraitement assure que les images Sentinel-2 sont alignées, homogènes et filtrées avant d'être utilisées pour la classification.

Les fichiers finaux permettent une analyse efficace de la dynamique forestière

PRODUCTION D'UNE ESSENCE FORESTIÈRES À L'ÉCHELLE DU PIXEL

Cette 4e partie du projet consiste à produire une classification supervisée des essences forestières à l'échelle du pixel à l'aide du modèle RandomForestClassifier (RF) de Scikit-Learn.

Ce modèle repose sur un ensemble d'arbres de décision, chacun entraîné sur un sous-échantillon de données, et combine leurs prédictions pour renforcer la robustesse et limiter le sur-apprentissage. Il exploite la sélection aléatoire de caractéristiques et le bagging (Bootstrap Aggregating) pour améliorer la précision et la stabilité.

Après son entraînement sur les données disponibles, le modèle est validé par validation croisée, puis appliqué pour générer une carte des essences forestières.

Objectif de la classification

Cette classification a pour objectifs :

- Entrainer un modèle Random Forest (RF) pour attribuer une classe d'essence forestière à chaque pixel ;
- Produire une carte géoréférencée des essences à 10m de résolution ;
- Évaluer la qualité du modèle à l'aide d'une matrice de confusion.

Sorties attendues

- Carte des essences forestières :
results/data/classif/carte_essences_echelle_pixel.tif
 - Matrice de confusion :
results/figure/matrice_confusion_peuplements.png
- + Graphique de la qualité du modèle (PNG)
results/figure/_qualites_CV5folds_stratified_group_x30times.png

Paramètres du Modèle Random Forest

Le classifieur RF est paramétré selon les recommandations suivantes :

PARAMÈTRE	VALEUR
max_depth	50
oob_score	True
max_samples	0.75
class_weight	balanced
n_estimators	Valeur par défaut

Tableau n° 7 : Liste des paramètres renseignés

Pourquoi ces choix de paramétrage ?

- `max_depth = 50` : Limite la profondeur de chaque arbre, évite le surapprentissage.
- `oob_score = True` : Active le calcul d'erreur Out-Of-Bag pour une validation interne.
- `max_samples = 0.75` : Utilise 75% des échantillons pour entraîner chaque arbre.
- `class_weight = balanced` : Corrige les déséquilibres entre classes.

Stratégie de Validation

Pour garantir la robustesse du modèle, une validation croisée à 5 folds est mise en place.

Trois stratégies sont possibles :

MÉTHODE	DÉTAILS
Stratifiée	Répartition équilibrée des classes dans les folds.
Stratifiée avec contrainte spatiale	Les pixels d'un même polygone sont tous dans le même fold.
Stratifiée + répétée 30 fois	La validation est répétée 30 fois et les résultats sont moyennés.

Tableau n° 8 : Liste des stratégies envisageables

Méthodologie utilisée

La stratégie utilisée est la validation croisée à 5 folds stratifiés avec contraintes spatiales (StratifiedGroupKFold) avec également une répétition de 30 fois pour moyenner les résultats et éviter la variabilité aléatoire.

Diagramme de flux

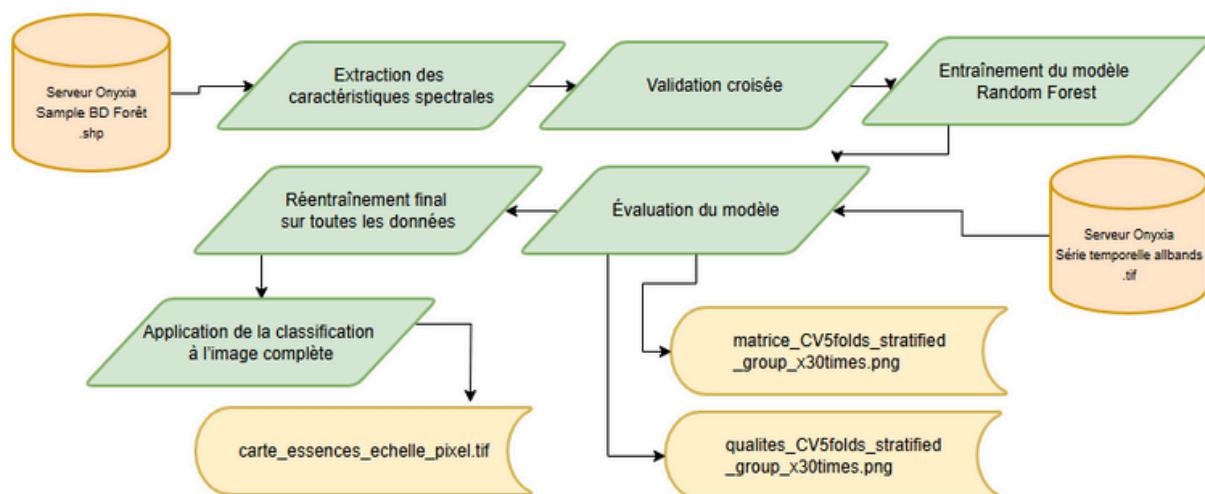


Figure n° 10 : Diagramme de flux du workflow de production d'une essence forestière à l'échelle du pixel

Processus de validation croisée :

1. Les données sont divisées en 5 sous-ensembles → 80% pour l'entraînement, 20% pour le test.
2. La répartition est stratifiée → Assure que toutes les classes sont bien représentées dans chaque fold.
3. Les pixels d'un même polygone restent dans le même fold → Empêche la fuite d'information entre l'entraînement et le test.
4. Le processus est répété 30 fois → Les résultats sont moyennés pour obtenir une évaluation stable.

Pourquoi cette approche ?

- Évite les biais spatiaux en s'assurant qu'un polygone entier ne soit pas à la fois en train et en test.
- Améliore la généralisation du modèle en l'évaluant sur plusieurs partitions des données.
- Donne une estimation plus fiable de la performance grâce à la répétition sur 30 itérations.

Explication des Choix Techniques

Utilisation de RandomForestClassifier

- Robuste et efficace pour des données haute dimension comme les images satellite.
- Gère les classes déséquilibrées avec `class_weight='balanced'`.

Validation croisée avancée

- Stratifiée et groupée par polygone pour éviter les biais d'apprentissage.
- Répétée 30 fois pour des résultats plus fiables.

Encodage en uint8 avec NoData = 0

- Optimise l'espace disque tout en gardant une bonne précision.

PRODUCTION D'UNE CARTE À L'ÉCHELLE DES PEUPLEMENTS

Cette dernière section vise à agréger les résultats de classification à l'échelle du pixel pour produire une carte des peuplements forestiers. Chaque polygone de la BD Forêt® se voit attribuer une classe de peuplement en fonction des proportions des classes d'essences présentes dans `carte_essences_echelle_pixel.tif`.

Objectif de la Classification des Peuplements

Les objectifs de cette partie sont :

- Assigner une classe de peuplement à chaque polygone de la BD Forêt® en fonction des proportions des classes d'essences forestières ;
- Produire une matrice de confusion pour évaluer la qualité de cette classification ;
- Générer un fichier vectoriel contenant les classes de peuplements assignées.

Sorties attendues

- Shapefile avec classes de peuplement assignées : `Sample_BD_foret_T31TCJ_stand.shp` ;
- Matrice de confusion : `matrice_confusion_peuplements.png` ;
- Tableau des pourcentages des classes d'essences par polygone : `pourcentages_classes_peuplement.csv`.

Méthodologie de Classification des Peuplements

Diagramme de flux

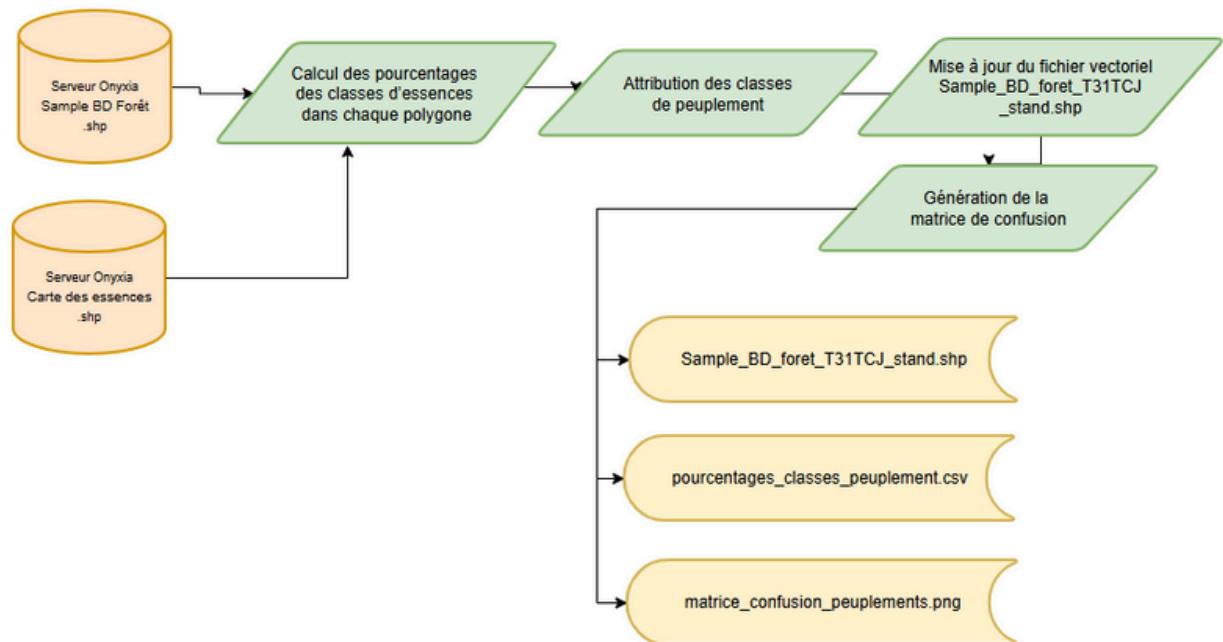


Figure n° 11 : Diagramme de flux du workflow de la classification des peuplements

Calcul des pourcentages des classes d'essences par polygone

Utilisation de Rasterstats pour croiser les polygones et les classes d'essences dans carte_essences_echelle_pixel.tif.

- Lecture du raster (rasterio.open) et du shapefile (gpd.read_file) ;
- Calcul des proportions de chaque classe dans chaque polygone ;
- Enregistrement des résultats dans un fichier CSV (pourcentages_classes_peuplement.csv).

Pourquoi cette étape ?

- Permet de convertir les informations à l'échelle du pixel en information à l'échelle des peuplements ;
- Les proportions des classes d'essences sont essentielles pour appliquer les règles de décision.

Attribution des classes de peuplement

Utilisation des règles de décision définies dans l'énoncé pour assigner une classe de peuplement.

- Chaque polygone reçoit une classe code_predit en fonction des proportions des classes d'essences ;
- Ajout d'une nouvelle colonne code_predit au shapefile Sample_BD_foret_T31TCJ.shp. qui est devenu Sample_BD_foret_T31TCJ_stand.shp.

Pourquoi cette étape ?

- Transforme une classification pixel en une classification objet adaptée aux peuplements.
- Garantit une correspondance avec la nomenclature de la BD Forêt®.

Production de la matrice de confusion

Comparaison entre les classes réelles (Code_Objet) et prédites (code_predit)

- Matrice de confusion générée avec sklearn.metrics.confusion_matrix
- Affichage des performances en identifiant les classes les plus et les moins bien prédites.
- Enregistrement de la matrice en image (matrice_confusion_peuplements.png).

Pourquoi cette étape ?

- Évalue la précision du passage de l'échelle pixel à l'échelle peuplement.
- Permet d'identifier les classes mal classées et d'ajuster les règles de décision.

Explication des Choix Techniques

Utilisation de Rasterstats

- Permet de croiser efficacement un raster de classification et un shapefile de polygones.
- Facile à utiliser pour calculer des statistiques par polygone.

Mise en place des règles de décision pour assigner une classe de peuplement

- Assure une conversion robuste entre l'échelle pixel et peuplement.
- Garantie d'une cohérence avec la nomenclature BD Forêt.

Utilisation de la matrice de confusion

- Permet une évaluation fiable de la classification.
- Identifie les classes les plus sujettes aux erreurs.

Conclusion

Ce traitement permet de convertir une classification à l'échelle du pixel en une classification à l'échelle des peuplements.

- L'utilisation de Rasterstats et de règles de décision assure une classification robuste.
- L'évaluation par matrice de confusion permet d'identifier les améliorations possibles.

RÉSULTATS ET ANALYSES

04

ANALYSE DES RÉSULTATS

Cette section présente une analyse factuelle des résultats obtenus à travers les différentes étapes du projet. L'analyse est divisée en deux parties :

1. Analyse des échantillons (distribution et variabilité des classes).
2. Analyse des cartes des essences et des peuplements (résultats de classification et performance du modèle).

Analyse des Échantillons

Objectif Évaluer la répartition et la représentativité des échantillons utilisés pour entraîner le modèle de classification.

1. Répartition des polygones par classe (Diag_baton_nb_poly_by_class.png)

- Certaines classes sont sur-représentées, tandis que d'autres sont sous-représentées. ;
- La majorité des polygones appartient à quelques classes dominantes, ce qui peut affecter l'équilibre du modèle.

2. Répartition des pixels par classe (Diag_baton_nb_pix_by_class.png)

- Certaines classes, bien que peu nombreuses en polygones, couvrent une grande surface en pixels.
- Cela souligne un déséquilibre entre la distribution des polygones et celle des pixels, pouvant influencer la classification.

3. Distribution du nombre de pixels par polygone (Violin_plot_nb_pix_by_poly_by_class.png)

- Forte variabilité intra-classe pour certaines classes, indiquant des disparités de taille entre les polygones.

Certaines classes ont des polygones de taille relativement homogène, tandis que d'autres montrent une forte dispersion.

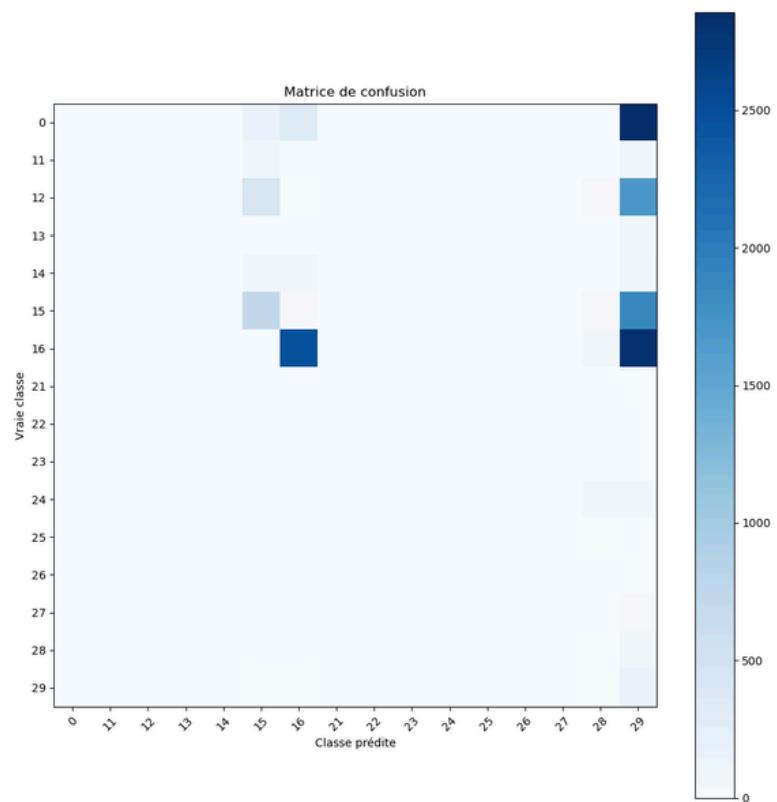


Figure n° 12 : matrice de confusion de la classification des peuplements

Interprétation

- Les échantillons sont inégalement répartis entre les classes, avec un déséquilibre notable entre le nombre de polygones et le nombre de pixels par classe.
- Certaines classes sont plus homogènes que d'autres, ce qui peut affecter la capacité du modèle à bien les différencier.

Analyse des Cartes des Essences et des Peuplements

Objectif Évaluer les performances de la classification des essences et des peuplements en s'appuyant sur les matrices de confusion et la répartition des classes prédites.

1. Carte des essences forestières (carte_essences_echelle_pixel.tif)

- La répartition des classes d'essences suit globalement les tendances observées dans les échantillons.
- Certaines classes sont bien représentées, tandis que d'autres sont peu visibles sur la carte, suggérant des confusions dans la classification.

2. Matrice de confusion des essences (matrice_CV5folds_stratified_group_x30times.png)

- Score global (Overall Accuracy, OA) : 30%.
- Forte confusion entre certaines classes, notamment celles ayant des signatures spectrales proches.
- Certaines classes sont bien classées, avec des taux de précision et de rappel élevés.

Constats

- Le modèle montre des performances limitées, avec des erreurs de classification fréquentes entre certaines essences.
- Les classes les plus représentées dans les échantillons sont généralement mieux prédites, tandis que les classes rares souffrent d'une forte confusion.

3. Carte des peuplements forestiers (Sample_BD_foret_T31TCJ_stand.shp)

- L'agrégation des classes d'essences en classes de peuplements produit une carte plus homogène.
- Certains peuplements dominent la carte, reflétant la distribution des essences classifiées.

4. Matrice de confusion des peuplements (matrice_confusion_peuplements.png)

- Des confusions persistent, bien que réduites par l'agrégation des classes d'essences.
- La transition de la classification pixel vers la classification par polygone améliore légèrement les performances globales.

Constats

- L'agrégation des classes d'essences en peuplements permet d'atténuer certaines erreurs de la classification à l'échelle du pixel.
- Les peuplements les plus représentés sont mieux classifiés, mais des confusions restent présentes entre certains types de peuplements.

Synthèse des Résultats

ANALYSE	RÉSULTATS CLÉS
Répartition des échantillons	Déséquilibre entre le nombre de polygones et le nombre de pixels par classe.
Variabilité intra-classe	Certaines classes ont des polygones très dispersés en taille, ce qui peut affecter la classification.
Performance de la classification des essences	Score global OA = 30%, avec des confusions marquées entre certaines classes spectrales proches.
Classification des peuplements	L'agrégation des classes améliore légèrement la classification, mais des erreurs persistent.

Tableau n° 9 : Tableau de synthèse des résultats

DISCUSSION DES RÉSULTATS

05

DISCUSSION DES RÉSULTATS

Analyse des confusions et limites méthodologiques

L'analyse des matrices de confusion et des graphiques de qualité révèle plusieurs points clés quant à la performance de la classification des essences forestières et des peuplements. L'accuracy globale du modèle à l'échelle du pixel est de 35% avec une faible dispersion, ce qui indique une stabilité du modèle mais un taux de classification correct limité. De nombreuses confusions sont visibles dans la matrice de confusion, notamment entre certaines classes forestières proches en termes de caractéristiques spectrales.

Ces confusions peuvent s'expliquer par plusieurs facteurs :

1. Spectres similaires entre certaines essences : Les classes qui se recouvrent spectrlement, notamment les formations mélangées, sont plus difficiles à discriminer.
2. Qualité et représentativité des échantillons d'entraînement : Un déséquilibre dans le nombre d'échantillons par classe, comme montré dans les graphiques d'analyse des échantillons, peut entraîner un apprentissage biaisé en faveur des classes majoritaires.
3. Résolution spatiale et mélange des pixels : À l'échelle du pixel, un même pixel peut contenir plusieurs types d'essences, ce qui entraîne des erreurs de classification dans les zones de transition.
4. Stratégie de validation croisée : Bien que la validation croisée stratifiée par polygone soit utilisée, elle ne garantit pas une séparation parfaite des données entre l'entraînement et le test, ce qui peut induire des résultats optimistes ou pessimistes selon les zones analysées.

Pertinence des résultats et limites des données

L'analyse des peuplements montre une amélioration par rapport à la classification au pixel, ce qui est attendu étant donné que lagrégation par polygone permet de lisser les erreurs locales. Cependant, des erreurs persistent, notamment pour les classes de peuplement en mélange, confirmant la difficulté de les différencier.

Les limites des données utilisées sont également un facteur clé :

- Qualité des images Sentinel-2 : Certaines bandes spectrales ont une résolution inférieure à 10m et nécessitent un rééchantillonnage, introduisant un risque de perte d'information.
- Temporalité des données : Une seule saison de données NDVI est utilisée, alors que certaines essences montrent une variabilité temporelle forte qui pourrait aider à leur distinction.
- Masque forêt : Les erreurs éventuelles dans la définition des zones forestières peuvent impacter la classification, notamment si certaines zones classées comme "forêt" comportent en réalité des couverts mixtes.

Interprétation et applications des résultats

Malgré ses limites, la carte des essences et celle des peuplements restent exploitables dans plusieurs contextes :

- Gestion forestière : La classification peut aider à identifier des tendances générales sur la répartition des essences et des peuplements, notamment pour des inventaires à grande échelle.
- Cartographie de la biodiversité : Les informations sur les types de peuplement sont essentielles pour les études écologiques et la gestion des habitats.
- Suivi du changement : La méthodologie pourrait être appliquée sur plusieurs années pour suivre l'évolution des forêts et détecter d'éventuelles perturbations.

Axes d'amélioration et perspectives

Plusieurs pistes peuvent être envisagées pour améliorer les résultats :

1. Améliorer la base d'échantillons : Une meilleure répartition des échantillons pour équilibrer les classes et éviter le surapprentissage sur certaines essences.
2. Exploiter d'autres indices spectraux : En complément du NDVI, d'autres indices (NDWI, EVI, etc.) pourraient enrichir la discrimination des classes.
3. Optimiser le modèle de classification : Tester d'autres classifieurs (XGBoost, SVM) ou ajuster les hyperparamètres du RandomForestClassifier pour réduire l'overfitting.
4. Inclure des données multi-temporelles : L'analyse des signatures phénologiques sur plusieurs saisons pourrait aider à mieux discriminer certaines classes.
5. Post-traitement de la classification : L'intégration de méthodes de filtrage spatial ou d'agrégation post-classification pourrait améliorer la cohérence des résultats.

En conclusion, bien que perfectibles, ces cartes constituent une première base exploitable pour une cartographie forestière automatisée et ouvrent la voie à des améliorations méthodologiques pour affiner la précision des classifications.

2025

